Meta-Augmented Human:
From Physical to Cognitive
Towards Affective State Recognition

Thesis approved by the Department of Computer Science
Technische Universität Kaiserslautern for the award of the Doctoral Degree
Doctor of Engineering (Dr.-Ing.) to **Shoya Ishimaru**

Date of Defence:	May 10th 2019
Dean:	Prof. Dr. Jens Schmitt (Technische Universität Kaiserslautern)
Reviewers:	Prof. Dr. Prof. h.c. Andreas Dengel (Technische Universität Kaiserslautern),
	Prof. Dr. Paul Lukowicz (Technische Universität Kaiserslautern),
	Prof. Dr. Koichi Kise (Osaka Prefecture University)

D 386

Abstract

This thesis investigates how smart sensors can quantify the process of learning. Traditionally, human beings have obtained various skills by inventing technologies. Those who integrate technologies into daily life and enhance their capabilities are called augmented humans. While most existing augmenting human technologies focus on directly assisting specific skills, the objective of this thesis is to assist *learning* – the meta-skill to master new skills – with the aim of long-term augmentations.

Learning consists of cognitive activities such as reading, writing, and watching. It has been considered that tracking them by motion sensors (in the same way as the recognition of physical activities) is a challenging task because dynamic body movements could not be observed during cognitive activities. I have solved this problem with smart sensors monitoring eye movements and physiological signals.

I propose activity recognition methods using sensors built into eyewear computers. Head movements and eye blinks measured by an infrared proximity sensor on Google Glass could classify five activities including reading with 82 % accuracy. Head and eye movements measured by electrooculography on JINS MEME could classify four activities with 70 % accuracy. In a wild experiment involving seven participants who wore JINS MEME more than two weeks, deep neural networks could detect natural reading activities with 74 % accuracy. I demonstrate *Wordometer 2.0*, an application to estimate the number of rear words on JINS MEME, which was evaluated in a dataset involving five readers with 11 % error rate.

Smart sensors can recognize not only activities but also internal states during the activities. I present an expertise recognition method using an eye tracker which performs 70 % classification accuracy into three classes using one minute data of reading a textbook, a positive correlation between interest and pupil diameter ($p < 0.01$), a negative correlation between mental workload and nose temperature measured by an infrared thermal camera ($p < 0.05$), an interest detection on newspaper articles, and effective gaze and physiological features to estimate self-confidence while solving multiple choice questions and spelling tests of English vocabulary.

The quantified learning process can be utilized for feedback to each learner on the basis of the context. I present *HyperMind*, an interactive intelligent digital textbook. It can be developed on *HyperMind Builder* which may be employed to augment any electronic text by multimedia aspects activated via gaze.

Applications mentioned above have already been deployed at several laboratories including Immersive Quantified Learning Lab (iQL-Lab) at the German Research Center for Artificial Intelligence (DFKI).

Acknowledgements

I am deeply grateful to Prof. Andreas Dengel, my dissertation advisor. I sincerely thank him for giving me a great opportunity to work at the University of Kaiserslautern and the German Research Center for Artificial Intelligence (DFKI). He always gives me insightful comments and suggestions.

I would like to give a warm thank you to Prof. Koichi Kise and Prof. Kai Kunze. They have supervised me by sharing interesting research topics since I was a Bachelor student at Osaka Prefecture University. Furthermore, I would like to thank the members of my PhD committee: Prof. Sebastian Michel and Prof. Paul Lukowicz.

I would like to express my gratitude to Dr. Nicolas Großmann for his kind coordination at Immersive Quantified Learning Lab. I would like to thank Brigitte Selzer, Dr. Thomas Kieninger, Dr. Syed Saqib Bukhari and Dr. Sheraz Ahmed for giving me valuable advice. They gave me insightful comments whenever I had problems.

I would like to acknowledge Prof. Jochen Kuhn, Jun. Prof. Pascal Klein, Dr. Carina Heisel, Dr. Stefan Küchemann, and Michael Theese for frequent supports on the context of U.EDU project. Special thanks to Prof. Andreas Bulling, Prof. Masahiko Inami, Prof. Yutaka Arakawa, Assoc. Prof. Olivier Augereau, Dr. Tilman Dingler, and Dr. Benjamin Tag for constructive discussions. I was inspired by their research ideas many times. In addition, JINS MEME hardware and software implementations were often supported by Yuji Uema and Katsuma Tanaka.

Most importantly, thanks to my students: Chris Allison, Emil Baitemirov, Iuliia Brishtel, Prerna Garg, Anirban Ghosh, Dayananda Herurkar, Jan Holub, Jawad Hussain, Ali Mert İnal, Fransisco Ito, Soumy Jacob, Nidhi Kamath, Sai Kumar, Hangfan Liu, Takanori Maruichi, Yuya Ohbayashi, Apurba Roy, Jayasankar Santhosh, Arka Sinha, Ann-Sophie Steinert, Ko Watanabe, and Kent Yamada. This thesis would not have been completed without their hard work and kind help.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

One of the roles of technology is to extend human abilities relating to different aspects. For instance, optical eyeglasses make it possible to see objects clearly regardless of their individual differences. Bicycles, automobiles, ships, and airplanes enable people to travel far away quickly. Computers support cognitive tasks including memory and calculation. The general motivation underlying my research is to invent such technologies that augment the abilities of human beings and make our daily lives comfortable.

Researchers in the field called *Augmented Human* have demonstrated that human beings and technologies perform better when they collaborate rather than when working individually [52]. However, there are still several scenarios in which technologies are of no help to people (e.g., for social activities, tasks requiring special attention, and those outside of radio waves or batteries). The reality is that the more people depend on technologies, the more their original performances may deteriorate over time [56]. This thesis concerns this problem and solves it by proposing a concept of *Meta-Augmented Human*. Compared to the general augmented human, it aims to amplify the ability of the human, not only temporarily, while using external devices, but permanently even if when there are not any direct supports.

This chapter presents the basic concepts of Meta-Augmented Human and introduces a summary of the research problems/contributions addressed in the thesis.

1.1 Basic Concepts

Creating and using a tool is not a limited ability for human beings. Many animals know that it is possible to augment their abilities with tools. For instance, sea otters break shellfish with stones. Chimpanzees hunt ants by stirring a twig in a nest. Green herons drop leaves into a pond to catch fish by making them believe the leaves are food. Then, how does the use of tools differ in human begins and animals? The answer is *meta-creation*, the ability to create more complex and advanced tools by combining existing tools [168]. As Arthur C. Clarke said in his book [34], "The old idea that Man invented tools is therefore a misleading half-truth; it would be more accurate to say that *tools invented Man*". Human beings have invented several tools and gradually developed their abilities.

For physical examples, ancient human beings learned how to use tools from lithic reduction. The inventions of automobiles, ships, airplanes, and spacecraft made it possible to connect widely distant places. For moving with perfect freedom, recent researchers have proposed an artificial arm that traces the movement of a leg [154] and a wearable flying device [169].

Tools have also contributed to improving human cognitive abilities. It is not too much to say that the character is one of the greatest inventions. By using characters, human beings have obtained deep insights. Characters have also enabled them to communicate with each other beyond time and place. Typographical printing contributes to spreading knowledge widely. By using computers and network technologies, we can access the latest news all over the world. Computers also play an important role as a delegation of cognitive tasks [25]. Much of the office work that human beings have done so far can be performed by computers faster and more accurately. Artificial intelligence using deep learning techniques in recent years has been outperforming human performances even in recognition and learning tasks [160].

1.1.1 Influences of the Augmentations to Human Minds

Andy Clark insists that we are already *Natural-Born Cyborgs* even if we do not embed any mechanical parts in our bodies [33]. We have obscured the boundary between our own bodies and tools by delegating confidence in abilities to instruments such as paper and pen. When we suddenly lose the tools, the ability fails as though we lost a part of our body. Due to the invention of various cognitive assistants, this trend is accelerating year by year. As a result of these dependencies, we are gradually losing abilities including memorization, mental map-making, and so on. Nicholas Carr summarized such examples of degeneration as *Glass Cage* [56]. Hundreds of notifications on smartphones have decreased the ability to concentrate. According to recent surveys, it takes 20 minutes to recover the concentration aborted by a notification, and 28 % of office hours are used for browsing social media. In order to avoid such problems, Apple and Google have released a function to track the usage time of applications and limit the use as a new function on their latest operating systems. On the other hand, tools may influence us in a positive way. According to Maryanne Wolf, we do not have specific parts of the brain that correspond to reading when we are born, but we obtain these advanced abilities by combining multiple functions in the brain through reading [187].

1.1.2 Meta-Augmented Human: Augmented Reading and Learning

On the basis of the positive and negative influences mentioned above, I believe that researchers should not only develop technologies which make our lives more convenient, but also investigate the impact of these technologies on our bodies and minds. Then, what kind of assistance leads us in a good direction, for instance? Through my doctoral study, I have come up with an idea that assistance systems improving reading habits and learning experiences could have a great potential and usefulness in the daily life.

A good habit makes humans healthier and smarter. It contributes to accurate decision-making and high productivity. Therefore, there are many applications and services that motivate people to stay physically fit. However, the approaches relating to cognitive activities have been investigated/implemented to a lower extent compared to physical activities. Since the relationship between cognitive benefits (e.g., vocabulary, academic scores, critical thinking) and reading habits, especially the increased reading volumes, has been well-explored [71, 137, 143], I focus on reading activity and propose a system that encourages people to read more. I believe that quantifying daily reading activities and giving feedback to people improve their cognitive habits and make them more intelligent.

Learning – the act of acquiring new knowledge, skills, abilities, and expertise – is one of the vital behaviors of human beings. In particular, people in the modern world are always required to learn new situational skills. The reason is that advances in technology are constantly changing their lifestyles and the ways they work. However, technology does not only require an acquisition of new skills; it can also provide an assistance by acquisition of these skills [43]. I propose an intelligent learning assistant that recognizes learners' behaviors and affective states according to several sensors and offers individual support for each learner.

Reading and learning are skills to master new skills, i.e., *meta-skills*. I believe that assisting them can augment in the long-term or permanently. Compared to the existing Augmented Human System, which extends abilities only when worn on the body, I define technologies assisting meta-skills as *Meta-Augmented Human*. One of the critical requirements towards the Meta-Augmented Human System in both reading and learning is recognition of behaviors by using sensors enough affordable to be worn at everywhere and everyday. I define such sensors as *smart sensors* and present several approaches and evaluations discussed in the next section.

1.2 Research Questions

I propose a research hypothesis: "Meta-skills can be quantified by smart sensors." Understanding human behavior is an especially important issue for which to develop systems supporting humans. It resembles the problem of grasping the hidden bottom shape of an iceberg. I propose the Iceberg Model of Activity Recognition (see Figure 1.1) and address following three research problems to claim my hypothesis.

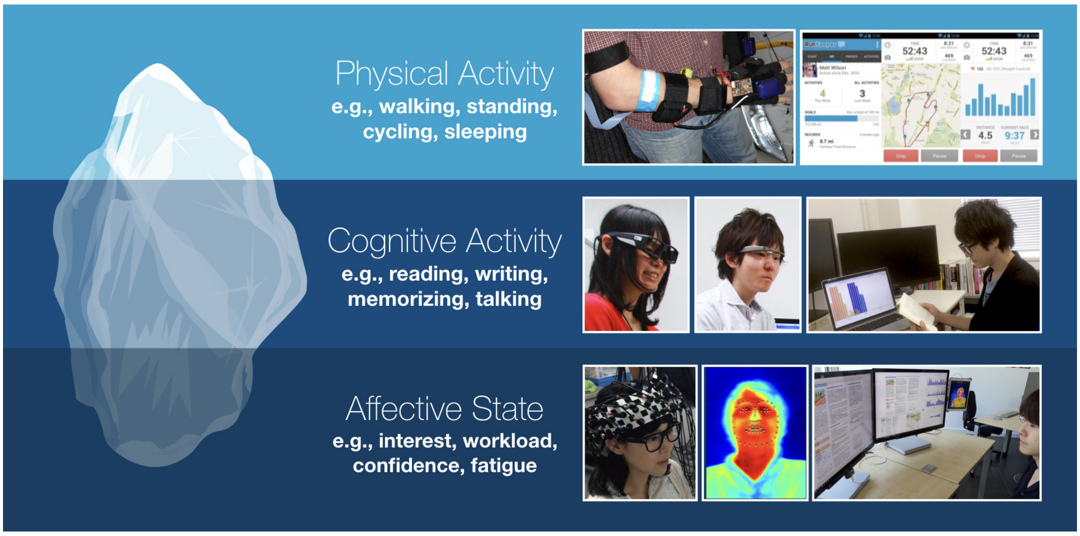

Physical Activity
e.g., walking, standing, cycling, sleeping
Cognitive Activity
e.g., reading, writing, memorizing, talking
Affective State
e.g., interest, workload, confidence, fatigue

FIGURE 1.1: Iceberg Model of Activity Recognition

How can smart sensors quantify reading activities in daily life? This problem is similar to when part of an iceberg is under the water and is invisible. To understand the shape, an underwater camera is required. As such, researchers have utilized advanced sensors to recognize cognitive activities. However, advanced sensors sometimes prevent users' natural behaviors and cannot be used in the wild. One of the problems addressed in this thesis is to investigate whether it is possible to quantify reading activities with sensors designed for everyday use.

How can smart sensors quantify affective states of learners? Affective states involving attention, interest, self-confidence, and cognitive load play important roles in the process of learning, and they are more difficult to recognize than activities.

This thesis proposes several affective state recognition methods through sensors that can be utilized in a classroom geared toward intelligent learning assistants.

How can a system augment reading/learning experiences? The last problem is about interventions. If we change documents dynamically/statically while reading, how will reading behaviors be influenced? By designing an intelligent digital textbook that displays contents dynamically for each reader, this thesis presents observations on reading behaviors of several conditions.

1.3 Contributions

In summary, contributions of this thesis include:

- An overview of state-of-the-art activity recognition and intervention
- Methods to recognize daily activities by using eyewear computers
- An application which quantifies reading to improve our cognitive abilities
- Methods to recognize learners' affective states using various sensors
- Development of an intelligent digital textbook and its ecosystem

Applications proposed in this thesis are deployed in several laboratories including Immersive Quantified Learning Lab (iQL-Lab) at the German Research Center for Artificial Intelligence (DFKI), Smart Sensor Room at Osaka Prefecture University (OPU), Ubiquitous Computing Systems Laboratory (UBI-Lab) at Nara Institute of Science and Technology (NAIST), StudySapuri Lab at Recruit Marketing Partners Co., Ltd., and THINK FUTURE English Academy in Risshikan Seminar Co., Ltd.

1.4 Outline of Thesis Chapters

Chapter 2 presents an overview of existing sensing methodologies, activity recognition, and interventions. Chapter 3 reports recognition of cognitive activities including reading in wearable devices which are designed for everyday use. *Wordometer 2.0*, a prototype of a reading tracker, is also demonstrated in this chapter. Chapter 4 presents quantified learning systems, which measure affective states of learners using several sensors. An eye tracker, an infrared thermal camera, and a physiological sensing wristband are utilized to recognize comprehension, interest, mental workload, and self-confidence. Chapter 5 presents *HyperMind*, an intelligent digital textbook, and the evaluation and a building tool. A negative influence of reading on digital media is also discussed in this chapter. Finally, Chapter 6 summarizes conclusions and future work.

Chapter 2

Background and Related Work

Since Douglas Carl Engelbart proposed the framework of *Augmenting Human Intellect* [52] and Mark Weiser created the terms *Ubiquitous Computing* [182] and *Calm Technologies* [183], many researchers have investigated the recognition of human activities by using several sensors for giving proactive assistance.

This chapter presents a literature survey about what types of sensors are utilized to recognize what types of activities and affective states. The survey starts from an overview of Activity Recognition research (Section 2.1). Then, details of eye tracking (Section 2.2) and physiological sensing (Section 2.3) are presented in separate sections because they are potential approaches that do not prevent activities of a user. Finally, Section 2.4 summarizes the applications.

2.1 Overview

The starting point of the research field of Activity Recognition was to recognize *what* a user is doing. As summarized in the survey by Lara and Labrador [115], several physical activities (e.g., walking, running, cycling, sleeping) can be recognized by motion sensors on the body [11, 26, 59, 184] or a smartphone [44]. On the other hand, recognition of cognitive activities (e.g., reading, writing, talking) is considerably restricted to the extent of the body movements. Frequently, cognitive activities occur without or with minimal bodily activity. In this case, additional sensors are required to recognize the activities that are taking place. One of the interesting approaches for solving this is to use eye movements. Eyewear computers hang in the balance between pervasiveness and potential [5]. For instance, Bulling *et al.* classified tasks including cognitive activities by using electrooculography sensors [21]. Shiga *et al.* recognized daily activities by mobile eye tracking glasses [159]. Biedert *et al.* presented a robust differentiation between reading and skimming [16].

The more sensors developed, the more researchers became interested in recognizing the context of human activities (i.e., *when*, *where*, *by whom*, and *why* the activity is performed). For example, first-person vision is utilized in the context recognition. Through an egocentric camera attached to the head or the body, activities and the contexts can be estimated from objects in front of the person [121]. The recognition of social interactions is also a key factor in understanding the context of talking activities [54]. Instead of the on-body sensors mentioned above, remote sensors (e.g., fixed cameras [45], microphones [163]) have also been employed to recognize contexts because they can record the interactions between humans and the environment.

The most abstract subject of recognition is *how* the activity is performed. An obvious approach to get inside how cognitive activities are performed is to use sensors that provide spatial or temporal resolutions of the brain activities. Magnetic resonance imaging (MRI) [37], electroencephalography (EEG) [62] and near-infrared spectroscopy (NIRS) [88] can be candidates, if we can accept the limitation of the recording environment. These limitations can be partly solved by an inclusion of physiological sensing. For instance, the autonomic nervous system (ANS) provides a good insight into the hidden mental processes [153]. Some of affective states (e.g., concentrations, mental workload, boredom) can be measured by changes in pupil diameter [100], nose temperature [1, 112], and Electrodermal activity (EDA) [19].

One of the critical issues in this research field is how to conduct experiments in natural settings for proposing robust methods. One major problem is the lack of unobtrusive technology to make long-term tracking possible. There are some datasets contributed by computer vision researchers working on egocentric vision, which are mostly camera recordings [40, 63]. Few datasets include eye gaze data [165].

2.2 Eye Tracking

Since readers use their eyes to understand text, measuring eye movements is a promising approach to understanding readers' behavior. Figure 2.1 shows an example of eye movements on text. Experiments conducted during the 19th century revealed that eye movements while reading are not always smooth but a series of rapid movements (saccade) and short stops (fixation) [181]. Small movements in a fixation (drift/tremor and micro-saccade) were observed by precise eye tracking methodologies in the 20th century. A micro-saccade is an involuntarily movement to keep vision stable during a fixation [122] and reflect attention [69]. Around 10 % of saccades while reading are moving in a direction opposite to the direction of reading (called as regression or re-reading) in order to understand the content of the text [104]. A blink – semi-autonomic rapid closing of the eyelid – has several strengths and frequencies depending on the activity, tiredness, and concentration [192]. In addition, a smooth pursuit occurs when a person tracks a moving object with a slow speed. However, this metric has not been considered in reading behavior analysis because most of the documents have static layouts.

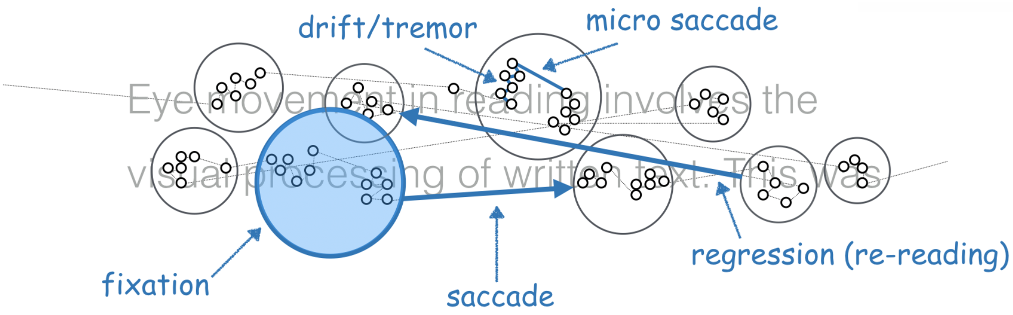

drift/tremor
micro saccade
Eye movement in reading involves the
visual processing of written text. This was
fixation
saccade
regression (re-reading)

FIGURE 2.1: Eye movements on text

2.2.1 Eye Tracking Methodologies

Several eye tracking methodologies have been proposed for several purposes (see Figure 2.2). This subsection explains the characteristics from the left to the right.

The search coil eye tracker requires a user to wear a contact lens with an electromagnetic coil, and the orientation of the eye boll is calculated by electrodynamics [147]. An advantage of this methodology is that it can measure highly precise movements. Although it has a long history, it has not been used in modern experiments because of the high demand for participants and experimenters. However, it is in the spotlight again with an appearance in virtual reality headsets [186].

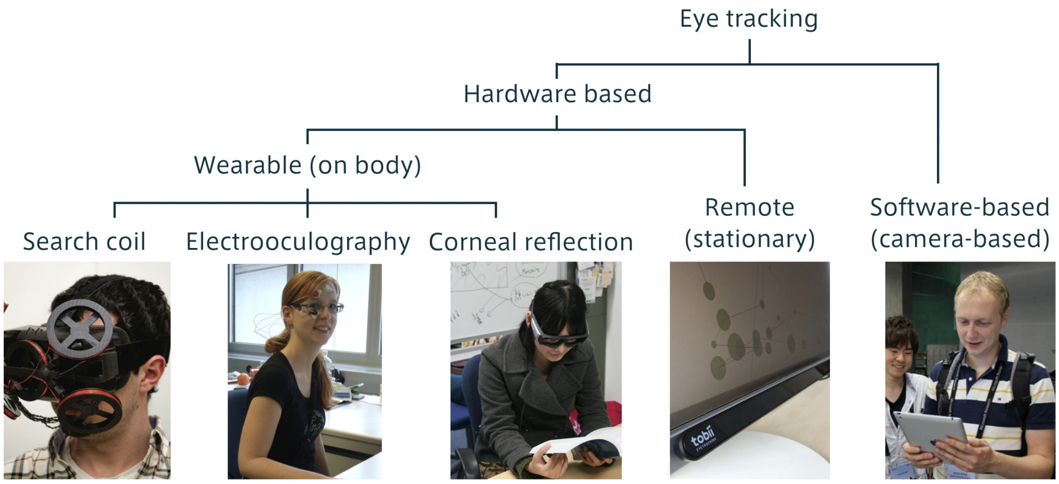


FIGURE 2.2: Eye tracking methodologies

The electrooculography (EOG) uses electrodes attached around an eye and measures voltages on the skin. It measures the corneo-retinal standing potential that exists between the front and the back of an eyeball. Traditional setups required four electrodes around an eye to recognize vertical and horizontal eye movements [22]. But recent sensing devices have revealed that signals from three electrodes on each nose pad and the forehead are enough for reading analysis [89, 90]

The corneal reflection eye tracker uses a light source to illuminate the eye, causing highly visible reflections, and a camera to capture an image of the eye, showing these reflections. It is integrated in mobile eyeglasses with a first-person perspective camera, and it projects the eye gaze of a user to the scene image. This metrology is used in several experiments because of its versatility and usability [139].

The remote (stationary) eye tracker that is attached to a display is a good option by which to analyze reading behaviors. The technical background of remote eye tracking is the same as the corneal reflection on glasses. However, it is attached to a display so that the researcher does not need to project eye gaze to a document (eye gaze is measured with a coordinate on a screen). Some vendors produce several remote eye trackers from 60 Hz to 2,000 Hz, depending on their purpose, which ranges from gaming to research.

The software-based (camera-based) eye tracking, i.e., estimating eye gaze without such specific hardware, is a challenging and hot topic in computer vision research. Researchers estimated the gaze on a mobile tablet by using a regression model [91], appearances of eyes [166, 194], and convolution of the neural network (CNN) with a crowdsourced dataset [106] or synthesized eye images [167, 195].

2.2.2 Eye Tracking while Reading

The relationship between cognitive tasks and eye movements have been well known in the field of cognitive science and psychology [161]. For example, Kliegl *et al.* explored eye movements in relation to cognitive tasks [104]. There are interesting findings in this line of research, including assessing expertise and other cognitive tasks using fixation features [32]. Rudmann *et al.* presented an extensive review of affective state detection using visual behavior [149].

Saccade speed and length with other measures achieved high accuracy in measuring human performance. On the other hand, Manuel *et al.* suggested a decrease in saccade speed indicated tiredness and an increase in the same indicated task complexity [12]. According to Rudmann *et al.*, the direction of saccades indicates repeated interest in an area and the importance of the area of interest in the current activity [149].

Fixation duration and fixation rate are indicators of an increase in attention on the current task [31]. They delved into the relevance of saccades in interpreting human mental effort in solving a task. They also found that an increase in blink interval and a decrease in blink rate indicated high mental effort and that studying the diameter of the pupil helps to realize the task difficulty and the cognitive effort.

The pattern of blinks is also one of the important features in recognizing activities. Bentivoglio *et al.* studied the relation between sitting activities and blink rates [13]. They described that the blink rate changes when participants are reading, talking, and resting. Acosta *et al.* presented the case that working with computers causes a reduction of blinks [3]. Haak *et al.* described that emotion, especially stress, affects blink frequency [68]. Orchard *et al.* also assessed the mental workload during reading by analyzing blinking patterns [136].

2.2.3 Eye Tracking while Learning

The behavior of a student who does not understand the contents of a document is characterized by low reading speed and frequent regressions [144]. Thai *et al.* proved that comprehension of a question by a student appears in his/her eye movement, for example, in the case of regressions of a question [175]. Okoso *et al.* investigated the relation between gaze patterns on difficult words [61] or difficult parts [135] of a document and comprehension.

Chen *et al.* collected students' responses to computer-based physics concept questions that were presented as either pictures or text. They guaranteed that students' eye movement behavior can predict computer-based assessment performance [30]. Martínez-Gómez *et al.* presented a formal framework to recognize the reader's level of understanding and language skill and gave measurements of reading behavior via eye gaze data [123]. Oliver *et al.* estimated the English skill of non-native English speakers from his/her eye behaviors in English test [8]. Klein *et al.* studied students' understanding and cognitive processing while they are solving multiple representation problems [102]. Daniel *et al.* recorded eye movements of students studying e-learning to investigate specific gaze patterns for predicting their concentration [41].

Moreover, about the relation between the behavior of eyes and self-confidence, it has been proved that low self-confidence is characterized by a frequent regressions of questions and long gaze on choices [105]. Yamada *et al.* estimated whether students answered confidently or not on multiple-choice questions [188].

2.3 Physiological Sensing

Another interesting approach to understand affective states is physiological sensing. Especially, the autonomic nervous system (ANS) controls smooth muscle and glads and its activity is largely unconsciously [153] and can be observed as several reactions including the heart rate, digestion, respiratory rate, pupillary response, urination, and sexual arousal (see Figure 2.3). This section focuses on some of the reactions that can be measured by sensors without affecting a user's behavior.

2.3.1 Pupil diameter

When the sympathetic nervous system is more active than the parasympathetic nervous system (e.g., when a person experiences a high workload), the diameter of the pupils increases. It can be measured with an eye tracker or a camera facing a user. Kucewicz *et al.* investigated a relationship between pupil diameter and memory [107]. In their experiment, they asked participants to memorize words displayed on a screen and found that pupil diameter was significantly larger while successfully recalled words were displayed than in other situations. Porta *et al.* observed that a decrease in pupil diameter at the end of the task indicated tiredness [141].

2.3.2 Nose Temperature

When the sympathetic nervous system is more active than the parasympathetic nervous system, blood vessels constrict and the temperature of the nose drops. It can be measured by an infrared thermal camera [86]. Abdelrahman *et al.* recorded temperatures of the nose and the forehead under different task difficulties and found significant changes [1]. The temperature can also be recorded by a small thermometer module attached to the nose. Yasufuku *et al.* developed glasses involving the module in order to detect stresses in daily life [190]. Kunze *et al.* measured the nose

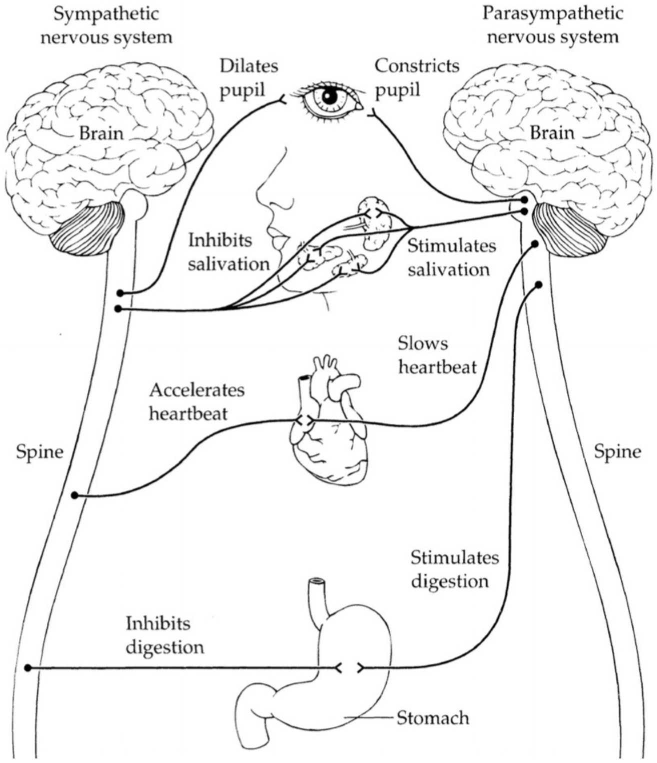


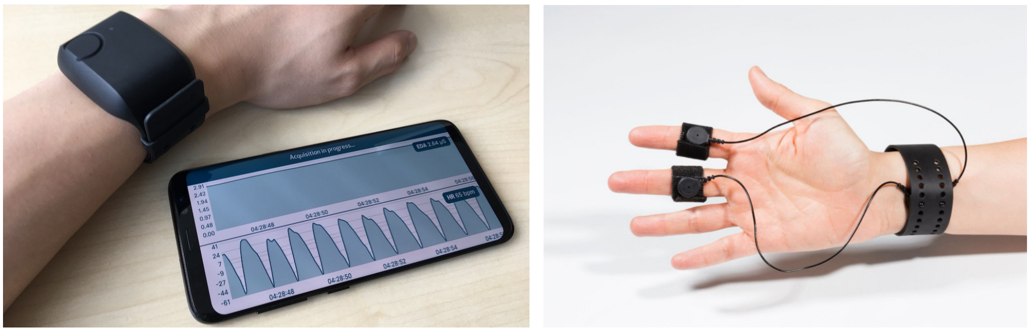
FIGURE 2.3: Autonomic nervous system. Image credit: [153]

temperature and classified engagements while reading into two classes (feeling interesting or boring) by combining it with eye blinks [112].

2.3.3 Electrodermal activity (EDA)

Electrodermal activity (EDA), also known as galvanic skin response (GSR), electrodermal response (EDR) and psychogalvanic reflex (PGR), is an electrical conductance of the skin and a sensitive physiological index of changes in ANS. It can be measured by electrodes on an arm or fingers (see Figure 2.4). Figure 2.5 shows an example of the sensor signal recorded while reading. EDA can be decomposed into two signals: the tonic component and the phasic component [38, 65, 67].

The tonic component, known as skin conductance level (SCL), refers to the baseline skin conductance level and spontaneous fluctuations in the component. Lazarus *et al.* showed that the SCL and the heart rate increased significantly during the presentation of violent films [116]. Nomikos *et al.* showed that even the expectation of an unpleasant event could cause a similar reaction in SCL as the event itself [133]. Multiple studies investigating the effect of the anticipation of electrical stimulation conversely assume that the rising of the SCL reflects an increased cognitive activity related to the avoidance of aversive events rather than an emotional component.



(A) Electrodes on an arm

(B) Electrodes on fingers

FIGURE 2.4: Experimental settings recording EDA

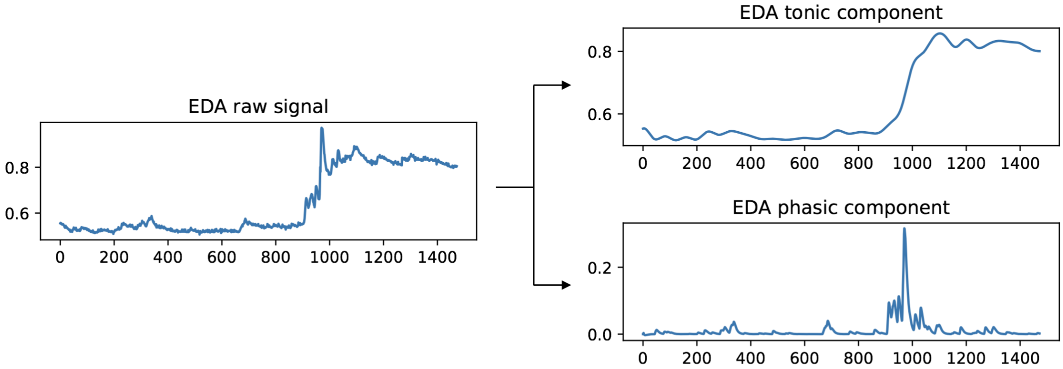


FIGURE 2.5: Decomposition of a raw EDA signal

The phasic component, also known as skin conductance response (SCR), is a high frequency phasic component reflecting the short-time response to the stimulus. The frequency of the non-specific SCRs reveals the emotional component of the stress reaction. Further studies used experimental settings that were closer to a real-life office environment than simple electrical stimuli. Several authors investigated how involuntary interruptions in the workflow due to long-system response times influenced the EDA. An increase of non-specific SCRs for long-system response times could be demonstrated. Jacobs *et al.* also showed that an increase in skin conductivity correlated with level of the mental stress [97]

Implicit emotional responses that may occur unconsciously (e.g., threat, anticipation, salience, novelty) can be examined using EDA. Setz *et al.* showed that the EDA peak height and the instantaneous peak rate depict the person's stress level [158]. Boucsein provides an extensive summary of EDA research in relation to stress [19]. He showed that the SCL and the non-specific SCRs are sensitive, valid indicators of stress, whereas other physiological measures (e.g., heart rate) do not show equal sensitivity [51].

2.3.4 Blood Volume Pulse (BVP)

Blood volume pulse (BVP) is the change in volume of blood over a given period of time. Certain emotions can trigger the release of hormones such as epinephrine and norepinephrine, which will increase blood flow to bring more oxygen to the muscles. BVP can be monitored using photoplethysmography (PPG), which is a non-invasive technique that relies on light absorption and reflection. The signals detected form a wave that represents the change in blood volume relative to heart rate. Adjacent local peaks in the wave indicate heartbeats, and the time interval between these peaks is the inter-beat interval (IBI). Heart rate variability (HRV), IBI, and the raw signal of BVP have been associated with frustration and anxiety [113].

The E4 wristband[1] has often been used in recent studies because both EDA and PPG sensors are integrated. In the context of reading behavior analysis, Matsubara *et al.* recognized emotional arousal while reading a comic [126], and Sanches *et al.* classified the categories (comedy, romance, or horror) [152].

2.4 Intervention and Applications

Researchers in Human-Computer Interaction (HCI) have proposed applications to change the behavior of a user based on the activities recognized by sensors.

The initial interactive eye tracking application in reading was implemented for entertainment [17] and the real-time usage in education has not been considered explicitly for long. The first gaze-oriented application focusing on educational aspect was *iDict* by Hyrskykari *et al.* [79]. It provides translations for comprehension problems detected in the reader's gaze patterns. Then Biedert *et al.* introduced an application for assisted and augmented reading called the *eyeBook* [14]. The idea behind the eyeBook is to create an interactive and entertaining reading experience which helps the reader to understand the text better. Biedert *et al.* created a framework to construct gaze-responsive real-time interactions to enhance the reading experience (e.g., displaying images, translations, footnote, and bookmarks) as *Text 2.0* [15].

Eye gaze has also been used in adaptive scrolling algorithms [108], for example, for continuous reading of newspaper articles in large public displays by Lander et al [114]. They used head-mounted eye trackers in a multi-user scenario to study the effect of this approach on the user's reading speed. Lee *et al.* proposed building a virtual tutor to support a student's learning [117]. This work proved that eye communications with a virtual tutor enhance the efficiency of learning.

Reading experiences are getting more and more interactive. For instance, Yannier *et al.* proposed haptic feedback to enhance reading experiences [189]. Gaze-oriented

[1] https://www.empatica.com/en-int/research/e4/

interventions have also a high potential in the virtual reality [27, 49]. But it is required to design and adapt reading experiences for this new environment again. According to an experiment by Dingler *et al.*, a majority of participants preferred white text on a black background as opposed to black text on a white background in the virtual reality [46].

The visual attention measured by eye gaze is useful to select which information should be shown to a user. Toyama *et al.* proposed *Attention-Aware Systems* including a gaze guided object recognition using a head-mounted eye tracker [174], an augmented reality reading assistant combining document retrieval and eye tracking [172], and a gaze-oriented personal assistant for museums and exhibits [173].

Presenting information to users changes their behavior not only voluntary but also involuntary. Futami *et al.* proposed *Success Imprinter* presenting a stimulus a reminding success using psychological conditioned information improves the mental performance [60]. Researchers in their group also demonstrated that if a system displays a false heart rate lower than an actual value, a user believes the value and it starts decreasing.

Chapter 3

Cognitive Activity Recognition

This chapter presents work towards a cognitive activity tracker aiming for making people smarter by motivating them to read more in their daily lives. The cognitive benefits of reading (e.g., better vocabulary skills) and the benefits of increased reading volumes, are well explored in the fields of education and cognitive science [39]. As people can be physically fit by monitoring step counts [128], tracking the volume of reading can be a starting point to improve their reading habits. Following three steps are required to realize such an application.

The first step is a classification. Classifying reading from other daily activities is challenging compared to classifying walking because reading does not always require a lot of body movements. In order to solve this problem, Section 3.1 proposes a blink detection algorithm using an infrared proximity sensor equipped on Google Glass and an activity recognition method with features from eye blinks and head motions [93]. Section 3.2 proposes an activity recognition method on JINS MEME: commercial electrooculography glasses. JINS MEME is light, visually familiar, relatively inexpensive compared to other wearable devices, and has a sufficient battery for all-day use [89, 90].

The second step is a reading detection in the wild environment. There is a strong gap between *controlled reading* designed carefully in the laboratory and *natural reading* without any limitations in the wild. Section 3.3 reports results of an experiment in the wild asking participants to wear JINS MEME more than two weeks (seven participants, 880 hours recording in total) to investigate the difficulty. This section proposes deep learning based approaches using the large scale dataset [85].

The last step is an estimation of the number of read words. Section 3.4 demonstrates word count estimation algorithm using JINS MEME [83, 87].

In addition, Section 3.5 demonstrates a quick labeling interface on Google Glass to correct ground truth labels of activities in the wild experiment as a late-breaking work toward large scale recording [94].

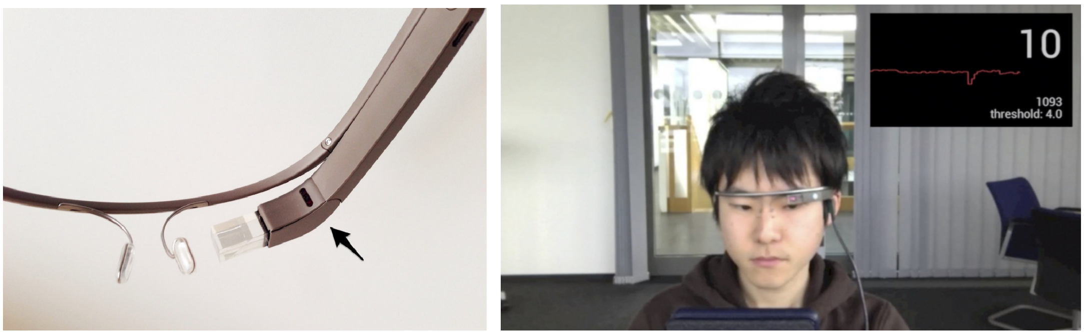
3.1 Eye Blink Based Activity Recognition on Google Glass

This section proposes a method to recognize cognitive activities by using Google Glass. While it is well known that eye movement is correlated with user activities [23], the aim of this section is to show that (1) eye blink frequency data from an unobtrusive, commercial platform which is not a dedicated eye tracker is good enough to be useful and (2) adding head motion patterns improves the recognition.

We adopt Google Glass for the first step to recognize cognitive activities. It has four sensors that could potentially be used: a camera, a microphone, an inertial measurement unit (IMU), and an infrared proximity sensor facing towards the users' eye as shown in Figure 3.1 (a), which can be used for blink detection. This section focus on the latter two (IMU and blink detection), which we argue to be most characteristic to the Google Glass platform.

There is also a lot of existing work on head mounted cameras [80]. However, one problem with computer vision approach is that it requires a lot of processing and thus might be impractical due to battery constraints. The motion sensors and the proximity sensor seem to be the most promising modalities. Especially, combining eye blink frequency and head motion patterns for daily activity recognition has so far not been studied in much detail.

Regarding a blink detection, Chau *et al.* applied an image processing [29] and Bulling *et al.* proposed eye tracking approaches [21, 23]. However, the situations of these existing blink detectors are limited. We focus on another approach which doesn't distract the user by bulky hardware. As far as we know, we are the first to use a simple proximity sensor embedded in a commercial wearable computing system for activity recognition and to combine it with head motion patterns.



(A) Device overview (B) Glass Logger

FIGURE 3.1: Infrared proximity sensor built into Google Glass

3.1.1 Approach

Figure 3.2 shows the overview of the proposed method. Our blink frequency based activity recognition is based on two stages. The first stage is the pre-processing blink detection which extracts the time stamps of blinks. Secondly, the main part of our algorithm calculates features based on the detected blinks.

Note that Google Glass doesn't have an official application programming interface (API) to provide the raw data of infrared proximity sensor. We obtain a root-permission of Google Glass on the basis of Glass hacking tutorial[1] and implement our own logging application[2] shown in Figure 3.1 (b) for the data recording.

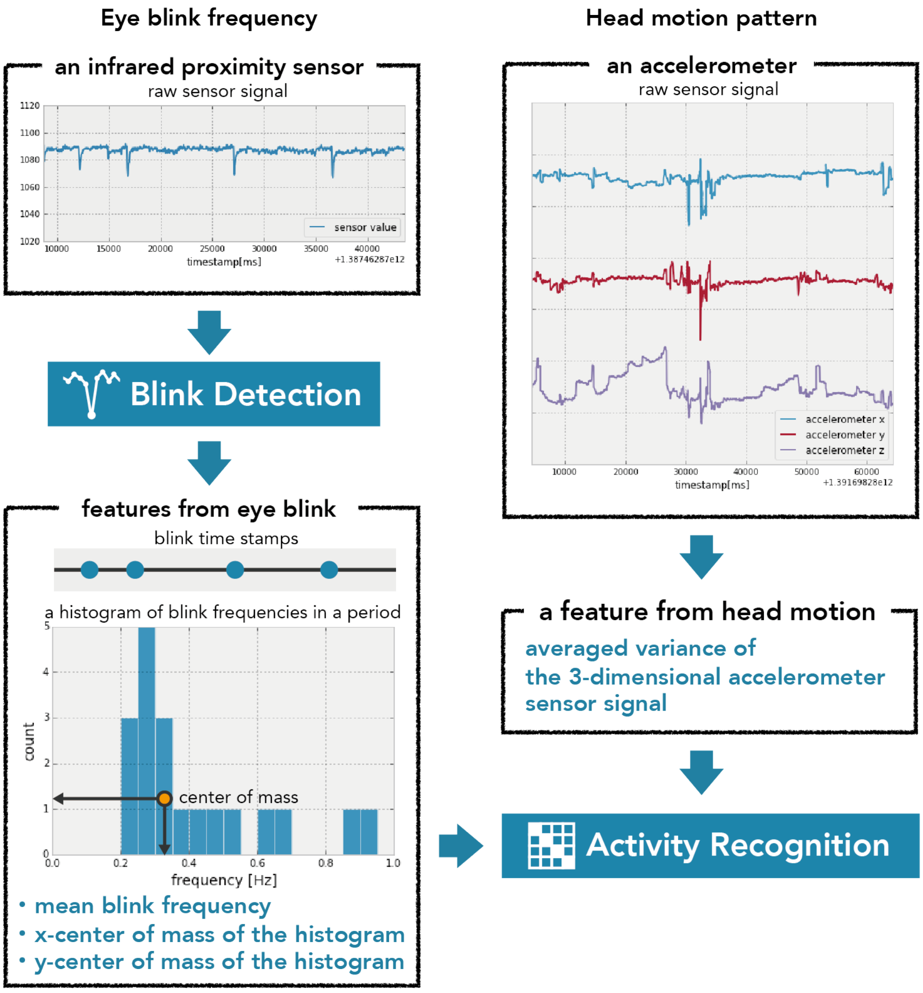


FIGURE 3.2: The overview of proposed activity recognition method

[1] https://developers.google.com/events/io/sessions/332704837

[2] https://github.com/shoya140/GlassLogger

Blink detection

Blinks are detected based on the raw infrared proximity sensor signal. A sliding window is moved on the sensor data stream and monitors whether the center point of each window is a peak or not by following definition. The distance from one sensor value of the center point in the window (p_5 in Figure 3.3) to the average value of other points (p_1, p_2, p_3, p_7, p_8 and p_9) is calculated. The preceding and subsequent points of the center (p_4 and p_6) are excluded from the average calculation because their sensor values are often affected by the center point. If the distance is larger than a threshold, the center point is a peak. Peaks in one second are combined and defined as a blink because one blink sometimes contains some peaks.

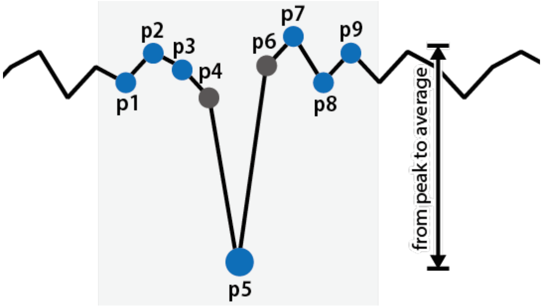


FIGURE 3.3: Blink detection based on the peak calculation

As the shape of the face and eye location vary, the best threshold for the peak detection varies for each user. Figure 3.4 with the same scale for each sub-graphic also demonstrates different signal variations for different users. The best threshold (in 0.1 steps ranging from 3.0 to 7.0) is calculated by evaluating the accuracy based on the ground truth information. This approach can be applied only in off-line evaluation. In on-line usage, the blink detecting application needs a few seconds for calibration before detection. During the calibration term, Google Glass urges the user to blink as matching some timing. The application gets sensor values and actual blink timing from calibration and evaluate the best threshold.

Blink Frequency Based Activity Recognition

As an output of the blink detection, timestamps of blinks are extracted. To recognize activities, a three-dimensional feature vector is computed by the timestamps. One is the mean blink frequency which describes the number of blinks during a period divided by the length of a period. Two other features are based on the distribution of blinks. Graphically, this can be understood as the histogram of the blink frequency. Figure 3.5 shows a histogram with a period of 60 seconds. The x-axis describes the mean blink frequency (0.0 - 1.0 Hz) and the y-axis describes the blink counts of each

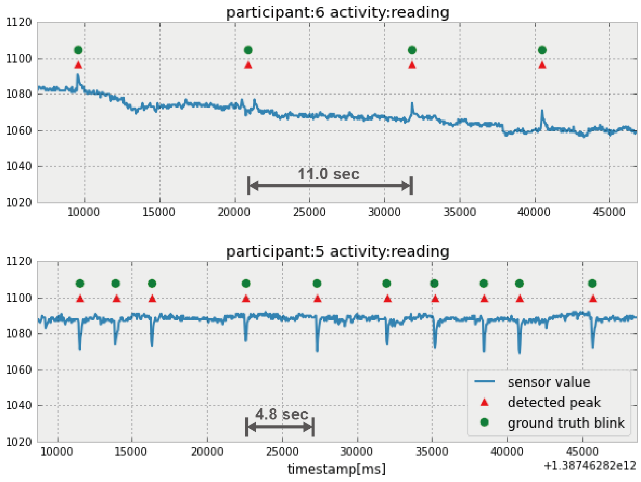


FIGURE 3.4: Proximity sensor values and ground truth of two participants.

frequency. The number of specified bins per histogram is 20 having a resolution of 0.05 Hz. The frequency value is calculated as inverse value of the interval between two blinks. The second and third features are defined as the x-center of mass and the y-center of mass of the histogram.

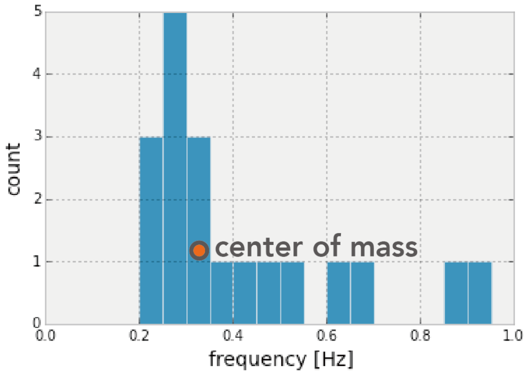


FIGURE 3.5: A histogram of blink frequencies during a period.

Head Motion Based Activity Recognition

The user's head motion pattern varies for different activities. A feature from head motion pattern is based on the degree of the head movement. It is computed as the averaged variance of the three-dimensional accelerometer sensor signals. Figure 3.6 shows the calculation process. Accelerometer sensor signals of three-dimension (x, y and z) in a period are recorded and each variance is calculated. The average value of three variances is computed and used as a feature.

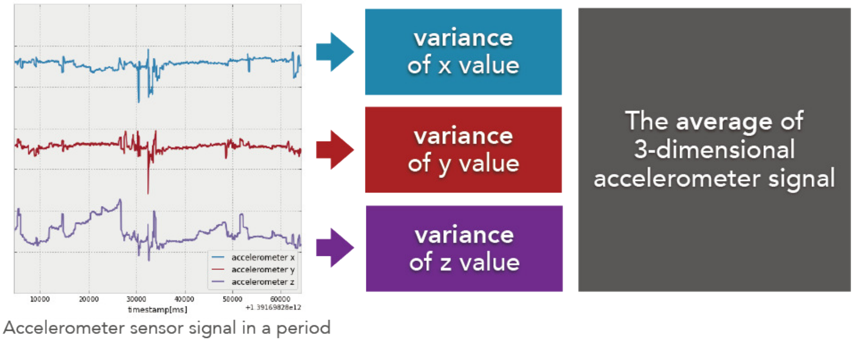


FIGURE 3.6: Calculation process of a feature from head motion

Combination of Blink Frequency and Head Motion

The advantage to use Google Glass is that we can easily get and combine several sensor signals for our activity recognition task. In the approach combining blink patterns and head motion we use the following four features: variance value of accelerometer, the mean value of blink frequency and the x-center and y-center of mass value of blink frequency histogram. We combine these features and compare the impact on activity recognition accuracy.

3.1.2 Experimental Design

We evaluated three different classification methods (using only eye blink frequency, using only head motion patterns and using all features from eye blink and head motion) on a data set containing five class activities and compared them.

We recruited eight participants to perform five activities each lasting five minutes while wearing the Google Glass. All of the participants were male. Five of them had unaided vision and three (participants 2, 3, and 4) were using contact lenses. The activities were defined as watching a movie on a Laptop, reading a book on an ebook reader, solving mathematical problems on paper (entrance examination for graduate school), sawing cardboard and talking with another person. Solving was intended as an example of mental tasks and sawing was selected as a physical task. The location and light condition was fixed for all experiment participants. The display of Google Glass was always turned off and did not attract the participant's attention during the experiment. We collected values of the infrared proximity sensor and the accelerometer. Each activity was recorded separately because activity spotting and segmentation is not yet implemented. Feature extraction and classification were applied to the data containing a single activity.

We also recorded the video of experimental scenes. Figure 3.7 shows the overviews of five different users performing five different activities. After the experiments, one person labeled all blinks by using the videos.

At the recognition part a sliding window was moved and each data in the window are classified. The window size was defined as 60 seconds with a step size of 10 seconds. The window size should be longer than max interval in dataset. The longest blink interval through all participants was 50 seconds (see Table 3.1 for details). We trained a user dependent J48 decision tree classifier which is implemented on Weka[3] and evaluated the classification accuracy by confusion matrices based on 10-fold cross validation.

TABLE 3.1: Dataset overview from ground truth

	Experiment Participants								
	1	2	3	4	5	6	7	8	Avg.
Total blink counts	230	161	420	313	381	309	207	414	**304**
Min frequency (Hz)	0.02	0.02	0.03	0.06	0.02	0.04	0.02	0.02	**0.03**
Max frequency (Hz)	0.96	0.96	0.99	0.99	0.99	0.98	0.99	0.99	**0.98**

3.1.3 Results and Discussion

Blink Detection

Figure 3.8 shows one participant's five different histograms based on the blink frequency distribution during five minutes for each activity. We evaluated the blink detection according to actual blink timestamps from ground truth videos. The average precision was 80 % with 78 % recall (see Table 3.2 for details). Each participant's blink detection results are based on the average value of five activities.

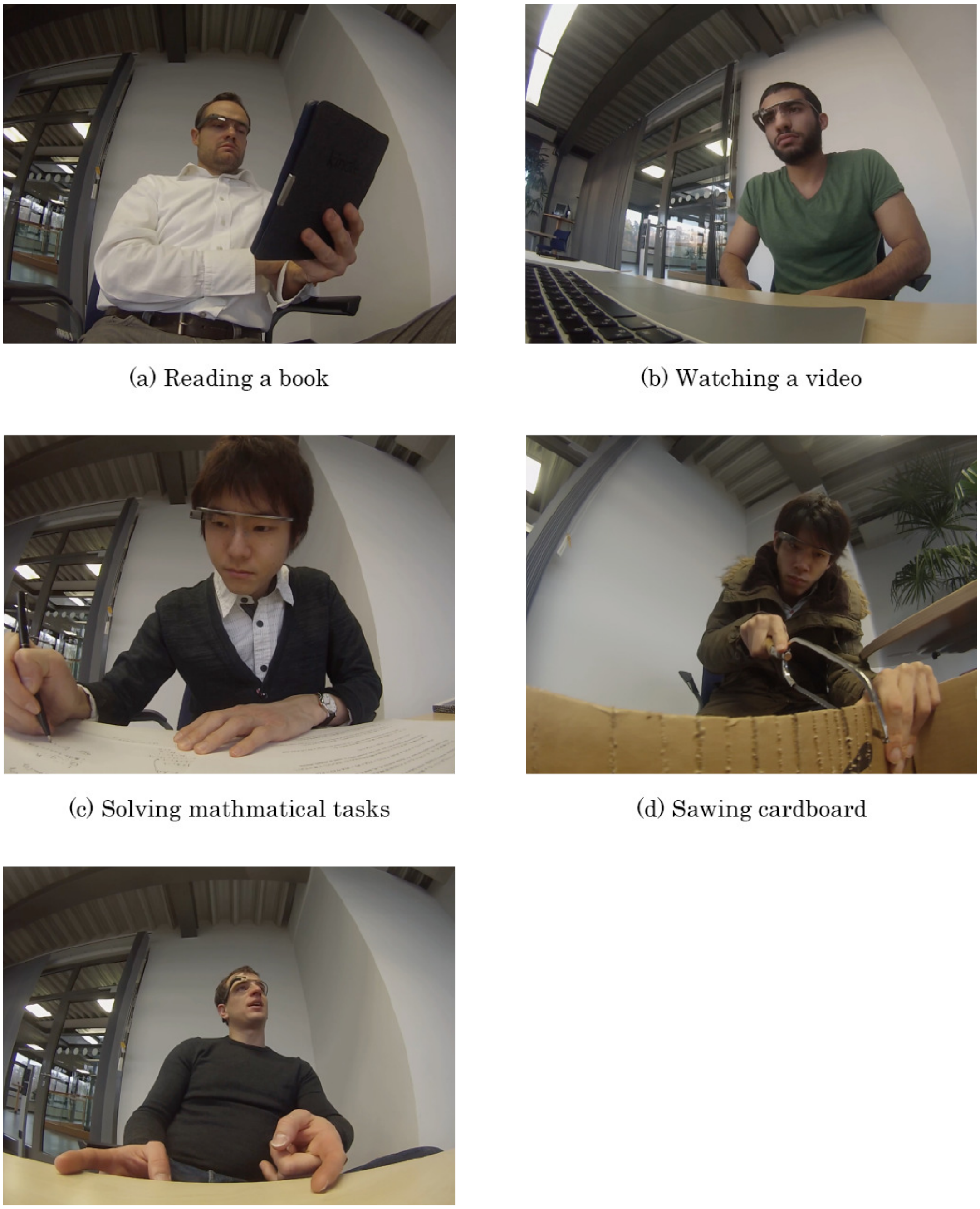
TABLE 3.2: Pre-processing blink detection results

	Experiment Participants								
	1	2	3	4	5	6	7	8	Avg.
Precision (%)	92	48	76	98	87	80	71	86	**80**
Recall (%)	64	71	72	98	89	86	72	74	**78**

Activity Classification

Solely based on blink frequency features and an experimental complexity of eight participants and five activity classes we achieved an average classification accuracy

[3] http://www.cs.waikato.ac.nz/ml/weka/



(a) Reading a book

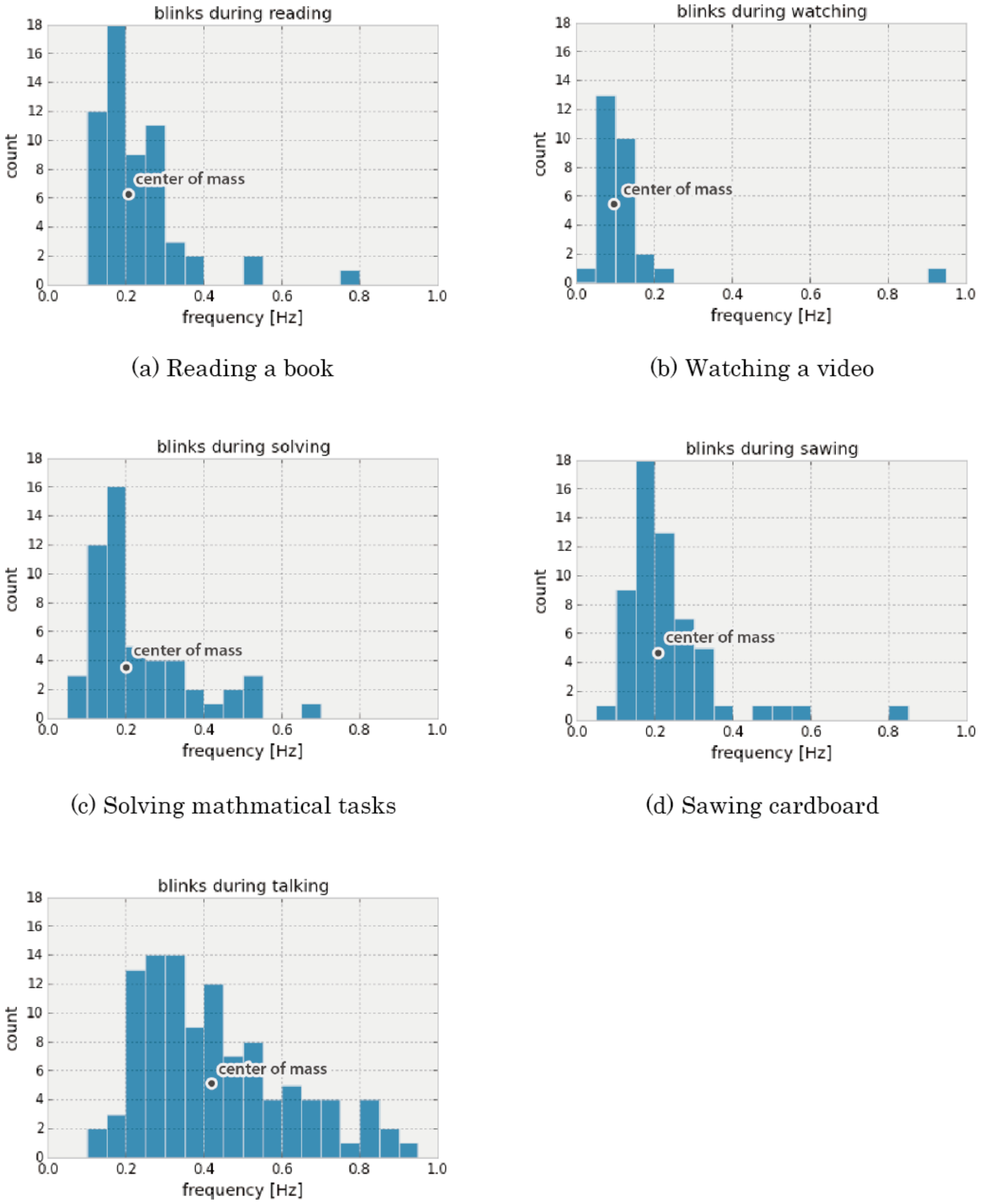
(b) Watching a video

(c) Solving mathmatical tasks

(d) Sawing cardboard

(e) Talking

FIGURE 3.7: Video-based ground truth image excerpts of the experiment scenes



(a) Reading a book

(b) Watching a video

(c) Solving mathmatical tasks

(d) Sawing cardboard

(e) Talking

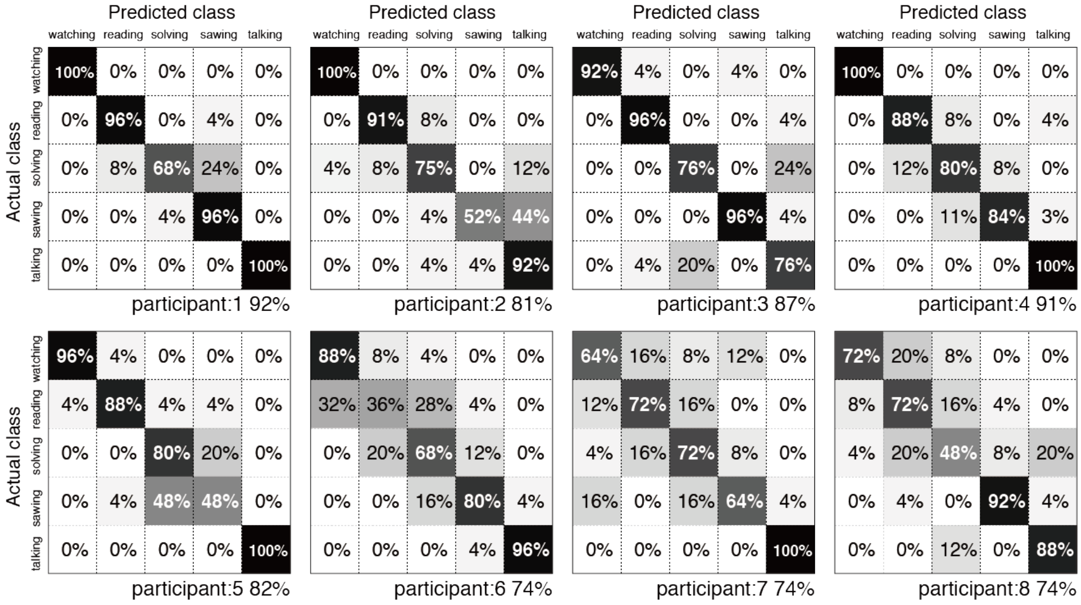
FIGURE 3.8: One participant's blink frequency histograms during five minutes recording.

of 67 % (see Table 3.3 for an overview of all participants) individually ranging from 52 % to 82 %. Solely motion feature based recognition underperformed with 63 % classification accuracy. When we combine the blink frequency based features with the motion based feature we achieve an average classification accuracy of 82 % (increased by 15 % compared to blink frequency based recognition). Figure 3.9 shows the individual confusion matrix results of eight experiment participants. These confusion matrices show correctly classified instances on the diagonal and wrongly classified instances in other areas.

The training duration per class and per person was only five minutes long. In future the input of the correct activity might be given during daily usage of Google Glass learning constantly from the user's activities and improving the classification constantly. We evaluated ten minutes of recordings of six participants (1, 4, 5, 6, 7 and 8) again. The classification based on blink frequency improved by 7 % an in combination with the motion feature improved by 9 % compared to the five minute long recording.

TABLE 3.3: Activity classification results

	Experiment Participants								
	1	2	3	4	5	6	7	8	Avg.
Blink frequency based classification (%)	70	52	82	76	70	54	69	64	**67**
Head motion based classification (%)	75	56	66	83	57	56	58	50	**63**
Combination based classification (%)	92	81	87	91	82	74	74	74	**82**



	watching	reading	solving	sawing	talking
watching	100%	0%	0%	0%	0%
reading	0%	96%	0%	4%	0%
solving	0%	8%	68%	24%	0%
sawing	0%	0%	4%	96%	0%
talking	0%	0%	0%	0%	100%

participant:1 92%

	watching	reading	solving	sawing	talking
watching	100%	0%	0%	0%	0%
reading	0%	91%	8%	0%	0%
solving	4%	8%	75%	0%	12%
sawing	0%	0%	4%	52%	44%
talking	0%	0%	4%	4%	92%

participant:2 81%

	watching	reading	solving	sawing	talking
watching	92%	4%	0%	4%	0%
reading	0%	96%	0%	0%	4%
solving	0%	0%	76%	0%	24%
sawing	0%	0%	0%	96%	4%
talking	0%	4%	20%	0%	76%

participant:3 87%

	watching	reading	solving	sawing	talking
watching	100%	0%	0%	0%	0%
reading	0%	88%	8%	0%	4%
solving	0%	12%	80%	8%	0%
sawing	0%	0%	11%	84%	3%
talking	0%	0%	0%	0%	100%

participant:4 91%

watching	96%	4%	0%	0%	0%
reading	4%	88%	4%	4%	0%
solving	0%	0%	80%	20%	0%
sawing	0%	4%	48%	48%	0%
talking	0%	0%	0%	0%	100%

participant:5 82%

watching	88%	8%	4%	0%	0%
reading	32%	36%	28%	4%	0%
solving	0%	20%	68%	12%	0%
sawing	0%	0%	16%	80%	4%
talking	0%	0%	0%	4%	96%

participant:6 74%

watching	64%	16%	8%	12%	0%
reading	12%	72%	16%	0%	0%
solving	4%	16%	72%	8%	0%
sawing	16%	0%	16%	64%	4%
talking	0%	0%	0%	0%	100%

participant:7 74%

watching	72%	20%	8%	0%	0%
reading	8%	72%	16%	4%	0%
solving	4%	20%	48%	8%	20%
sawing	0%	4%	0%	92%	4%
talking	0%	0%	12%	0%	88%

participant:8 74%

FIGURE 3.9: Confusion matrices of all participants.

Feature Visualization

Figure 3.10 represents all participant's feature plot. Talking and watching is easily distinguished by other activities. But it is difficult to classify sawing, reading and solving by only blink patterns. Head motion feature helps to distinguish especially those classes. Conversely, reading and watching can not be distinguished easily only by head motion. The dispersion of head motion during solving is larger than other activities because solving contains 2 statuses, concentrating to write the answer and looking at the assignment on another paper.

3.1.4 Conclusion

This section proposed a method to recognize high-level activities by using eye blink frequency and head motion patterns delivered from Google Glass. The method was evaluated on a data set containing five activities (reading, watching, solving, sawing, talking) of eight participants showing 67 % recognition accuracy for eye blink only and 82 % when extended with head motion patterns. We have shown how the infrared proximity sensor from the standard Google Glass can be used to acquire user's eye blink statistics and how such statistics can be combined with head motion pattern information for the recognition of high-level activities.

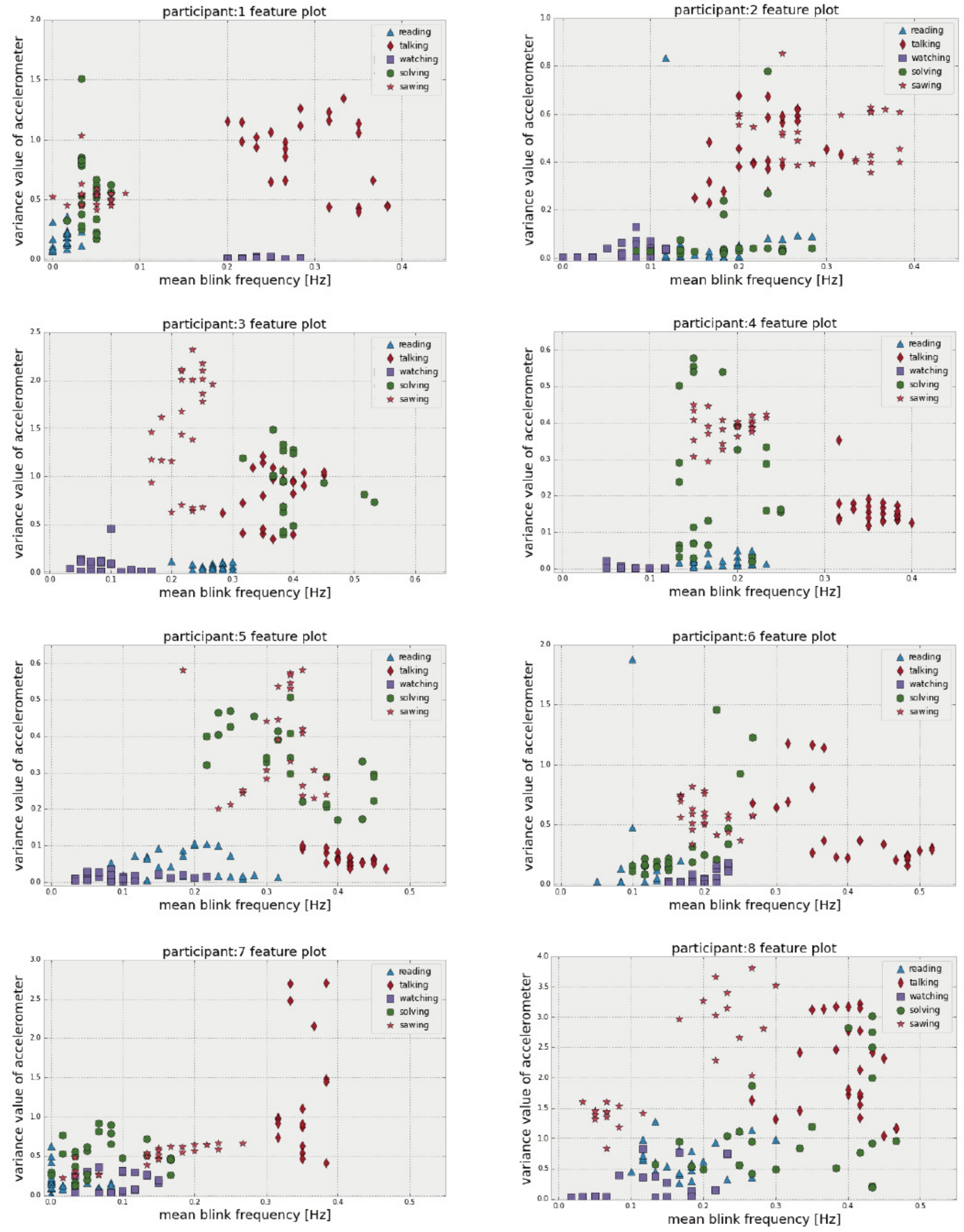


FIGURE 3.10: Feature representation of eight participants and five different activity classes.

3.2 EOG Based Activity Recognition on JINS MEME

Research and industry get more and more interested in Eyewear Computing, from Google Glass, over Epson Moverio, the Oculus Rift to the Sony Smart Glasses. However, most of these designs still look a bit clunky and emphasize displays (augmented or virtual reality), not the sensing aspect. This section focuses on a device which is really designed for everyday use and can be used for daily activity tracking.

We define a *smart eyewear* as a device which doesn't require much load to a user and has an enough long-life battery for all day use (see Figure 3.11). While high spec head-mounted wearable devices are suitable for specific scenarios such as entertainment, industrial, educational and medical field [131, 185], smart eyewears are better for tracking daily activities.

The aim of this section is to investigate the potential of smart eyewear. We show the feasibility of JINS MEME commercial electrooculography (EOG) glasses to detect simple eye movements and apply them to activity recognition involving reading. Contributions of this section are two-fold: (1) motivating that smart eyewear is interesting for ubiquitous computing applications, as it enables to track activities that are hard to observe otherwise, especially in regard to cognitive tasks and (2) evaluating specific smart glass, a prototype of JINS MEME for their use for activity recognition tasks. We show a signal level evaluation and simple classification task of four activities (two participants five minutes per activity). Both indicate that the device can be used for more complex scenarios.

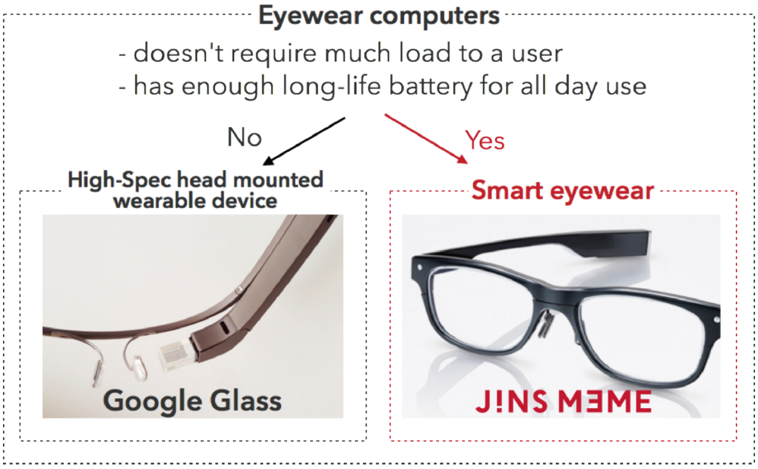
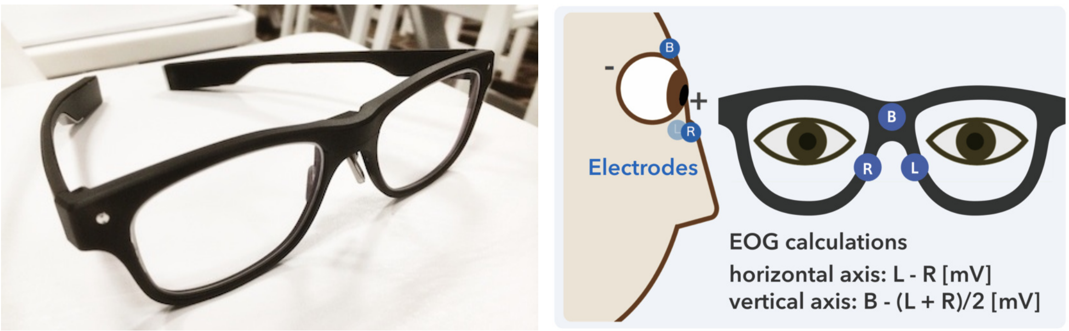

Eyewear computers
- doesn't require much load to a user
- has enough long-life battery for all day use
No
Yes
High-Spec head mounted wearable device
Smart eyewear
Google Glass
J!NS M∃ME

FIGURE 3.11: Definition of a smart eyewear in this thesis.

3.2.1 Approach

Figure 3.12 (a) shows the overview of JINS MEME. It is unobtrusive and looks close to normal glasses. It is equipped with 3 electrodes to detect eye movements and inertial measurement unit (IMU), as well as a Bluetooth Low Energy module to stream the data to a computer or smartphone. The electrodes sample data with 100 Hz, the motion sensor with over 50 Hz. It has a long-life battery which runs over 16 hours.

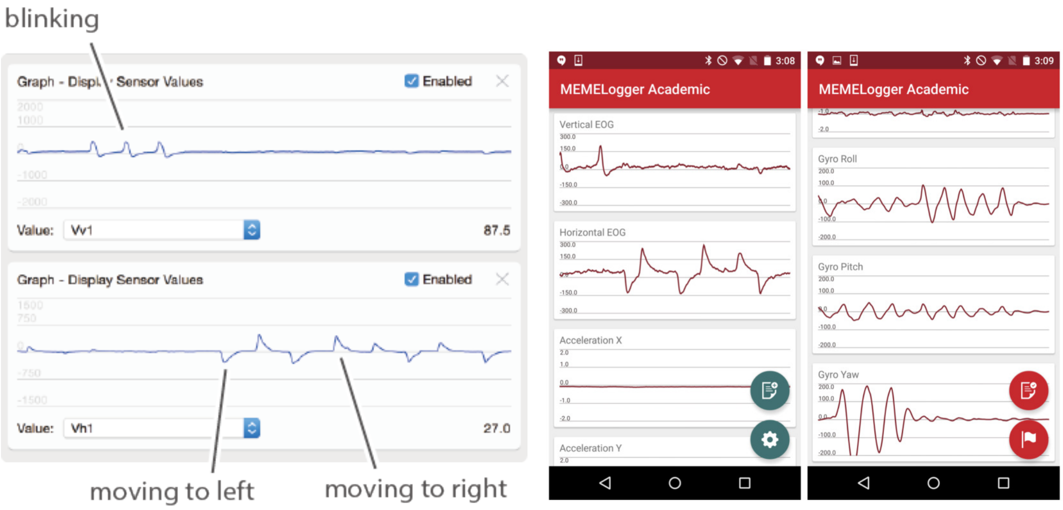
As shown in Figure 3.12 (b), one ground electrode (*B* in the figure) and two active electrodes (*R* and *L* in the figure) are equipped on JINS MEME. The EOG vertical component is calculated as the difference between *B* and an average of *L* and *R*, and the horizontal component is calculated as the difference between *L* and *R*.


B
-
+
L R
Electrodes
B
R
L
EOG calculations
horizontal axis: L - R [mV]
vertical axis: B - (L + R)/2 [mV]

(A) Overview of the hardware

(B) Three electrodes measring eye movements

FIGURE 3.12: JINS MEME



(A) MEMELogger for macOS

(B) MEMELogger for Android

FIGURE 3.13: MEMELogger

We record sensor signals by using applications developed by us shown in Figure 3.13. Figure 3.13 (a) represents EOG vertical (the top part of figure) and horizontal (the bottom part of figure) component including some eye blinks and left/right eye movements. They are detected as peaks on the sensor signals.

From recorded data, we create a six second sliding window (overlap: two second) and calculate seven features shown in Table 3.4 in the window. Then *K*-Nearest Neighbor classifier (k-NN; *k*=5) is applied to classify the activity for each window.

TABLE 3.4: Features for k-NN activity classification

No.	Feature
1-2	{mean, variance} of vertical EOG component
3-4	{mean, variance} of horizontal EOG component
5	variance of acceleration x
6	variance of acceleration y
7	variance of acceleration z

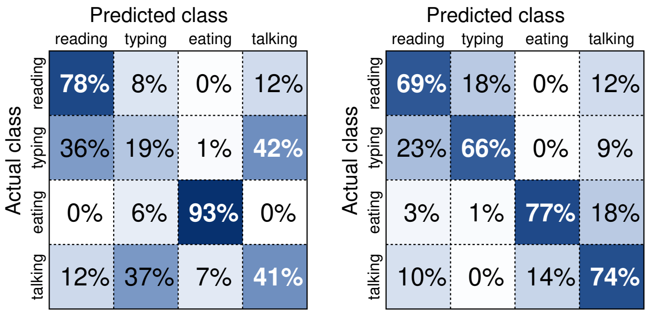
3.2.2 Experimental Design

We evaluated if the sensor data from the glass can distinguish more complex activity recognition tasks. We assume that modes of locomotion etc. can easily be recognized by the motion sensors alone. Therefore we concentrated on tasks performed while sitting in a common office scenario. We included four activities: typing a text in a word processor, eating a noodle dish, reading a book and talking to another person.

Two participants were asked to wear JINS MEME and perform the four activities, each activity for five minutes two times, in total 80 minutes. Before starting recording, we adjusted the electrodes on the device toward the facial features of the user to be sure to capture a clean EOG signal. This initial setup step needs need to be done only once per user. We applied the windowed feature extraction and a user-independent classification, i.e., training with the data of one user and evaluating with the another user.

3.2.3 Results and Discussion

We reached classification rates of 58 % by using only EOG signals and 70 % if IMU signals are integrated. The confusion matrices are given in Figure 3.14. Strengthened by the good performance distinguishing the four activities for two users in a user-independent approach. One interesting finding from the results is that there are a lot of confusions between typing and talking when we utilize only EOG signals but they were classified by the combination of EOG and IMU. This might be because both activities contain frequent vertical eye movements.



Actual class \ Predicted class	reading	typing	eating	talking
reading	78%	8%	0%	12%
typing	36%	19%	1%	42%
eating	0%	6%	93%	0%
talking	12%	37%	7%	41%

(A) EOG only (acc.: 58 %)

Actual class \ Predicted class	reading	typing	eating	talking
reading	69%	18%	0%	12%
typing	23%	66%	0%	9%
eating	3%	1%	77%	18%
talking	10%	0%	14%	74%

(B) EOG and IMU (acc.: 70 %)

FIGURE 3.14: Confusion matrixes of activity recognitions using EOG and IMU

Based on previous work in Section 3.1, it is possible to classify different activities offline using just blink detection and head motion. We extend this work, and implement an online reading/talking detection using JINS MEME. We tested a prototype of our reading/talking detection system on 12 people. They performed 15 instances of each reading and talking as well as a 5-6 instances of other unrelated activities (drinking water, eating etc.). By using the same features, only 16 instances were wrongly classified leaving us with 91 % of accuracy. No instances of other activities were classified wrongly, indicating that our online system works as well as the offline classifier implemented on Google Glass data.

3.2.4 Conclusion

This section presented interaction and recognition demonstrations using an unobtrusive EOG glasses prototype. The applications are meant to show the potential of the device category. We show the feasibility of detecting eye movements with our prototype. Our pilot study with two participants and four activity tasks (reading, typing, eating, and talking) have shown that the EOG and IMU signals can classify the four activities with 70 % accuracy. We have not designed application cases and suitable eye gestures people might want to use, this is left for future work.

3.3 Reading Detection in the Wild by Deep Neural Networks

Reading in real life occurs in a variety of settings, involving various devices and document layouts that would result in irregular eye movements. Our assumption is that there is a substantial difference between controlled reading and natural reading, meaning that the reading detection methods that work in labs may not necessarily be usable in the wild.

We believe that developing less obtrusive optical eye trackers is key to achieving reading quantification in real life settings. In this regard, using commercially available electrooculography (EOG) glasses seems promising since they are relatively light, visually familiar (looking like conventional eyewear) and have sufficient battery life for all-day use. Their cost is also relatively low, making them suitable for conducting large-scale data recording [5]. In this work, we record natural reading activities using commercial EOG glasses (see Figure 3.15) and evaluate the accuracy of the detection algorithm in the wild.

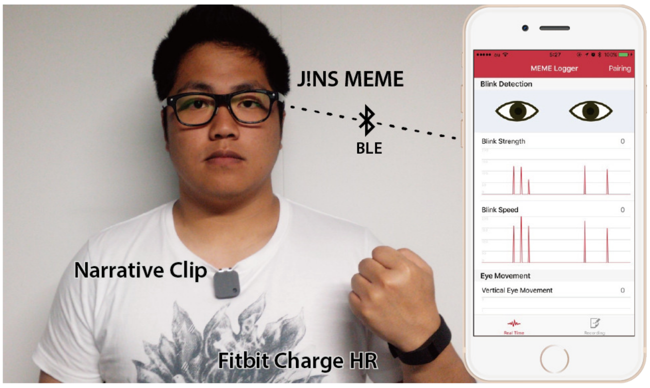


FIGURE 3.15: The recording setup. Participants wore JINS MEME, Narrative Clip, Fitbit Charge HR and a smartphone every day for more than two weeks.

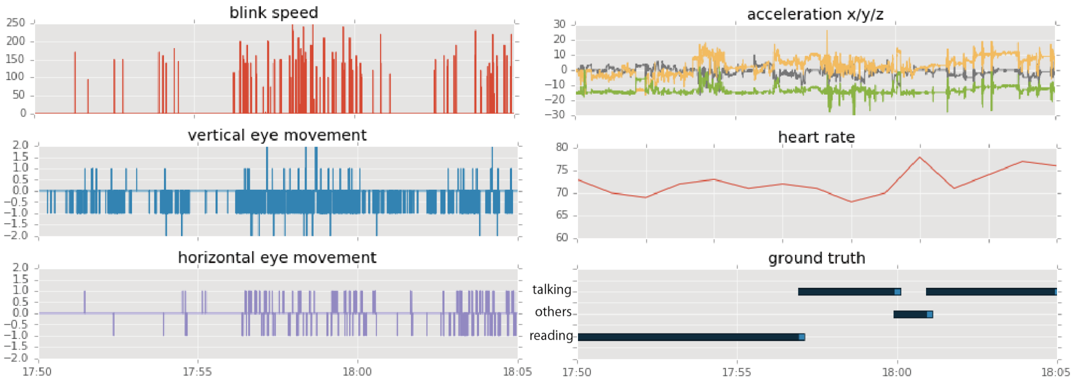


FIGURE 3.16: An overview of the sensor signals exported by JINS MEME. The ground truth is annotated by the participant at the end of the day with reviewing images.

3.3.1 Approaches

We propose three types of reading detection approaches. The first is a manual feature extraction based approach. In this approach, we analyze the data obtained from the devices to find characteristic sensor patterns during reading, and select the features for manual classification. The second and third approaches are automatic feature extraction based. We designed a convolutional neural network (CNN) and recurrent neural network with Long short-term memory (LSTM) for classifying the raw data. They extract best features by training with large-scale data.

We utilize JINS MEME[4] for the sensing. The device is equipped with three electrodes for eye movement detection and a 6-axis inertial measurement unit (IMU) for head movement detection. It is developed by JIN CO., LTD. The company has released two models of the device: the developer's version and the academic version. We used the former for this research, as it's widely available to consumers. JINS MEME calculates basic eye movements (blink speed, blink strength, two-step strength of up/down/left/right eye movements) internally on the device itself as shown in Figure 3.16, and stream them with IMU data to a smartphone via Bluetooth Low Energy. The sampling rate is 20 Hz. Battery run time is 18 hours, which is sufficient for gathering data all day during the day. Data are recorded on the iOS application, MEMELogger[5] and sent to a hosted server every day.

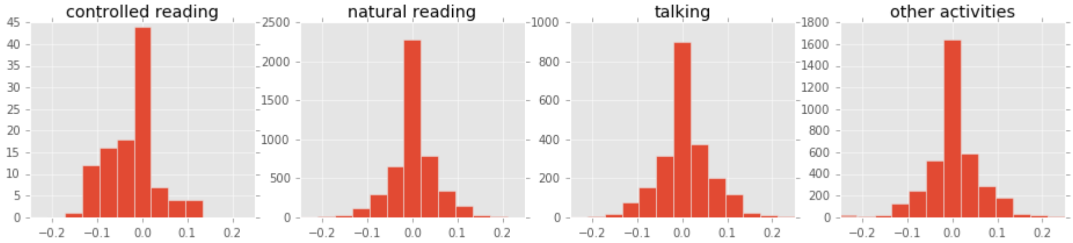


FIGURE 3.17: Histograms of horizontal eye movements.
Positive values represent movements to the right.

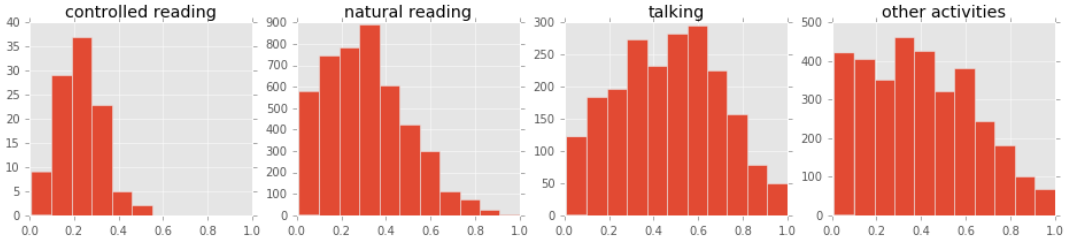


FIGURE 3.18: Histograms of blink frequencies

[4] https://jins-meme.com/en/

[5] https://itunes.apple.com/en/app/memelogger/id1073074817

SVM Based Reading Detection

We employ 16 statistical features from the sensor signals of JINS MEME (10 features from eye movements and 6 from head movements) as shown in Table 3.5. The frequencies of eye blinks and eye movements are calculated as inverse values of the duration between two blinks or eye movements. Acceleration x, y and z are raw signals. We create samples with the window size of 60 seconds. After the data were normalized and whitened, we calculate the mean and standard deviation for each of the sensor values in the window. A Support Vector Machine (SVM) with a radial basis function kernel (RBF) kernel is used for the learning. After the classification, we applied a majority vote for 5 minutes' worth of data to smooth out the results.

TABLE 3.5: Features for the SVM based reading detection approach

No.	Feature
1-2	{mean, SD} frequency of eye blinks
3-4	{mean, SD} frequency of eye move up
5-6	{mean, SD} frequency of eye move down
7-8	{mean, SD} frequency of eye move left
9-10	{mean, SD} frequency of eye move right
11-12	{mean, SD} raw signal of acceleration x
13-14	{mean, SD} raw signal of acceleration y
15-16	{mean, SD} raw signal of acceleration z

CNN Based Reading Detection

An overview of the CNN architecture is shown in Figure 3.19. The network receives raw sensor values from JINS MEME as inputs, and classifies the gaze activity as either reading or not reading. For the input layer, 6 maps with a size of 400×1 were created from 400 frames of 6 sensors' values, including blink speed, vertical eye movement, horizontal eye movement, acceleration x, y, and z. To increase the number of training samples, we employed different input window size (20 seconds) compared to SVM. There is no overlap between windows. The network has two convolution layers, each followed by a pooling layer. For the first convolution layer, the approach utilize a filter with size 12×1 with step 2 that exports 8 maps. Since the convolution is done without zero-padding, the window goes from 400 to 195. Then the approach utilize an max pooling with a stride of 3 to the 8 maps, thus maps with size 65×1 are exported. The same process with filtering size 11×1 and max pooling stride 2 are applied for the second convolution and pooling. Finally, 10 maps with size 14×1 are fully linked to 100 units, and fully linked to the output channel with 2 units: reading or not reading. Activation functions are rectified linear units (ReLU). We employ dropout with dropping rate 0.5 in each pooling and full connecting.

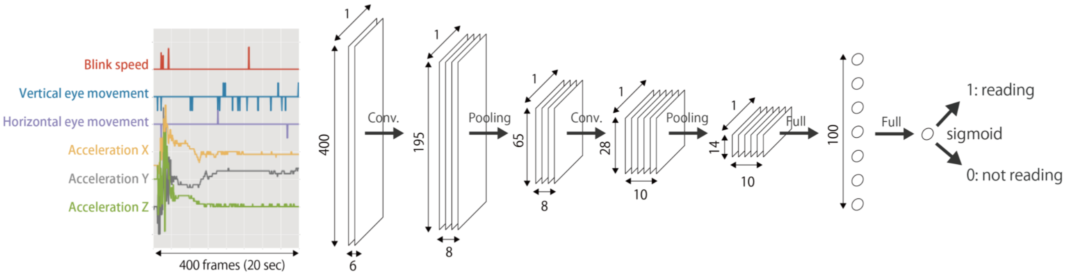


FIGURE 3.19: The CNN architecture for the reading detection

LSTM Based Reading Detection

By utilizing the advantage of the characteristics of time series data, we have also designed the network architecture including LSTM [76]. The input shape and parameters of the architectures are described in Figure 3.20. The parameters of the network were selected by random search. Since our purpose is to quantify reading activities and give feedback to a user later in the same way with physical activity tracker, a real time analysis is not necessarily required. Therefore Bidirectional LSTM is utilized to precede high accuracies.

After the both of classifications, we apply majority voting for 5 minutes of data (as we did in the SVM approach) to smooth the results.

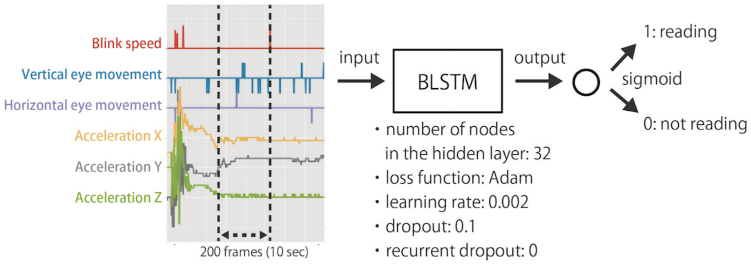


FIGURE 3.20: The LSTM architecture for the reading detection

3.3.2 Experimental Design

We evaluated the reading detection approaches on our long-term dataset with user-independent and user-dependent learning. This section presents procedures of the evaluation and classification results.

Data Recording

We asked 7 participants to record their habits using the following commercial sensors: JINS MEME, Fitbit Charge HR, Narrative Clip and Tobii eyeX (see Figure 3.15). Note that Fitbit Charge HR and Tobii eyeX are not used in this experiment. All of the participants were college students studying computer science, who worked on computers most of time. They used the tracking/recording devices during the day and charged them while they slept for more than two consecutive weeks. The dataset contains 22 hours of controlled reading, 427 hours of natural reading, 156 hours of social interactions and 375 hours of other activities.

We did not place any limit on the participants' activities. Therefore various types of reading activity are included in the dataset. Participants, for example, read texts on computers, smartphones, e-book readers, as well as paper. Browsing web pages and typing on a computer were also labeled as natural reading.

To record enough labeled reading activities, we also conducted a controlled experiment. We prepared 60 documents and asked the participants to read them from beginning to end. They read 15 English documents on paper, 15 English documents on a screen, 15 Japanese documents on paper, and 15 Japanese documents on a screen. Reading on paper was recorded with JINS MEME, and reading on a screen was recorded with JINS MEME and Tobii eyeX. We did not prohibit them from reading back during the recording, but most of them read documents continuously without vertical movements. Figure 3.21 represents the example of the difference of eye movements while natural reading and controlled reading.

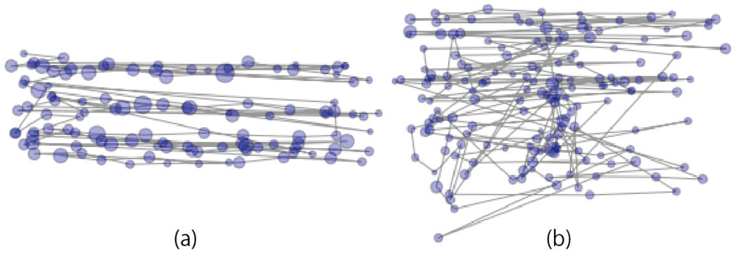


FIGURE 3.21: Eye gazes during a minute of (a) controlled reading and (b) natural reading. Data were collected by Tobii eyeX and classified into fixations (circles) and saccades (lines).

For the purpose of collecting ground truth, the participants added annotations to all data as shown in Figure 3.22. They were asked to apply one of the three labels (*reading*, *talking*, and *other activities*) to every 1 minute of data from 0:00 to 23:59. To help with the labeling tasks, we provided each participant with a Narrative Clip [6], a small life-logging camera which can be clipped to one's clothing. Narrative Clip takes a picture every 30 seconds. Participants reviewed the pictures at the end of each day and manually labelled their activities. In order to reduce ambiguities of the labels among participants, we asked them to label activities if pertinent objects (e.g. book, display, person) appeared in more than two consecutive pictures (= one minute). They submitted the annotated pictures after removing some of them for

[6] http://getnarrative.com/

privacy reasons. The reason we asked them to label their activities at the end of each day instead of during the recording is to make the dataset *wild* as much as possible. Regularly asking participants to provide ground truth labels leads to a well annotated dataset but might change their regular behaviors.

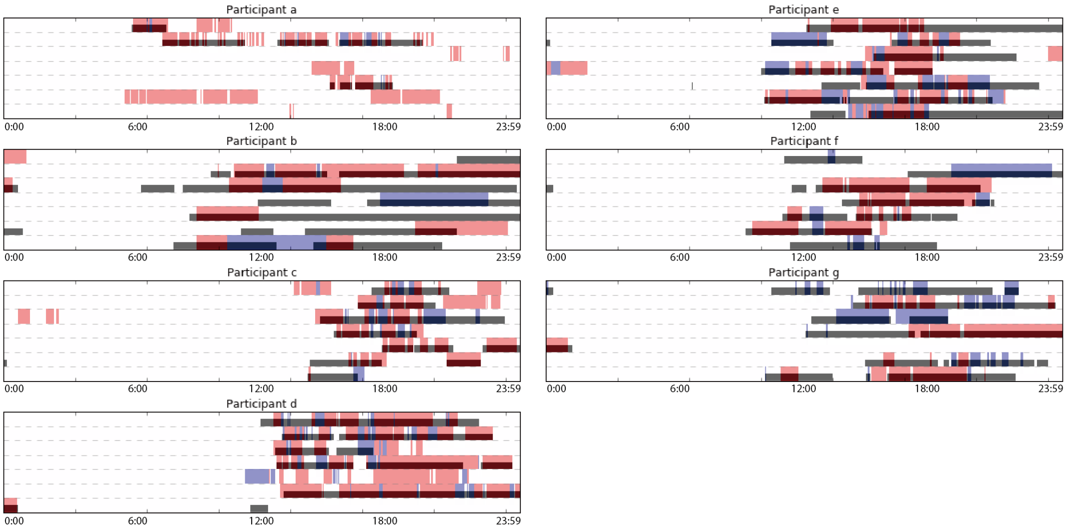


FIGURE 3.22: Activity labels of last seven days. Each row represents one day's activities. Periods filled in red are *reading*, blue are *talking*, and white are *other*. Periods participants were wearing JINS MEME are under lined in black.

Evaluation Design

For user-independent learning, training and testing data were separated by leave-one-participant-out cross validation. Samples of one participant were utilized as testing data, and samples of others were utilized as training data.

For user-dependent learning, training and testing data consist of samples from one participant. During our experiment, a new CSV file was created every time when a participant started recording. We shuffled the order of files and divided them to two groups equally. Samples in one groups were utilized as training data and the other were utilized as testing. The reason we employed this way is to prevent carelessly mixing training and testing samples. Applying cross validation with all samples is the easiest way. But it might lead to incorporation of very similar samples into training and test folds in the analysis of time series data [70].

The mean and standard deviation value of results were calculated over all 7 participants. Because the number of samples in each class is unbalanced, *class weight* functions implemented in machine learning frameworks (scikit-learn for SVM based and Keras with TensorFlow for CNN and LSTM based) were utilized during training the model.

3.3.3 Results and Discussion

Classification Performance

Table 3.6 shows results comparing the SVM, CNN, and LSTM based approaches. The SVM based approach is more accurate than other two approaches to detect controlled reading. Although the differences are small, deep learning approaches performed better to detect natural reading.

TABLE 3.6: Means and standard deviations of classification accuracies over seven participants (controlled/natural reading vs. not reading)

	controlled reading detection		natural reading detection	
	user-independent	user-dependent	user-independent	user-dependent
SVM	**80.7±8.0 %**	**92.2±7.2 %**	68.5±7.2 %	73.1±5.3 %
CNN	66.2±20.6 %	80.2±12.3 %	**69.6±7.1 %**	70.0±5.4 %
LSTM	74.3±17.5 %	90.4±5.8 %	67.1±10.1 %	**73.8±6.0 %**

Confusion matrices of the natural reading vs. not reading classification on the user-independent approach are shown in Figure 3.23. For most of the participants, except for Participant *c*, the results show high precision (true positives divided by true positives + false positives) and low recall (true positives divided by true positives + false negatives). This result indicates that there are some reading activities that are still difficult to be detected by the three approaches.

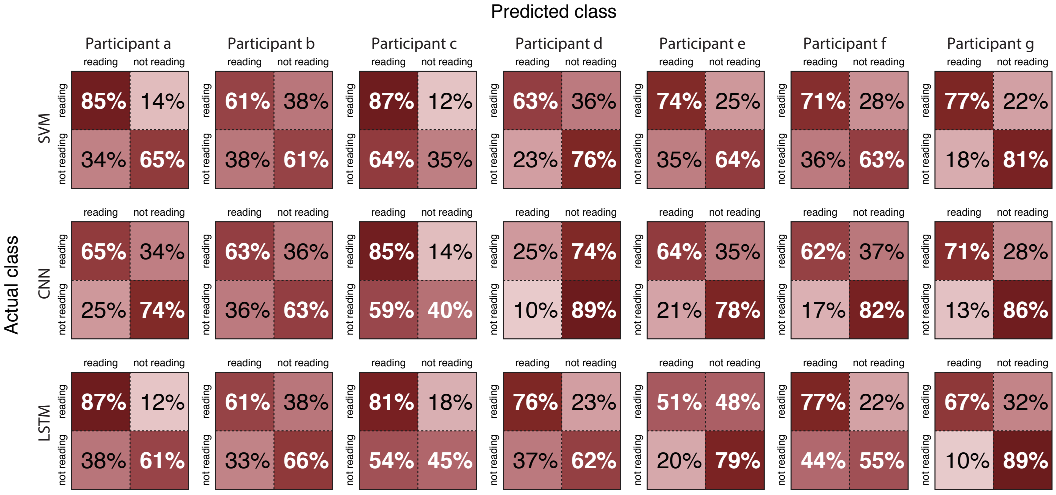


FIGURE 3.23: Confusion matrices of natural reading vs. not reading on the user-independent approaches over seven participants.

Observation of the Classified Samples

By reviewing the pictures taken by Narrative Clip, we identified some cases in which an activity can be misclassified. For example, while all the participants labeled Web browsing (see Figure 3.24 (a)) as *reading*, this activity was sometimes misclassified by the CNN and LSTM as not reading. This may have been caused by the combination of multiple factors, such as the web page layout that combines structured and non-structured texts (e.g., short text passages, banners, ads, etc.) as well as the actions that accompany web browsing, such as clicking on the embedded URIs. An interesting case of false positive occurred when one of the participants was watching a video (see Figure 3.24(b)). The participant himself labeled this activity as *not reading*, but our CNN and LSTM based user-independent approach classified it as reading. The participant was watching the video on *www.nicovideo.jp*, a popular video sharing service in Japan, which famously shows many floating subtitles in the videos. This has likely provided some irritations for the classifier.

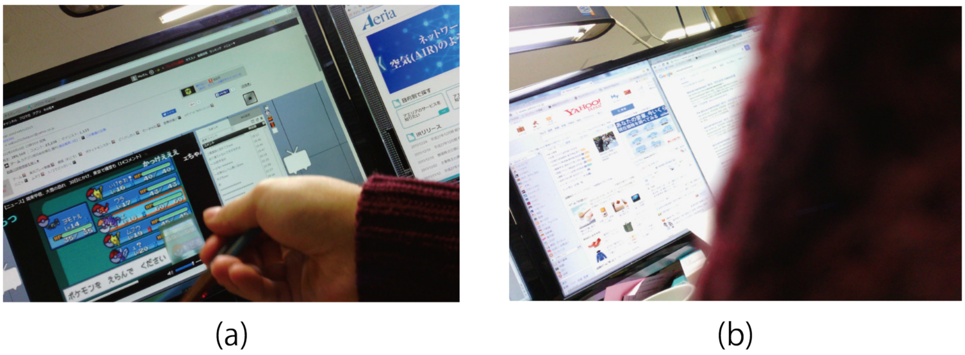

(a) (b)

FIGURE 3.24: Samples of errors in natural reading detection. (a) false negative: a user is browsing web pages. (b) false positive: a user is watching a video.

A major problem we found through this experiment is in labeling ground truth accurately for natural reading. Because the act of reading differs in kind (e.g. reading a paper book, browsing web pages, skimming texts, etc.), classifying activities into the simple two classes (reading vs. not reading) can be difficult even for humans.

3.3.4 Conclusion

In this work, we recorded natural activities in a daily life setting with unobtrusive, commercially available devices. By sacrificing accuracy to a degree, the amount data reached to more than 980 hours. The recorded data revealed that *natural reading* is a complex activity that includes many factors, as reading plain texts and browsing websites for instance involve different kinds of eye movements. We proposed three approaches to reading detection and found that the deep learning based approaches are superior to the SVM-based approach to detect natural reading activity. By investigating error samples, we have uncovered some of the challenges in detecting natural reading, including how to collect large-scale data with ground truth.

We continue exploring on data we gathered but did not use for the purpose of the present study in future work. Such data include recordings of the eye gaze while reading on a screen with Tobii eyeX and the heart rates while reading with Fitbit Charge HR. It should be interesting to see the relationship between JINS MEME's data and the data obtained by other sensors, and estimate the user's affective state such as the level of attention, concentration, and understanding of the contents.

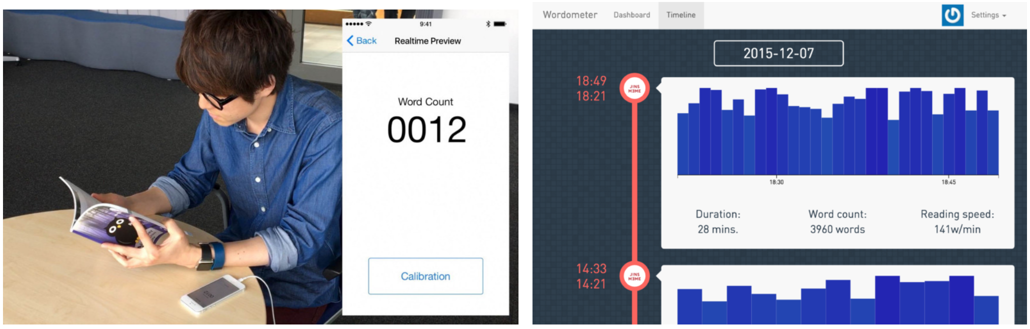
3.4 Wordometer 2.0: Estimating the Number of Read Words

Among several cognitive activities, reading is especially important activity because most of knowledge we have is from what we read. This section focuses on reading activity and proposes the method to quantify daily reading habits by using commercial electrooculography (EOG) glasses. The tracking result will be summarized and visualized as shown in Figure 3.25 (b). Because it is well known that increased reading volume is associated with numerous cognitive benefits including improved vocabulary skills [39], tracking and visualizing the amount of reading can help people improve their cognitive lifestyle.

The idea of estimating the number of read words by tracking eye movements has already been proposed by Kunze *et al.* as *Wordometer* [111]. They have introduced word counting algorithms based on mobile eye tracking glasses and medical EOG sensors. However, although the goal is to track daily reading habits, their setups are too bulky to be worn regularly (e.g., the devices are expensive; batteries are not enough long to cover a whole day; and cables prevent a user from moving naturally).

Compared to the their work, this research aims to quantify reading activities with technologies that are completely affordable. We utilize commercial EOG glasses because they are inexpensive, do not require significant user load, and have an long enough battery life for all-day use [89]. Additionally, there is no limitation to use the device from the viewpoint of privacy because it doesn't equip a camera. Features for the estimation are optimized for the device. We define our new word counting system which is designed for everyday use as *Wordometer 2.0*.

The contribution of this section is to show that: (1) Sensor signals from JINS MEME are good enough to detect specific eye movements during reading (forward- and backward-saccades). (2) The number of read words can be estimated by features from the forward- and backward-saccades.



(A) A user wearding JINS MEME (B) A web dashbord

FIGURE 3.25: Overviews of Wordometer 2.0

3.4.1 Approach

The word counting method consists of three processes: obtaining a user's eye movements, detecting forward- and backward-saccades, and estimating the number of words he/she read.

One reference electrode and two active electrodes are equipped on JINS MEME. In order to simplify the problem, we limit the subject of quantification to English text and utilize only a horizontal component. Figure 3.26 shows an overview of the EOG horizontal component in a one-minute recording that includes reading activity. Negative values represent eye movement right to left, and positive values represent left to right. Regular patterns of eye movement appear during reading activity because of line breaks.

Figure 3.27 shows outputs of the algorithm of detecting forward- and backward-saccades. Peak detection for forward-saccades is applied after applying a median filter to remove noises. Backward-saccades are detected if the sensor value is lower than a threshold. The threshold is calculated dynamically as the difference between the mean and variance of sensor values in a small window. The window size is one second, which was decided experimentally.

The number of words a user read is estimated by support vector regression. Four features are calculated for the regression: the total number of forward-saccades, the mean EOG signal value of forward-saccades, the total number of backward-saccades, and the mean EOG signal value of backward-saccades.

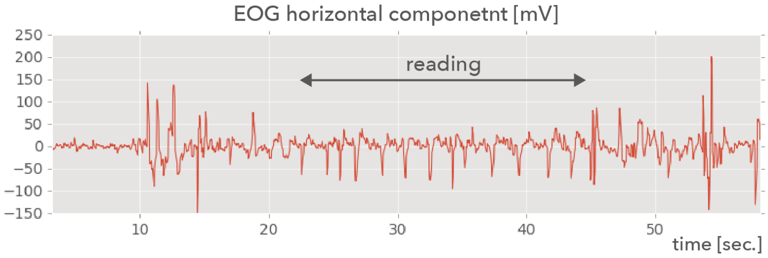


FIGURE 3.26: EOG sensor signal in one-minute recording including reading activity.

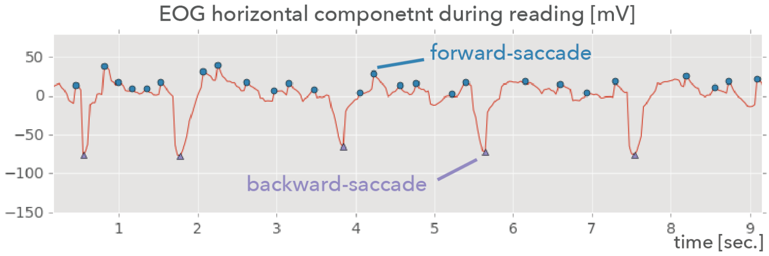


FIGURE 3.27: EOG sensor signal during reading activity. Circle and triangle markers are outputs of forward- and backward-saccades detection.

3.4.2 Experimental Design

Because it is hard to collect accurate ground truth in the wild experiment, the method was evaluated in the controlled experiment. We asked five participants to read English essays on an iPad wearing JINS MEME (see Figure 3.28). We developed an iPad reader app which highlights each paragraph of the essays on the basis of the scroll position. By using this hughlight function, participants informed the begging and ending time of reading each paragraph to the application. Every participant read 38 paragraphs, so the total amount of paragraphs in the dataset was 190 (minimum: 27 words; maximum: 120 words; average: 60 words in one paragraph). Because we used a prototype device, the sampling frequency of the EOG signal was 11 Hz.

Training and testing were done by leave-one-participant-out as a user-independent approach. We evaluated errors in the estimations with two measurements. One is an average of absolute error rates for each paragraph. This evaluation is valid for short-term recordings involving reading speed estimation. The other is an absolute error through all paragraphs, which is calculated as the total error of all recordings. This evaluation is valid for long-term recordings including a total count of words read in a day.

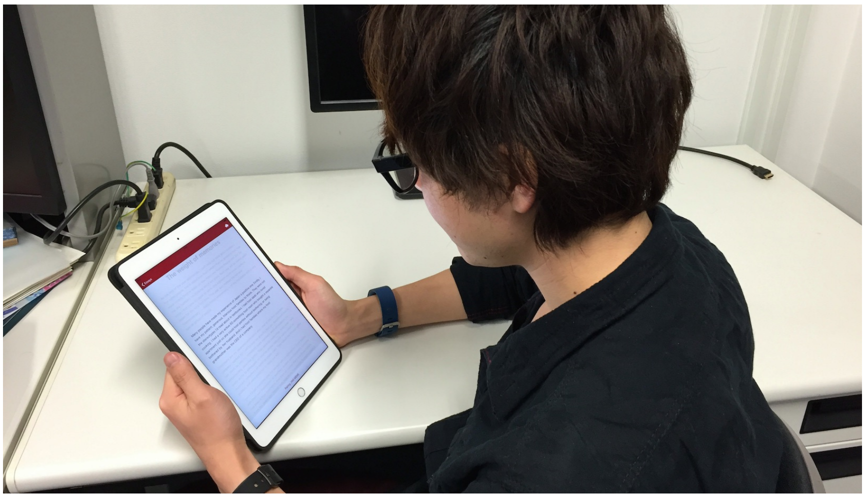


FIGURE 3.28: Experimental condition with JINS MEME and iPad

3.4.3 Results and Discussion

The estimation errors are shown in Table 3.7. An average error of five participants with user-independent training for each paragraph was 18 % and decreased to 11 % when extended with all 38 paragraphs. The discussion focuses on participant *b* and participant *c* to find the reasons of their inaccurate estimations.

TABLE 3.7: Word count estimation errors

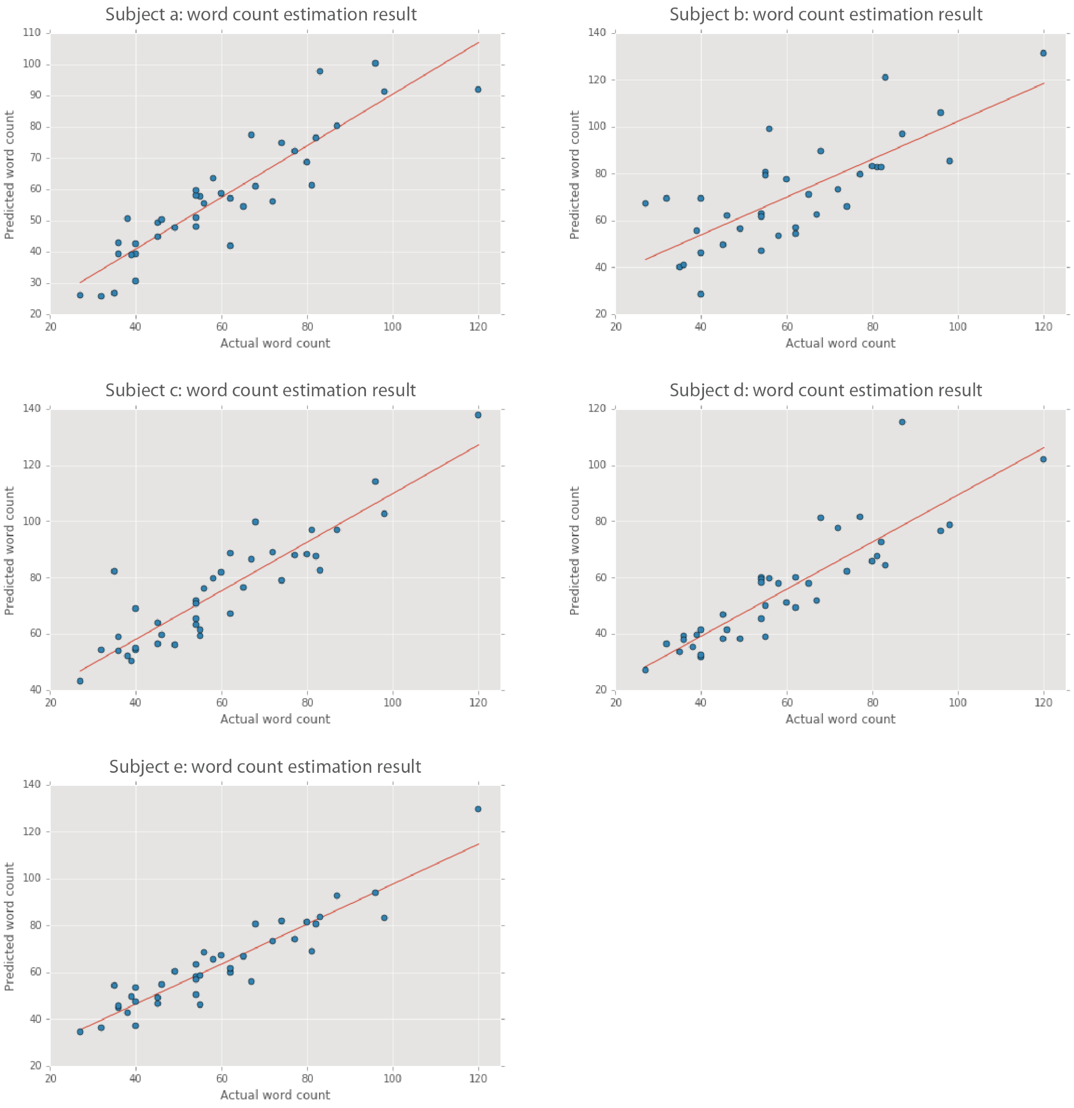
	User-independent		User-dependent	
Participant	Each paragraph	All paragraph	Each paragraph	All paragraph
a	17 %	11 %	14 %	2.2 %
b	24 %	9.5 %	25 %	6.4 %
c	24 %	18 %	15 %	2.4 %
d	15 %	7.5 %	15 %	1.8 %
e	15 %	11 %	12 %	2.4 %
Average	18 %	11 %	16 %	3.0 %

Figure 3.29 represents estimation results of each recordings with user-independent training. Plots of participant *b* is dispersed and this is the reason why estimation error for each paragraph is high. However a regression line of all plots looks so accurate. Participant *b*'s estimation error with combining all paragraph is low. It represents the truth that this system doesn't work well for the user whose reading style is not stable, but it works well with a scenario of long-term recording.

For participant *c*'s case, most of predicted word counts are higher than actual word counts and it means participant *c*'s eye movement during reading is unique. Figure 3.29 shows that estimation errors will decrease by user-dependent training. It means that estimation for the first use is not so accurate, but the accuracy will be increased with adapting to a user's behavior as he/she uses the system every day.

3.4.4 Conclusion

This section has described the method to quantify the amount of reading with JINS MEME. The estimation accuracies increase with extending recordings and could also be improved by applying user-dependent training. Because JINS MEME is designed for everyday use, we should follow the condition and create better estimating algorithm with long-term recordings.


Subject a: word count estimation result

Subject b: word count estimation result

Subject c: word count estimation result

Subject d: word count estimation result

Subject e: word count estimation result

FIGURE 3.29: Plot of estimations on user-dependent approach

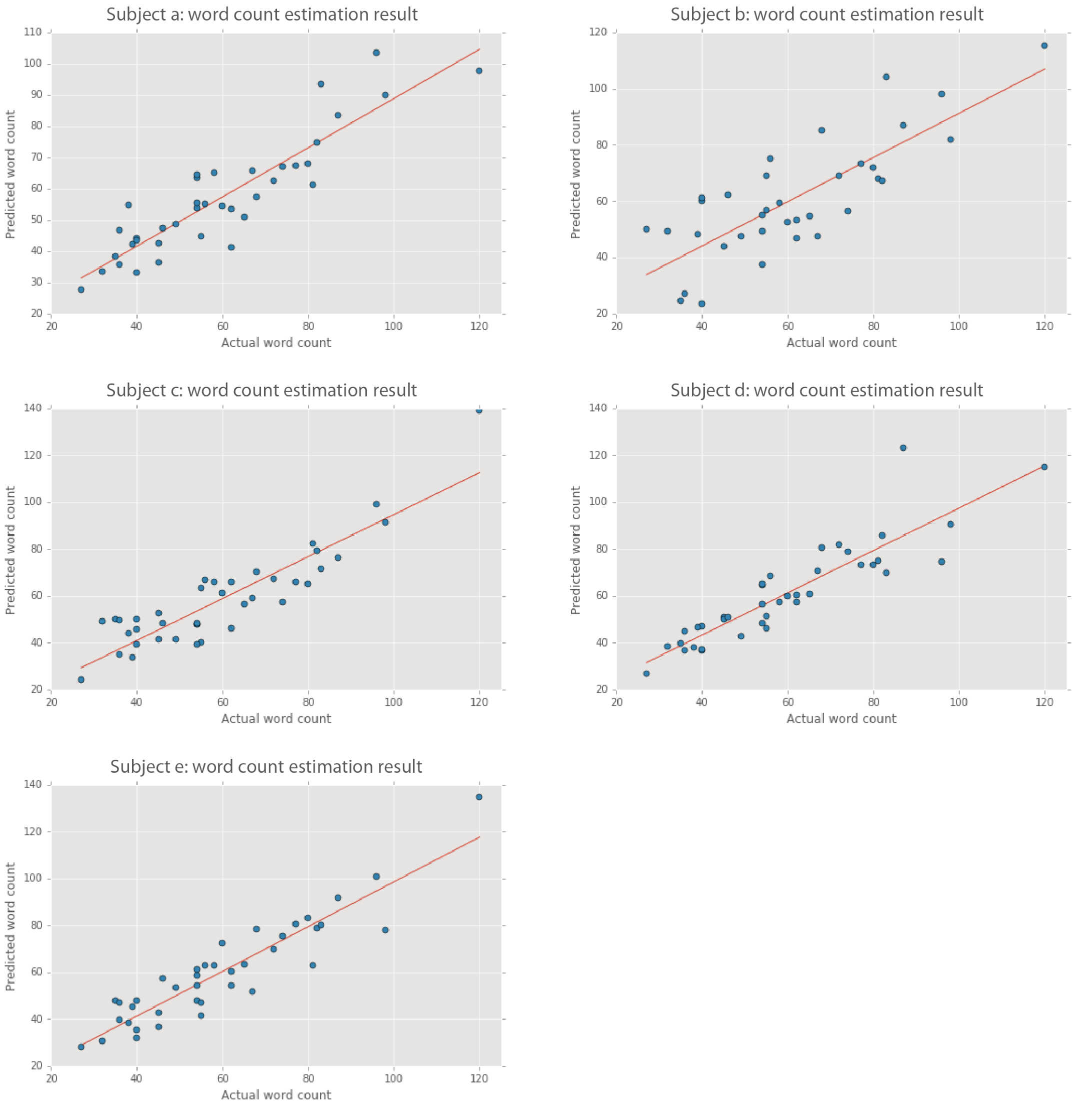


FIGURE 3.30: Plot of estimations on user-independent approach

3.5 Quick Activity Labeling Tool using Swipe Gestures

Collecting large data of daily activities with ground truth is necessary for better recognition. Eyewears which can be performed fast and do not disturb a user so much during the activities are perfect platform for ground truth labeling. This section proposes an interface for quick activity labeling on Google Glass with multi touch swipe gestures. The contributions of this section are: (1) Presenting an UI design concept targeting quick data entry during situations with a high cognitive load. (2) Evaluating the UI design in a prototype implementation.

The default interface of Google Glass works very well for many types of tasks. An exception is the selection of items from larger lists, in particular when voice interaction is not desired and the user is moving (e.g. walking through a shop or an office). The default interface involves *hitting* the right item through a controlled *analogue* motion (the selected item depends on how fast/far the finger moves) on the touch pad. Such analogue motions are difficult to perform exactly when the user is moving and require a degree of concentration. As an alternative we propose a *digital* interface that codes the item selection through hierarchical combinations of multi touch gestures (number of fingers, number of swipes, direction of swipe) supported by appropriate representation of the choice on the screen. The advantage of this approach can be explained by Fitt's law. It makes one dimension of the target size infinite. The user does not need to scroll to the *right* card, but makes his selection over the number of fingers and swipe direction.

There is some related work on wearable interaction focusing on gestures and multi touch [129]. Harrison *et al.* show how to use a depth sensor to enable multi touch interaction on everyday surfaces [72]. Complementary to our work, Thomas *et al.* evaluate the usefulness of touch pad interactions for wearable computers [171].

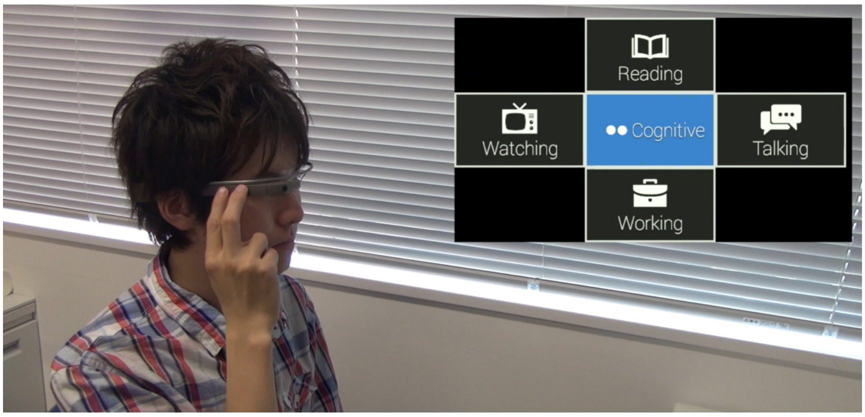


FIGURE 3.31: A user labeling his activity by using the proposed interface

3.5.1 Approach

Given these considerations, we came up with the following requirements. The user should not necessarily need to look at the screen to select an activity. The selection process should be done with a quick gesture only, as fast as possible. We emphasis that un-obtrusiveness and speed are preferred over accuracy regarding the selection. Lacking accuracy can be filtered out by mere numbers. However, we believe at once a wearable application is just a small hindrance, people will avoid using it.

Based on this discussion we designed the following interface as shown in Figure 3.32 (and using in Figure 3.31). The upper screen displays the categorical overview with hints how to access them (one dot = one finger, ...). As soon as one, two or three fingers touch the touch pad on the right side of Google Glass the corresponding label selection matrix appears. To select a label the user keeps his fingers on the touch pad and swipes to the front, back, up or down. Therefore a user can select out of 3 x 4 customizable activities using swipe gestures. More are currently not feasible due to the size of the Google Glass touch pad.

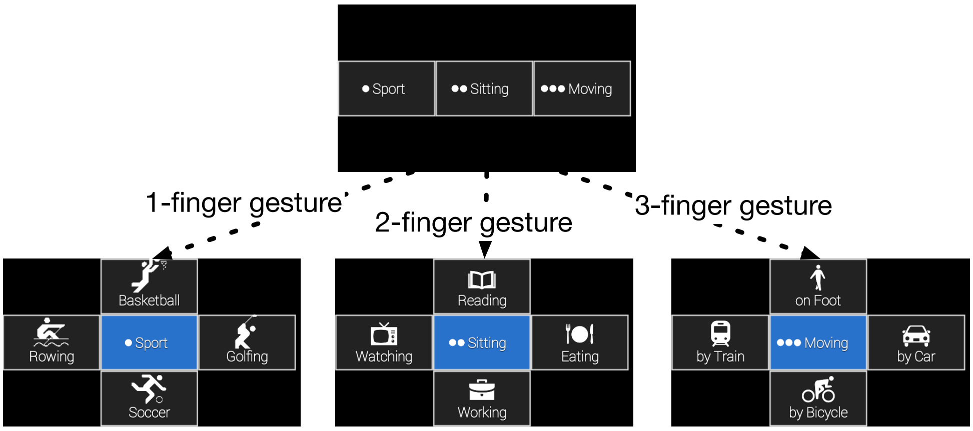


FIGURE 3.32: Google Glass swipe gesture input method screenshots.

3.5.2 Experimental Design

We asked 11 participants to use both input methods subsequently. 7 of 11 participants are employed in the area of computer science and 4 were employed in other areas and were not familiar with wearables. We studied the hierarchical (Google Glass card style) input method as shown in Figure 3.33 and ours referred to the multi touch swipe gesture input.

The experimental evaluation measures the time needed for an entry and the accuracy of a selected label with the given input-method. We first introduced the participant to Google Glass with focus on touch pad interactions. We then always presented the classic input method to the current participant first. The participant was only told how to use the Google Glass touch pad but was not familiarized with the set of labels (elements to select) before. We considered the main target of Google Glass: Performing quick look ups and input while underway, especially while walking: The participant was instructed to continuously walk during the experiment to achieve a natural environment with distractions i.e. watching their steps to avoid obstacles (the experiment area covered 10 by 5 meters). While the participant was in motion the experiment observer communicated one label after another to be inserted with the classic method. We repeated this 40 times with random labels. The experiment observer was automatically notified by the application when the input was accomplished and then communicated an automatically randomly selected label. After the first run the swipe-gesture user interface application was started with the same experimental conditions than on the run before. The labels were again communicated fully randomized one after another for 40 iterations.

During our experiment we ran a server application (1) generating a new label on demand, storing the label and timestamp and (2) logging the receiving label together with the timestamp of the selected label and the type of input (classic or swipe method) on Google Glass. The Google Glass applications we created established a TCP connection to our server and transmitted each label input. The server script allowed to retrieve a randomly chosen label which could then be communicated by voice to the participant. The observer monitored the status on a laptop connected to the server and was informed automatically when the current input was finished.

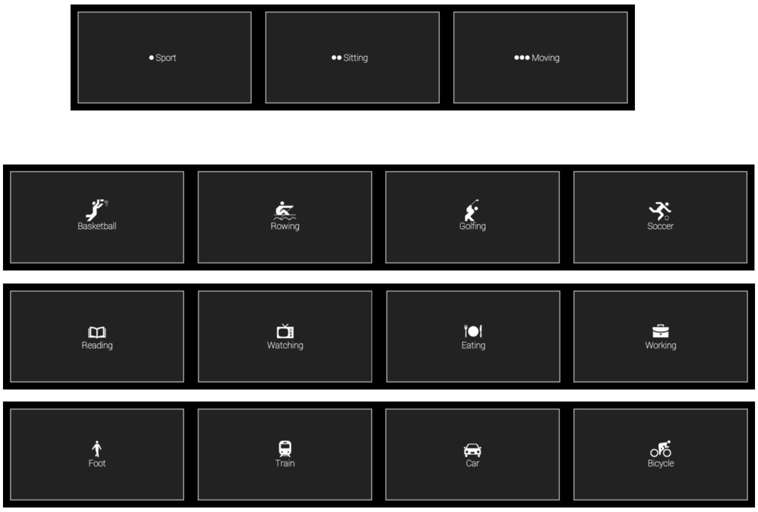


FIGURE 3.33: Screenshots of default hierarchical input method on Google Glass

3.5.3 Results and Discussion

Table 3.8 shows the experimental results. Figure 3.34 displays the reaction time vs the correctness of the input. For all participants the reaction time is decreased significantly ($p < 0.05$). The reaction time was reduced by 1.7 seconds on average. But, the mean accuracy decreases to 75 % (leaving out the worst 78 %) for the swipe interface versus 95 % on average for the classic interface. Figure 3.35 shows the change of annotation speed with training The finding from this result is that users could perform quicker and quicker with the proposed system by iterated practices.

TABLE 3.8: Experimental data set and results overview

Participant id	1	2	3	4	5	6	7	8	9	10	11
Classic (card based) input method:											
Average duration of response (sec)	5.7	8.2	7.7	6.1	7.7	5.7	5.2	5.7	5.4	8.0	5.3
Annotation Accuracy	97 %	95 %	98 %	100 %	92 %	93 %	95 %	95 %	98 %	100 %	96 %
Duration slope	+1.6	-142.8	-66.9	-94.7	-137.4	-56.7	-6.7	-124.8	-50.9	-89.7	-16.8
Swipe gestures input method:											
Average duration of response (sec)	3.9	4.7	5.4	5.4	5.4	5.5	4.8	5.0	3.8	5.4	4.1
Annotation Accuracy	82 %	64 %	68 %	42 %	95 %	85 %	86 %	93 %	67 %	64 %	75 %
Duration slope	-3.4	-31.1	+15.5	+6.2	-27.5	-47.7	+23.8	-47.6	-13.0	-29.5	-82.3

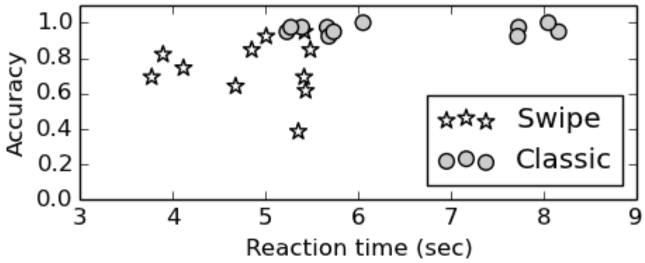


FIGURE 3.34: The diagram displays the reaction time vs the correctness of the input. Each symbol depicts one experiment run (40 samples).

3.5.4 Conclusion

This section presented a multi touch swipe interface for activity labeling on Google Glass. It is significantly faster compared to the standard interface. The speed increases in encouraging although we need to improve the accuracy. The current task *walking* might not so complex and does not require a lot of attention. We should evaluate our interface during more demanding tasks. We assume that the time decrease in selection will outweigh the accuracy decrease for challenging activities. We could observe a potential learning effect, as we are using the interface now for a longer time, it seems the accuracy increases and it's possible to select tasks without looking on the HMD. However, this needs to be evaluated in a larger study.

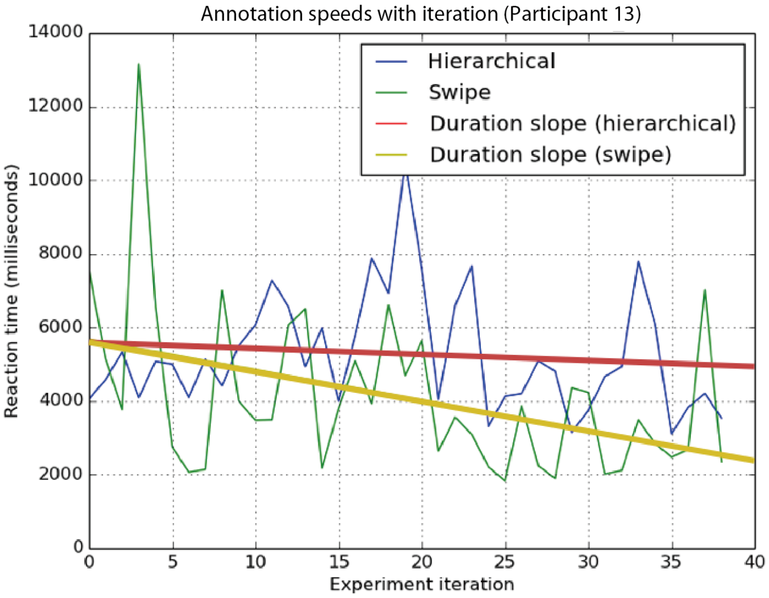


FIGURE 3.35: The change of annotation speeds with iteration

Chapter 4
Affective State Recognition

Quantified learning – sensing learning behaviors for giving effective feedback based on the contexts to each learner – has a high potential in the era of digitalized education [43]. This chapter presents qualifications on several media including textbook, video lecture, news article, multiple-choice questions, vocabulary spelling tests.

The starting point of this research field is comprehension recognition because improving the performance of learning should be a clear objective among all students. Section 4.1 presents comprehension recognition methods on a textbook: one of the most major learning materials [81, 82]. Section 4.2 presents investigation on a video lecture: a modern material [134].

Section 4.3 presents results of a pilot study finding correlations between sensor signals and affective states collected by a subjective survey after reading a textbook in Physics [86]. In order to create estimation models, it is essential to collect large data with distributed conditions. Therefore we selected an activity of reading newspaper articles to collect various reading behaviors. In Section 4.4, reading behaviors are classified into four class of levels by using an eye tracker [95] and a physiological sensing wristband [96].

Sections 4.5 and 4.6 present self-confidence estimation on multiple-choice questions [92] and handwriting vocabulary spelling tests [124]. Self-confidence is one of the most important affective states because if there is a gap between their confidence and level of understanding, they lose chance to re-cap a subject correctly. On the other hand, a high self-confidence in a learning subject causes positive learning feedback loops.

The following sections in this chapter are based on collaborative work with students. Section 4.2: Yuya Ohbayashi, Section 4.3: Apurba Roy, Section 4.4: Soumy Jacob, Section 4.5: Kent Yamada, and Section 4.6: Takanori Maruichi.

4.1 Comprehension Recognition on a Textbook

Reading a textbook is an important way to obtain new knowledge. We investigate students' reading behavior on a textbook in order to find specific patterns which are related to the context including situations and comprehensions. For this purpose, we prepared a document on "Basic Phenomena in Acoustics" including both, text and related exercises. In addition, we recorded their eye gaze during the period when only the text page is shown and subsequently when both, texts and exercises, are shown on a display.

There is a whole bunch of research investigating efficient visualizations and representations to improve students' skills of understanding and solving in Physics Education Research [127, 179]. However, most of them obtain students' insights from only answering sheets afterwards and do not care about their learning process while reading texts and solving tasks.

We follow preliminary research from Mozaffari *et al.* [130], where students' eye gaze is recorded while solving tasks of physics by using a remote eye tracker mounted to a tablet computer. The authors have revealed that students prefer different representations (vector, table and diagram) depending on their skill level. We follow their basic idea. But compared to their work, we are interested in students' natural reading behavior on a textbook. We do not optimize the text and tasks so much for the recording.

The contributions of this section are two-fold. (1) We find rough relations between levels of students' expertise and their reading behavior on a textbook. (2) We recorded the data from 6-grade students at school. It realized some limitations to using eye-tracking devices in a realistic educational scenario.

4.1.1 Approach

We believe that extracting the part where students pay attention is the first step to investigate their reading behavior. We propose a method to extract attention by using an eye tracking device. The method consists of three steps.

Mapping Eye Gaze Coordinates on a Document

We utilize mobile eye tracking glasses and record a student's reading behavior. Because the output of the device are coordinates of eye gaze on a scene camera, we need to map them to the document under consideration. As shown in Figure 4.1, we detect the position of the document in a frame of the camera by using SIFT features [120], and calculate a homograph to map the gaze point to the document.

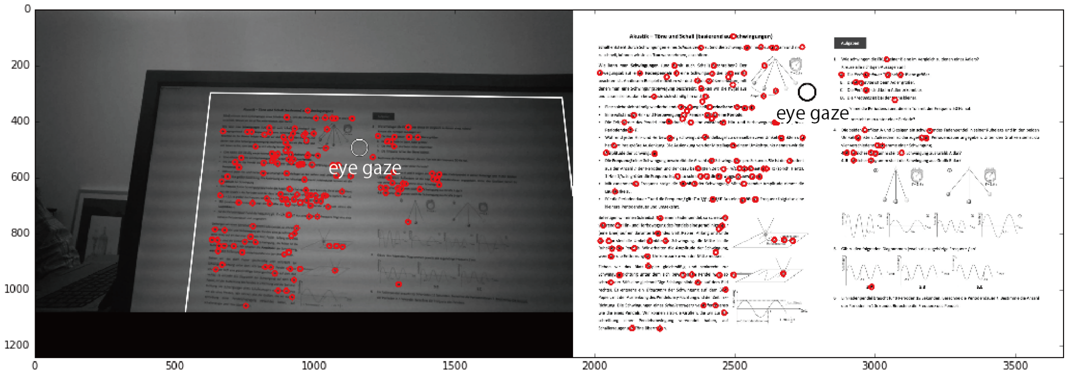
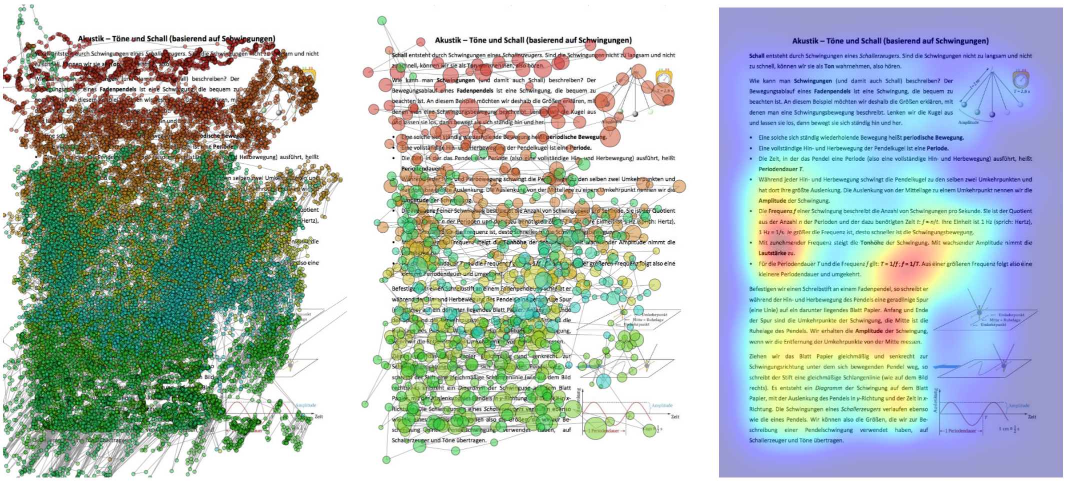


FIGURE 4.1: One of the results of gaze mappings. Red circles of both scene and document image represents feature points of SIFT. The documents is extracted as white rectangle on a scene image. The gaze point on document is estimated as a black circle.

Detecting Fixations

The raw gaze data on the document is classified into fixations and saccades. Fixations appear, when the gaze pause in certain position – normally lasting between 200 and 400 ms. Saccades are the jumps of the gaze between two fixations taking 10-20 ms. We apply the fixation-saccade detection algorithm proposed by Buscher *et al.* [24]. Figure 4.2 shows input and output. The radius of each circle corresponds to the fixation, and the line between two circles represents a saccade. Noises and drifts on raw data are also filtered in this step.



(A) Raw gaze (B) Fixations and saccades (C) Fixation duration heatmap

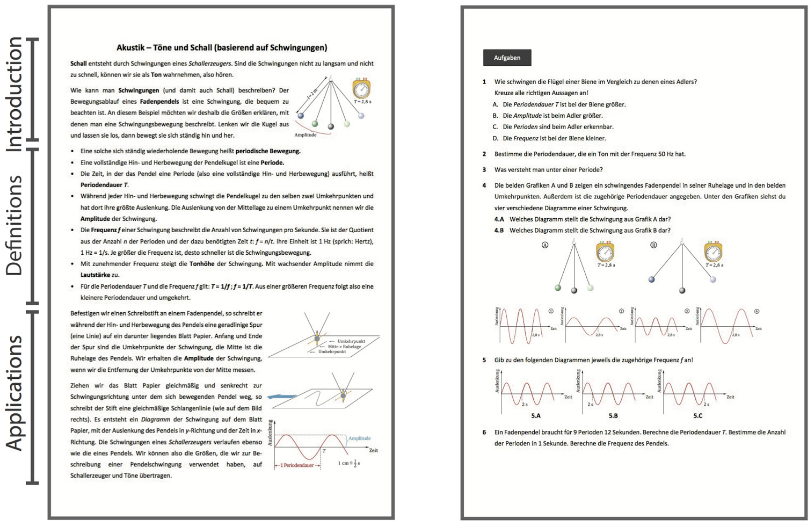
FIGURE 4.2: Eye gaze on the document while a student read a textbook. Colors in (a) and (b) represent the order of eye gaze (red to blue)

Calculating Features for Each Area of Interest

We divide a text beforehand based on the roll (e.g., the introduction, definitions, applications on the document shown in Figure 4.3.) We focus on the period of time a student needs to reads the content to obtain knowledge. Thus, for each area a sum of fixation durations is calculated, which is divided by the size of area to be normalized.

AOI Based Attention Extraction

We believe that extracting the part where students pay attention is the first step to investigate their reading behavior. We divide a text beforehand based on the roll (e.g., the introduction, definitions, applications on the document shown in Figure 4.3.), then focus on the period of time a student needs to reads the content to obtain knowledge. Thus, for each area of interest (AOI) a sum of fixation durations is calculated, which is divided by the size of area to be normalized.


Introduction
Definitions
Applications
l = 2 m
T = 2,8 s
Amplitude
Umkehrpunkt
Mitte = Ruhelage
Umkehrpunkt
Auslenkung
Amplitude
T
Zeit
1 Periodendauer
1 cm ≙ ½ s
A
T = 2,8 s
B
T = 2,8 s
1
2
3
4
Auslenkung
Zeit
2,8 s
Auslenkung
Zeit
2 s
5.A
5.B
5.C

Akustik – Töne und Schall (basierend auf Schwingungen)

Schall entsteht durch Schwingungen eines *Schallerzeugers*. Sind die Schwingungen nicht zu langsam und nicht zu schnell, können wir sie als **Ton** wahrnehmen, also hören.

Wie kann man **Schwingungen** (und damit auch Schall) beschreiben? Der Bewegungsablauf eines **Fadenpendels** ist eine Schwingung, die bequem zu beachten ist. An diesem Beispiel möchten wir deshalb die Größen erklären, mit denen man eine Schwingungsbewegung beschreibt. Lenken wir die Kugel aus und lassen sie los, dann bewegt sie sich ständig hin und her.

- Eine solche sich ständig wiederholende Bewegung heißt **periodische Bewegung.**
- Eine vollständige Hin- und Herbewegung der Pendelkugel ist eine **Periode.**
- Die Zeit, in der das Pendel eine Periode (also eine vollständige Hin- und Herbewegung) ausführt, heißt **Periodendauer *T*.**
- Während jeder Hin- und Herbewegung schwingt die Pendelkugel zu den selben zwei Umkehrpunkten und hat dort ihre größte Auslenkung. Die Auslenkung von der Mittellage zu einem Umkehrpunkt nennen wir die **Amplitude** der Schwingung.
- Die **Frequenz *f*** einer Schwingung beschreibt die Anzahl von Schwingungen pro Sekunde. Sie ist der Quotient aus der Anzahl *n* der Perioden und der dazu benötigten Zeit *t*: $f = n/t$. Ihre Einheit ist 1 Hz (sprich: Hertz), 1 Hz = 1/s. Je größer die Frequenz ist, desto schneller ist die Schwingungsbewegung.
- Mit zunehmender Frequenz steigt die **Tonhöhe** der Schwingung. Mit wachsender Amplitude nimmt die **Lautstärke** zu.
- Für die Periodendauer *T* und die Frequenz *f* gilt: $T = 1/f$; $f = 1/T$. Aus einer größeren Frequenz folgt also eine kleinere Periodendauer und umgekehrt.

Befestigen wir einen Schreibstift an einem Fadenpendel, so schreibt er während der Hin- und Herbewegung des Pendels eine geradlinige Spur (eine Linie) auf ein darunter liegendes Blatt Papier. Anfang und Ende der Spur sind die Umkehrpunkte der Schwingung, die Mitte ist die Ruhelage des Pendels. Wir erhalten die **Amplitude** der Schwingung, wenn wir die Entfernung der Umkehrpunkte von der Mitte messen.

Ziehen wir das Blatt Papier gleichmäßig und senkrecht zur Schwingungsrichtung unter dem sich bewegenden Pendel weg, so schreibt der Stift eine gleichmäßige Schlangenlinie (wie auf dem Bild rechts). Es entsteht ein *Diagramm* der Schwingung auf dem Blatt Papier, mit der Auslenkung des Pendels in *y*-Richtung und der Zeit in *x*-Richtung. Die Schwingungen eines *Schallerzeugers* verlaufen ebenso wie die eines Pendels. Wir können also die Größen, die wir zur Beschreibung einer Pendelschwingung verwendet haben, auf Schallerzeuger und Töne übertragen.

(a) text

Aufgaben

1 Wie schwingen die Flügel einer Biene im Vergleich zu denen eines Adlers? Kreuze alle richtigen Aussagen an!

A. Die *Periodendauer T* ist bei der Biene größer.
B. Die *Amplitude* ist beim Adler größer.
C. Die *Perioden* sind beim Adler erkennbar.
D. Die *Frequenz* ist bei der Biene kleiner.

2 Bestimme die Periodendauer, die ein Ton mit der Frequenz 50 Hz hat.

3 Was versteht man unter einer Periode?

4 Die beiden Grafiken A und B zeigen ein schwingendes Fadenpendel in seiner Ruhelage und in den beiden Umkehrpunkten. Außerdem ist die zugehörige Periodendauer angegeben. Unter den Grafiken siehst du vier verschiedene Diagramme einer Schwingung.

4.A Welches Diagramm stellt die Schwingung aus Grafik A dar?
4.B Welches Diagramm stellt die Schwingung aus Grafik B dar?

5 Gib zu den folgenden Diagrammen jeweils die zugehörige Frequenz *f* an!

6 Ein Fadenpendel braucht für 9 Perioden 12 Sekunden. Berechne die Periodendauer *T*. Bestimme die Anzahl der Perioden in 1 Sekunde. Berechne die Frequenz des Pendels.

(b) tasks

FIGURE 4.3: A document with text and tasks in physics. These two figures are in one page on a display (text on the left and tasks on the right) during the experiment.

AOI Based Expertise Prediction

We apply a support vector machine (SVM) in order to predict students' expertise. According to AOI based fixation duration described as above, each duration in AOI is calculated as feature of the training and the testing. From the document in Figure 4.3, for example, three features (durations on the introduction, definition, and application) are used. Note that since the this method requires a student's reading behavior from the beginning to the end of a document, it can only be applied as an offline analysis.

Subsequence Based Expertise Prediction

On the other hand, an online analysis is required in order to change the content dynamically while reading. Therefore, we also investigate whether a subsequence (e.g., 1 minute of reading) is enough useful to predict students' expertise In this approach, we calculate four features (mean and standard deviation of fixation durations and saccade lengths) in a subsequence and apply SVM based classification. These features are selected according to some work investigating eye movements as reflections of comprehension processes [135, 145].

4.1.2 Experimental Design

We asked 8 participants to wear eye-tracking glasses, to read a physics textbook and to solve respective exercises. The participants were 6-grade students at a German high school (around 12 years old). The document we prepared is shown in Figure 4.3. It consists of three parts: the introduction, itemized definitions, and applications. To analyze the students' natural behavior while acquiring knowledge, we selected content that they had not yet learned in class.

Only an explanation of about the content (the left page in Figure 4.3) was displayed at first. After they understood the content, they could make tasks appear by pressing the space-key on a keyboard. They could go back to read the content to help them in their solving tasks. In this section, we define these two steps as *reading* and *solving*.

Figure 4.4 shows the overview of the experiment. To evaluate whether our proposed method works with different eye tracking devices, two types of eye tracking glasses were used during the experiment. We used *Tobii Pro Glasses 2* with five participants (*a, b, d, e, f*). The glasses record eye gaze at a sampling frequency 100 Hz and a scene video at 25 Hz. We applied one-point calibration with a marker before starting each recording. The data of the other three participants (*c, g, h*) were recorded with *SMI Eye Tracking Glasses 2*. The glasses record eye gaze at a sampling frequency 60 Hz and a scene video at 30 Hz. We apply three-point calibration with this device.

4.1.3 Results and Discussion

Attention Extraction

Table 4.1 shows the percentages of time students paid attentions for the introduction, definitions, and the applications on the document. The sum of these three values is

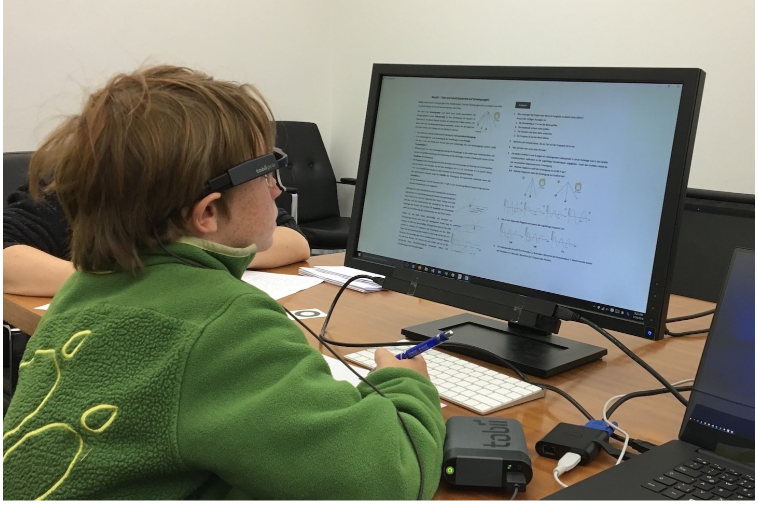


FIGURE 4.4: An overview of the experiment. A participant is solving questions on a display with wearing *Tobii Pro Glasses 2*.

100 %. We calculated the percentage depending on the each situation while reading a text and solving tasks. Note that the data in Table 4.1 is sorted by the number of correct answers. We categorized the eight participants to three comprehension levels according to the score: novice (the score is four or less), intermediate (the score is five), and expert (the score is six or more).

TABLE 4.1: Percentages of time students paid attentions

Participants	Score	Expertise	While reading text [%]			While solving tasks [%]		
	(out of 14)		Intro.	Def.	Appl.	Intro.	Def.	Appl.
a	3	Novice	14	49	37	13	59	28
b	4	Novice	17	43	40	17	48	35
c	5	Intermediate	7	51	42	4	44	52
d	5	Intermediate	31	41	28	21	49	30
e	5	Intermediate	23	37	40	27	40	33
f	6	Expert	16	47	37	12	60	28
g	7	Expert	34	50	16	25	56	19
h	7	Expert	28	64	8	22	70	8

By calculating mean values for each comprehension level as shown in Figure 4.5, it has become obvious that students with high-level comprehension do not pay attention to the applications part while both reading and solving tasks compared to other levels. They understand that the applications part is useful for understanding the content, yet there is not much information that can be used as hints for solving tasks. They preferred to read definitions part because there are direct hints (principles, formulas, etc.). Intermediates and novices spend much time to paying attention to the application part while both reading and solving.

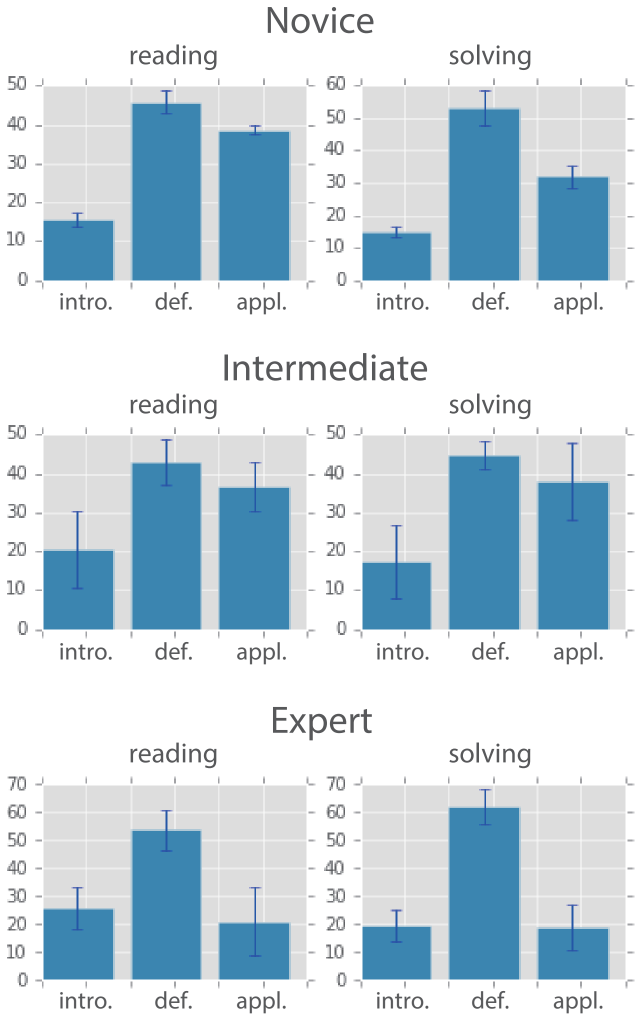


FIGURE 4.5: Histograms of the time students paid attentions [%].
Error bars represent standard deviations.

While percentages of attention in the reading and solving phase are similar for all levels, their reading behavior is deferent for each situation. Figure 4.6 represents average fixation-based heat maps from novice, intermediate, and expert while reading and solving. High-attention part during reading is almost same for all three skill levels, while the participants spend much time to read left side of the definitions part. Note that novices and intermediates look graphics carefully while solving, and this is one of the reasons their attention for this application is higher than for experts.

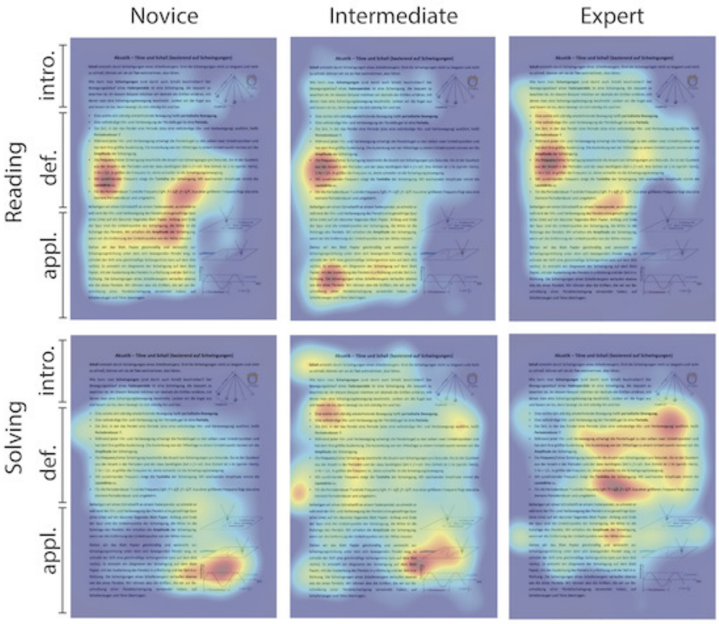
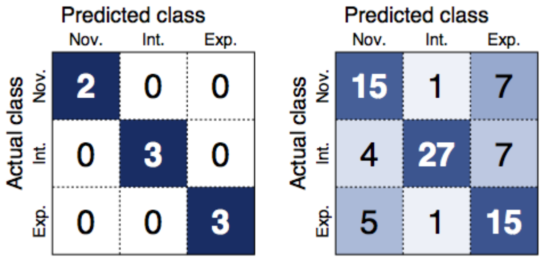


FIGURE 4.6: Fixation duration based heat maps while reading and solving

Expertise Predictions

By using the categories (novice, intermediate, and expert) as ground truth, we estimated participants' expertise. Figure 4.7 represents confusion matrices of the estimation results. The AOI based approach succeeded to estimate all expertise of participants. The estimation accuracy of the subsequences based approach was 70 %.



Actual class \ Predicted class	Nov.	Int.	Exp.
Nov.	2	0	0
Int.	0	3	0
Exp.	0	0	3

(A) AOI based

Actual class \ Predicted class	Nov.	Int.	Exp.
Nov.	15	1	7
Int.	4	27	7
Exp.	5	1	15

(B) subsequence based

FIGURE 4.7: Confusion matrices of the expertise prediction

Figure 4.8 shows all participants' feature plot in subsequences based approach. The higher the participant's expertise is, the larger mean saccade length is measured. Novice students read a textbook with large fixation duration and small saccade length, and intermediate students read with small fixation duration and large saccade length. It cleared that novice students read a textbook slowly with small steps. The distribution of data plot from experts is larger than others. They sometimes skip their eyes on the text, focus on the content they are interested in.

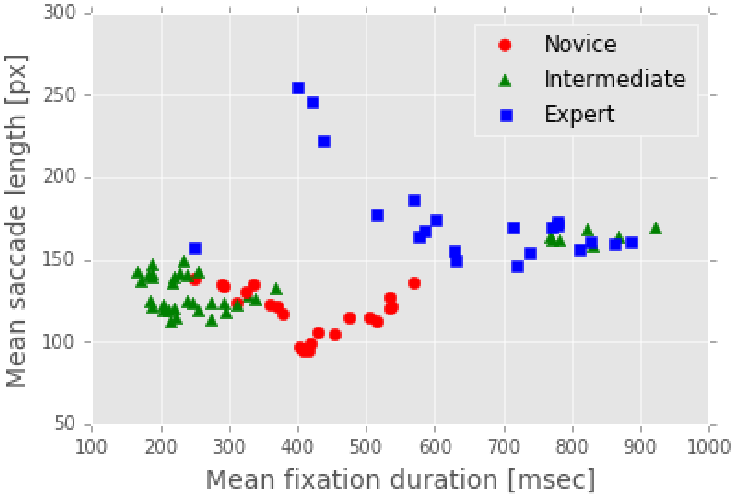


FIGURE 4.8: Feature representation of all participants' data in subsequences based approach. Each dot represents a data segment of one minute.

Sensing Modalities

There is another option to use remote eye tracker (device which is attached to a display) instead of mobile eye tracking glasses. Compared to a remote eye tracker, mobile eye tracking glasses have the advantage that the device can record eye gaze not only on display but everything on a scene camera. However, the big issue during this experiment was that students touched glasses unconsciously, and lose an accuracy of eye tracking adjusted during calibration. We actually recorded data with help of 10 students, but data from two of them are too noisy to analyze. Therefore using remote eye tracker may be better than mobile glasses for young participants.

Compared to a remote eye tracker, mobile eye tracking glasses have the advantage that the device can record eye gaze not only on display but everything on a scene camera. However the big issue during this experiment was that students sometimes touch glasses unconsciously, and lose an accuracy of eye tracking adjusted during calibration. Therefore using remote eye tracker might be better than mobile glasses for young participants.

4.1.4 Conclusion

In this section, we present an initial method to extract students' attention by using gaze data. By applying the approach to activities including reading a text and solving tasks, it is revealed that reading behavior is related to students' comprehension. Expert students, for example, tend to pay attention on definition part to understand the content. We also predicted students' expertise (ground truth was calculated by the score of tasks) by two approaches. One is attentions on AOI based, and the other is features from gaze subsequence based prediction. The former one works better than the later one, but it requires the recording of reading from beginning to end. We found that features from sub pattern of gaze data in one minute can enough classify students' expertise into three classes with 70 % accuracy.

There are roughly divided two feature work. One is a more detailed analysis of cognitive students while reading. Eye tracking can measure students' attention, but it is still difficult to distinguish it is because of their interest or confusing. To use other sensing modalities (e.g., heart rate, face temperature) may work to recognize them. The other is to apply dynamic changing. We are going to implement some dynamic changing as described in the basic concept section of this section, and investigate whether the changing helps students learning ability significantly.

4.2 Comprehension Recognition on a Video Lecture

Recently, video lecture based e-learning has become popular [176]. Compared to traditional human-to-human lectures, there are many advantages in this lecture style. For instance, students can select when, where, which subject, and by whom to learn by themselves. They can also pause and replay lecture videos anytime on the basis of the level of understanding. However, as a disadvantage, it is hard to keep their attention and motivations because teachers are not in front of students. Students may continue to learn inefficiently without enough understanding subjects.

One of the solutions to the problem is to delegate the observation task to sensors [81]. For example, there is extensive research in cognitive science and pervasive computing utilizing eye movements to recognize daily activities including reading and watching [23] and affective states involving comprehension and attention [99, 144]. The motivation of our work is to build *Comprehension-aware learning assistant* that gives students feedback in real-time with a comprehension estimation based on tracking their watching behaviors with sensors. It saves students' times to solve exercises, motivates students to visualize daily statistics scores calculated automatically, and identifies which part of videos students should watch again or skip.

Most of the comprehension estimation work is sensing reading behaviors on a static document [191], and watching behavior analysis on multimedia such lecture videos is still challenging task. Thus we involve other sensors to increase the performance in addition to eye tracking. For the first step, we survey what type of sensors are used in cognitive science. Then we propose methods to identify effective features and comprehension estimation. Finally, we report experimental results on a dataset involving 10 participants watching video lectures and answering exercises.

Contributions of this section are two-fold. (1) Enhancing the range of comprehension estimation from static documents to multimedia such as video lectures. (2) Investigating which sensor signals are correlated with comprehension.

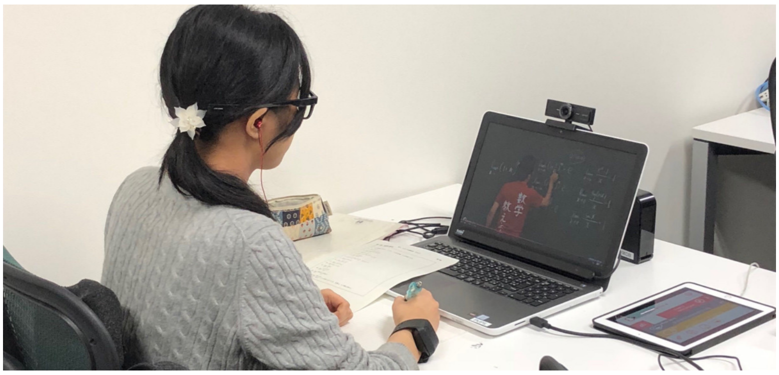


FIGURE 4.9: A student is watching a video lecture with wearing sensors

4.2.1 Approach

We utilize Tobii eye tracker 4C, E4 wristband, and JINS MEME to measure students' learning behaviors. In the following, we present the specifications of the devices, feature calculations, and classification methods.

Feature Calculations

TABLE 4.2: List of features from an eye tracker

No.	Feature Name	Description
1-2	FIX_{X, Y}_VAR	variance of {X, Y} axis of gaze point of fixation
3-7	FIX_D_{SUM, AVE, SD, MAX, MIN}	{sum, ave, std, max, min} of duration of fixation
8-12	SAC_L_{SUM, AVE, SD, MAX, MIN}	{sum, ave, std, max, min} of length of saccade
13-17	SAC_D_{SUM, AVE, SD, MAX, MIN}	{sum, ave, std, max, min} of duration of saccade
18-22	SAC_V_{SUM, AVE, SD, MAX, MIN}	{sum, ave, std, max, min} of velocity of saccade
23	FIX_COUNT	the number of fixation
24	FIX_RATIO	percentage occupied by fixation duration
25	SCREEN_DURATION	duration for which Tobii caught the subject's eyes

TABLE 4.3: List of features from a physiological wristband

No.	Feature Name	Description
26-29	TEMP_{AVE, SD, MAX, MIN}	{ave, std, max, min} of subject's temperature
30-33	EDA_{AVE, SD, MAX, MIN}	{ave, std, max, min} of electrodermal activity
34-36	EDA_DIFF_{SD, MAX, MIN}	{std, max, min} of difference of electrodermal activity
37	EDA_DIFF_PEAK	the number of peaks of difference of electrodermal activity
38-40	E4_ACC_{X, Y, Z}_SD	std of {X, Y, Z} axis of acceleration
41-43	E4_ACC_{X, Y, Z}_LPF_PEAK	the number of peaks of {X, Y, Z} axis of acceleration low path filtered
44-47	HR_{AVE, SD, MAX, MIN}	{ave, std, max, min} of heart rate

TABLE 4.4: List of features from EOG glasses

No.	Feature Name	Description
48-50	JINS_ACC_{X, Y, Z}_SD	std of {X, Y, Z} axis of acc.
51-53	JINS_ACC_{X, Y, Z}_LPF_PEAK	the number of peaks of {X, Y, Z} axis of low path filtered acc.
54-56	JINS_GYRO_{X, Y, Z}_SD	std of {X(pitch), Y(roll), Z(yaw)} axis of angular acc.
57-59	JINS_GYRO_{X, Y, Z}_LPF_PEAK	the number of peaks of {X, Y, Z} axis of low path filtered angular acc.
60-61	EOG_{H, V}_SD	std of {horizontal, vertical} axis of ocular potential

Features No. 1 - 25 shown in Table 4.7 are calculated by Tobii eye tracker 4C. The device extracts raw gaze coordinates on a screen (px) with 90 Hz frequency. By utilizing an approach proposed by Buscher *et al.* [24], we pre-process raw gaze into fixations and saccades.

Features No. 26 - 47 shown in Table 4.8 are from the E4 wristband. TEMP (human temperature), EDA (electrodermal activity), ACC (acceleration) and HR (heart rate) are raw signals from the device.

Head movements and EOG signals obtained from JINS MEME are also used for features as shown in Table 4.4. The reason why we use JINS MEME in addition to

Tobii eye tracker 4C is to record eye movements while students are not watching a video. Students usually take notes with a pen and they don't look at a screen while note taking. Features No. 37, 41 - 43, 51 - 53 and 57 - 59 are calculated on signals low-pass filtered by SciPy because they include many noises.

Classification

We apply binary classification of whether incorrect or correct for each participant. We use SVM and random forest and compare their performances. Regarding SVM, penalty parameter C of the error term is 1, kernel is *rbf*, and class weight is *balanced*. Also, regarding random forest, the number of trees in the forest is 10, the function to measure the quality of a split is *gini*, and class weight is *balanced*. These are default parameters in scikit-learn. We train the model for each participant, i.e., user-dependent approach.

Statistical Analysis

We could infer that the samples are drawn from different distributions. From the 61 features, we apply Welch's t-test for investigating significant features between two classes (correct answer and incorrect answer). We calculate p-values individually in each participant and utilize the mean. In this section, features whose the mean p-value is less than 0.05 are selected and visualized.

4.2.2 Experimental Design

We asked 10 participants to watch lecture videos with sensors as shown in Figure 4.9. We prepared a Tobii eye tracker 4C and a display on the desk. Participants wore JINS MEME EOG glasses, and wear the E4 wristband. After all of the devices are switched on, we start a recording session.

Participant watched video lectures on *Study Sapuri*, one of the most popular e-learning services in Japan. While watching videos, we permitted them to take notes of a blackboard. Every time they finished to watch one lecture, the participant solved related performance exercise. One exercise has 2-6 questions, and all of the questions have already explained how to solve by teachers in the lectures.

After finishing recording session, we scored the exercise. As previously mentioned, one exercise has 2-6 questions. If for one question, both of the answer procedure and the final answer are correct, we regarded this question as correct. For each question, we associated the question with a part of lecture videos which is the basis of answer for subjects to take the exercise. In fact, one video has about 10-30 minutes, and one question has about 2-6 minutes. Also, the participants took

6-19 videos and 19-60 questions on exercise. The number of videos and questions is different depending on subjects because the learning progress speed is different for each student. Another reason is that even if the participants have already taken some lectures, there are several missing data due to inadequate of the experiment like device malfunction.

To increase the number of samples, we divided the questions into windows shown as Figure 4.10. For one question, we let one window size be 30 seconds. The start of the first window is the start of the question. After that window, we make overlap for 10 seconds. Since one question is not necessarily a multiple of 10 seconds, we discarded the last remaining part of the question in this case. As the results, we gained about 400-600 windows as an instance. Our methods were evaluated in both of *leave-one-video-out cross-validation* and *leave-one-question-out cross-validation*.

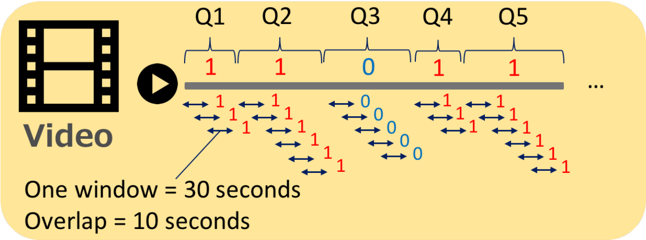
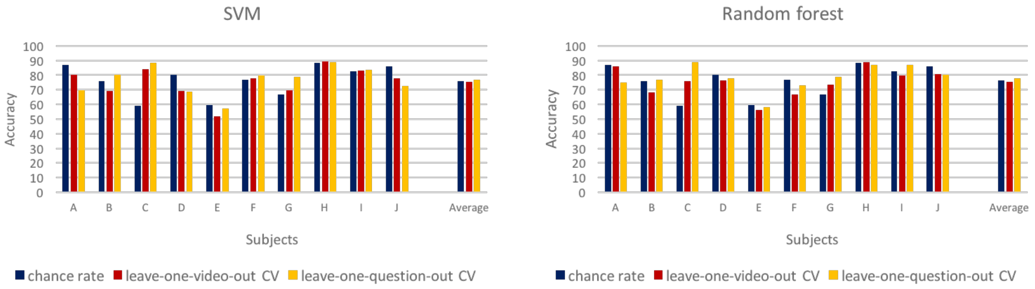


FIGURE 4.10: The process of separating data into training and testing samples

4.2.3 Results and Discussion

Classification accuracies using SVM and Random Forest are shown in Figure 4.11. On both of them, leave-one-video-out cross validation is slightly lower, and leave-one-question-out cross validation is slightly higher than chance rate.



(A) Support Vector Machine

(B) Random forest

FIGURE 4.11: User-dependent classification accuracies

Figure 4.12 shows six features those mean p value are less than 0.05. EDA_AVE was selected with $p < 0.01$ but this might be because baselines of EDA in two participants were strongly higher somehow. Thus we should exclude this from significant features. According to the differences of SCREEN_DURATION, FIX_COUNT, and FIX_D_AVE, participants attract more attention to video lectures in the correct class. This difference has also appeared in EOG_V_SD and JINS_GYRO_Y_SD. Participants moved their head and eyes vertically to take a note frequently in the incorrect class. There are two possibilities on the results: students watching difficult lectures desperately take notes to understand, or the high attention on note taking make students miss important points in the lectures.

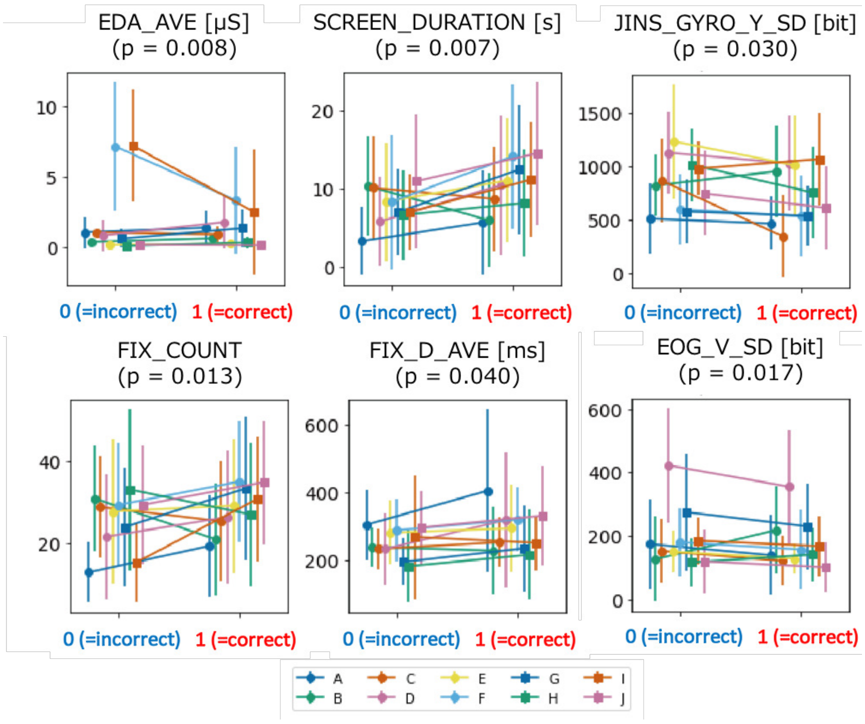


FIGURE 4.12: Mean and standard deviation values calculated for each participant and each class. The six features out of 61 features are selected as $p < 0.05$.

4.2.4 Conclusion

As discussed above, we have obtained five significant features out of 61 features. However, we couldn't get a much higher accuracy of classification than chance rate.

There, we have two strategies to raise accuracy. The first strategy is to add features of not only content independent but also content dependent. Namely, it is conceivable to use new features made by using both human behaviors while watching e-learning and the actual content. Compared with using only human behavior, there is a new possibility of that. The second strategy is to label experimental data not objectively but subjectively. In this experiment, the binary class is objectively labeled by whether subjects can correctly or incorrectly answer the exercise. Instead, we can also make subjects response whether they can understand content after watching it in such as a questionnaire. By doing so, labeling is more suitable for the state of the subjects, so improvement of accuracy will be expected.

After improving the classification accuracy, we are going to implement interventions on video lectures (e.g., suggesting students to watch specific parts in lectures again to eliminate their misunderstandings, preparing exercises students will not be able to solve according to their behaviors, etc). Evaluating whether the interventions can improve learning performance or not is also in future work.

4.3 Affective State Recognition on a Textbook

This section presents a pilot study, finding correlations between sensor signals and affective states collected by a subjective survey after reading a textbook on physics. We select the combination of eye tracking and thermal image analysis to measure affective states because they can be sense without bothering readers and do not interfere with each other. As a learning material, we prepare "Basic Phenomena in Acoustics and Pendulum" The specific contributions of our work are two-folds: (1) Demonstrating that the combination of commercial infrared thermal camera and facial landmark detection is accurate enough to measure students' nose temperatures (2) Investigating effective features measured by an eye tracker and an infrared thermal camera which are related to affective states collected as self-assessments.

Competent handling of multiple representations is supposed to be significant for learning and problem solving in physics [4, 47]. A psychological model for understanding the cognitive processes while working with multiple representations is offered by the Cognitive Theory of Multimedia Learning (CTML; [157]) and the Cognitive Load Theory (CLT; [28]). Referred to as CTML, the generation of a mental model of the learning content requires an active part in information processing. The presentation format of the learning material is essential and can be structured into text/picture or classified according to dynamics and interactivity [64]. Students' learning is improved by presenting text/equations and pictures/graphs/videos instead of learning with text/equations alone. While using the pictorial and verbal/auditory channels simultaneously, sensory and representational differentiations are connected. As a result, cognitive load is reduced. Therefore, greater capacity of working memory is available for forming mental representation models according to CTML and, therefore, learnability is increased.

Besides enhancing cognitive variables, our approach involves that students read textbooks actively through personalized feedback to learn in their way. Therefore, they experience autonomy, which is said to foster motivation in general [150, 151] as well as curiosity, as special features of motivation, in particular [156, 180].

4.3.1 Approach

Eye Tracking

We utilize a remote eye tracker which can attached to a display to track eye movements. Eye gaze data is composed of two metrics - fixations and saccades. A fixation occurs when the gaze falls on something of interest to the screen area and usually lasts for about 100 - 150 ms. The rapid movement of the eye between fixations is called a saccade. As preprocessing, we filter raw eye movements to fixations and saccades on the basis of the approach proposed by Buscher et al [24]. Figure 4.13 shows one of the examples of the filtering.

The average of left and right pupil diameters at any time instant is used as the pupil diameter feature for this work. The duration of each fixation during each question is aggregated to get the fixation duration feature. The length of a saccade is derived from the known values of fixation duration of the eye at a particular two-dimensional coordinate on the screen, at a given timestamp. Similar to the fixation duration feature, the summation of the saccade length corresponding to each participant for each question is used to obtain the feature value. The mean and standard deviation of fixation durations and saccade lengths are calculated as features.

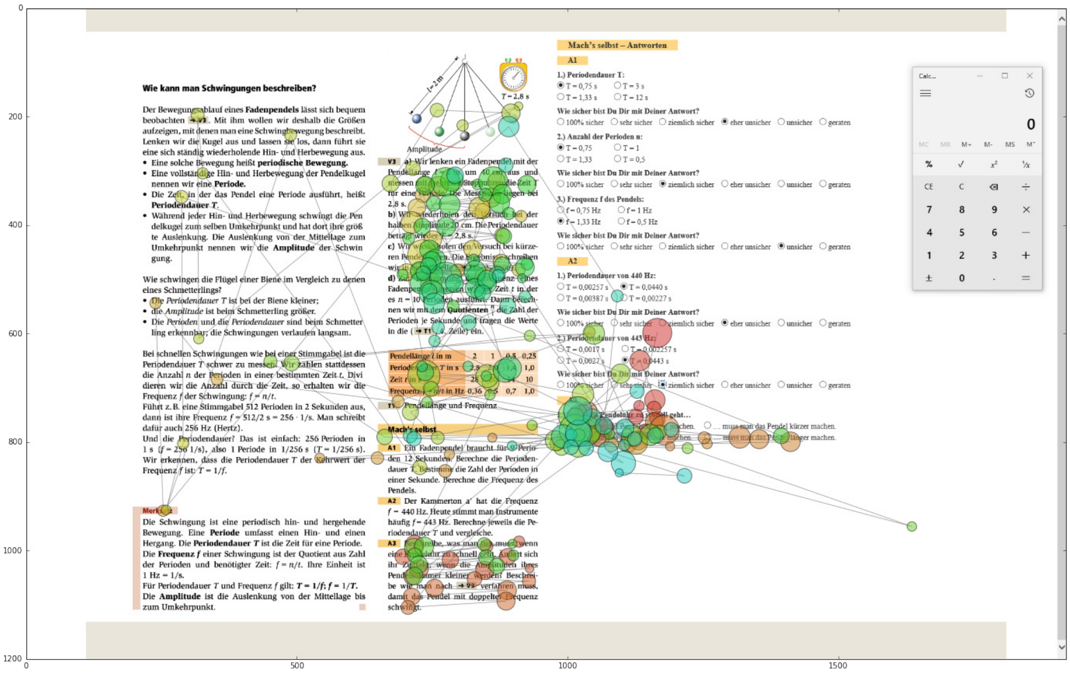
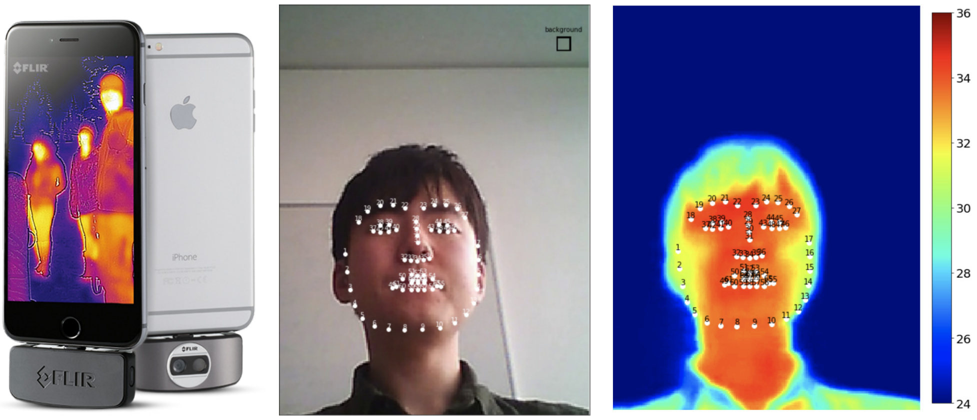


FIGURE 4.13: Fixations on a display during solving a task. The colors represent the order (from red to blue) and the durations are visualized as radiuses.

Nose Temperature Tracking

We utilize *FLIR One for iOS*, a commercial thermal camera which can be attached to a smartphone or a mobile tablet to measure face temperatures (shown in Figure 4.14). We have developed the sensor logging application of the device by ourselves and record the changing of temperatures as a video. Positions of the face and the nose on each frame are detected by using a method proposed by Baltrusaitis *et al.* [10].

The temperature data consists of the nose temperature of each participant at a given time during the experiment. From an initial analysis, we have found that generally, the temperature increases when the students read the textbook and decreases when they start solving exercises (see Figure 4.15). From this data, the slope and the standard deviation of the participant's temperature during the process of solving each question are calculated. Finding out the slope and standard deviation serves to measure the ascend/descend and the fluctuations in temperature.



(A) FLIR One for iOS

(B) RGB image

(C) Thermal image

FIGURE 4.14: An overview of nose temperature sensing

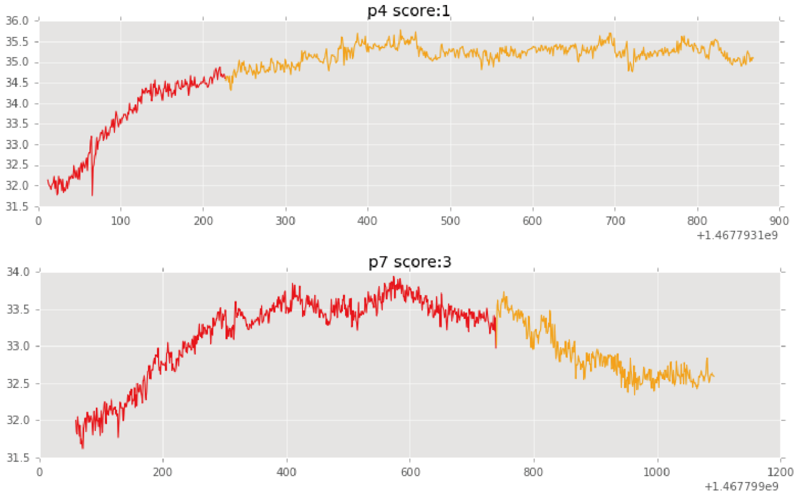


FIGURE 4.15: Examples of changing of nose temperature of two participants (top: low workload, bottom: high workload) during they are reading the textbook (red) and solving the tasks (orange). The x-axis represents timestamps (sec.).

4.3.2 Experimental Design

Figure 4.20 shows an overview of the experimental setup. The SMI 60 Hz remote eye tracker was set up alongside a normal computer desktop to record the eye gaze data and FLIR One for iOS was set to capture the thermal energy from the face. The eye tracker uses a reflection of infrared light to measure eye gaze. We made sure that there is no significant affect in thermal images before starting the experiment. We asked fourteen sixth-grade students (11 or 12 years old) to participate in the experiment. They read a Physics textbook on a screen and solved eight exercises related to the content. As shown in Figure 4.16, the content textbook was displayed on the left page on the screen, and exercises were displayed on the right page. Participants are allowed to use a calculator while solving the exercises. Eye movements on the calculator were excluded in the analysis. Note that seven participants read the text first and other seven participants read questions first. But we treat them as the same condition because there are no significant differences in their performances calculated by the score of the exercises.

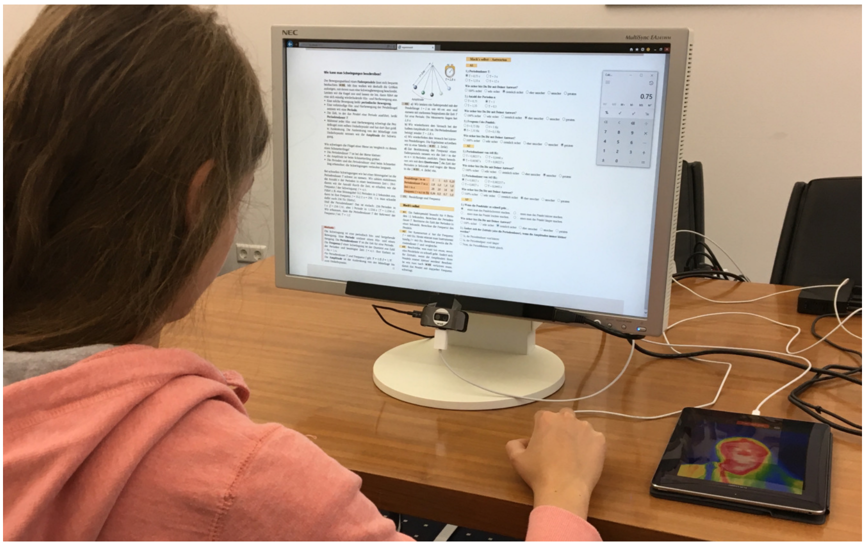


FIGURE 4.16: An experimental setup using an eye tracker and a thermal camera.

To collect ground truth of affective states, we asked participants to answer surveys on a paper form after the recording. We prepared the surveys as shown in Table 4.5 from the viewpoint of Physics education research. They can be categorized with two indexes: the type and the scope. We asked three types of subjective affective states (interest, confidence, workload) plus one objective measurement (expertise). Some of the surveys are general questions about Physics learning (macro) and the others are specific questions about the content (micro). The survey brought to light the interest these students had in learning and researching about physics.

The ratings ranged from one to six, six being "I agree completely and wholly" and one being "I do not agree with it at all". Note that there are two differences about the survey between the original form used in the experiment and reported in this section. (1) The order of the surveys in the form during the experiment was s7, s3, s8, s6, s12, s11, s1, s10, s4, s2, s9, s5, and s13. In this section, they are sorted by their semantic. (2) The surveys were written in German during the experiment. They are translated into English in the table for readers of this section.

We reluctantly exclude two of 14 participants' data as outliers. Their reading and solving time were too fast than other participants, they seemed to select the answers randomly, and their score of exercises were zero. They could not be attentive enough or understand the purpose of the experiment.

TABLE 4.5: Thirteen surveys

No.	type	scope	survey
s1	interest	macro	I enjoy solving physics problems.
s2	interest	macro	I am concerned about homework with topics dealing with physics.
s3	interest	micro	I like the content of the textbook.
s4	interest	micro	I am interested in learning more about the subject of the textbook as well as lectures and homework.
s5	interest	micro	I would like to know more on the topic of textbook in school.
s6	confidence	macro	I am good at physics more than other subjects.
s7	confidence	micro	The textbook text was easy to understand.
s8	confidence	micro	I knew what I had to answer during solving the tasks.
s9	workload	micro	I had to make an effort to solve the questions.
s10	workload	micro	It was difficult to find the right information to solve the questions in the text.
s11	workload	micro	I needed more assistance while reading the textbook.
s12	workload	micro	The textbook made me curious to know more about vibration and acoustics.
s13	expertise	macro	My physics record is about....

4.3.3 Results and Discussion

Table 4.6 represents the Pearson correlation and p-values (in brackets) between the features and surveys. High correlations with p-values less than 0.05 are highlighted as bold fonts. From these values, we have found three insights. First, surveys related to workload including s10 "It was difficult to find the right information to solve the questions in the textbook." and s9 "I had to make an effort to solve the questions." can be measured by a decrease of the nose temperature during solving exercises (p = 0.001 and p = 0.012). Second, increase of pupil diameter represents a student's interest including s3: "I like the content of the textbook." and s1 "I enjoy solving physics problems." (p = 0.008 and p = 0.030 during reading; p = 0.006 and 0.013 during solving). Third, students who read a textbook and exercises with small saccades felt high confidence in their understandings reflected in s7 "The textbook was easy to understand" and s8 "I knew what I had to answer during solving the tasks" (p = 0.025 and p = 0.035).

TABLE 4.6: Pearson Correlation and p-values features and thirteen surveys.

	interest					confidence			workload				expertise
feature	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13
nose slope reading	-0.0 (.97)	0.3 (.37)	0.4 (.23)	0.0 (.88)	0.0 (.89)	0.2 (.60)	-0.2 (.49)	0.1 (.75)	0.1 (.73)	-0.5 (.09)	-0.2 (.50)	0.2 (.53)	-0.2 (.61)
nose slope solving	-0.3 (.39)	0.3 (.40)	-0.4 (.25)	-0.0 (.96)	0.2 (.53)	**0.6 (.04)**	0.1 (.74)	0.4 (.22)	**-0.7 (.01)**	**-0.8 (.00)**	-0.2 (.48)	0.3 (.41)	-0.4 (.16)
nose std reading	0.0 (.92)	0.3 (.38)	0.4 (.18)	0.1 (.73)	0.1 (.84)	0.0 (.88)	-0.2 (.55)	0.3 (.40)	0.2 (.51)	-0.4 (.15)	-0.1 (.76)	-0.1 (.73)	0.1 (.85)
nose std solving	**0.6 (.04)**	-0.4 (.22)	0.0 (.90)	-0.1 (.76)	-0.3 (.42)	-0.2 (.58)	-0.4 (.19)	**-0.6 (.03)**	0.1 (.73)	0.3 (.29)	0.2 (.57)	0.1 (.70)	0.1 (.71)
pupil mean reading	**0.6 (.03)**	**-0.6 (.03)**	**0.7 (.01)**	0.1 (.72)	0.5 (.14)	0.1 (.80)	0.1 (.88)	**-0.7 (.01)**	0.3 (.39)	0.5 (.10)	0.5 (.08)	-0.0 (.99)	-0.2 (.50)
pupil mean solving	**0.7 (.01)**	-0.6 (.06)	**0.7 (.01)**	0.1 (.80)	0.5 (.12)	0.1 (.76)	0.1 (.77)	**-0.7 (.02)**	0.3 (.27)	0.5 (.12)	0.5 (.10)	-0.2 (.63)	-0.1 (.71)
pupil std reading	0.4 (.16)	-0.4 (.23)	**0.8 (.00)**	0.4 (.22)	0.5 (.14)	-0.1 (.67)	-0.0 (.92)	-0.4 (.20)	0.5 (.07)	0.5 (.09)	0.4 (.25)	0.0 (.99)	-0.1 (.69)
pupil std solving	0.2 (.57)	-0.4 (.21)	**0.6 (.03)**	0.6 (.06)	0.4 (.25)	-0.3 (.40)	-0.0 (.99)	-0.2 (.47)	0.5 (.13)	**0.6 (.03)**	0.3 (.40)	0.0 (.91)	-0.2 (.57)
fixation mean reading	-0.4 (.25)	0.3 (.36)	-0.4 (.18)	-0.1 (.87)	-0.1 (.80)	0.0 (.97)	0.3 (.42)	0.5 (.08)	-0.1 (.85)	-0.1 (.83)	-0.2 (.58)	-0.5 (.14)	0.3 (.37)
fixation mean solving	-0.3 (.29)	0.0 (.91)	-0.4 (.16)	0.0 (.99)	-0.2 (.51)	-0.2 (.44)	0.5 (.14)	0.4 (.19)	0.1 (.81)	0.3 (.32)	-0.4 (.24)	-0.3 (.33)	0.2 (.59)
fixation std reading	-0.3 (.30)	0.2 (.63)	-0.3 (.40)	-0.1 (.86)	-0.0 (.89)	-0.0 (.97)	0.2 (.51)	0.4 (.24)	-0.1 (.73)	-0.1 (.87)	0.1 (.76)	-0.3 (.31)	0.2 (.63)
fixation std solving	-0.4 (.26)	-0.1 (.84)	-0.1 (.83)	-0.1 (.83)	-0.1 (.68)	-0.2 (.44)	0.5 (.11)	0.1 (.70)	0.2 (.52)	0.4 (.26)	-0.2 (.53)	0.1 (.84)	-0.1 (.75)
saccade mean reading	0.1 (.81)	**-0.6 (.05)**	-0.1 (.82)	0.2 (.50)	-0.1 (.75)	-0.0 (.93)	**-0.6 (.02)**	-0.5 (.14)	-0.5 (.12)	-0.0 (.95)	0.6 (.05)	**0.7 (.02)**	-0.4 (.23)
saccade mean solving	0.3 (.28)	-0.3 (.41)	0.4 (.22)	-0.3 (.41)	-0.3 (.29)	-0.3 (.42)	-0.5 (.10)	**-0.6 (.03)**	0.3 (.43)	0.1 (.74)	0.3 (.41)	0.2 (.50)	0.2 (.60)
saccade std reading	0.0 (.96)	-0.5 (.10)	-0.1 (.80)	0.3 (.28)	-0.1 (.68)	-0.1 (.87)	**-0.7 (.01)**	-0.3 (.29)	-0.4 (.26)	-0.0 (.94)	0.5 (.10)	**0.7 (.01)**	-0.4 (.24)
saccade std solving	0.3 (.35)	-0.2 (.61)	0.4 (.18)	-0.0 (.96)	-0.2 (.49)	-0.2 (.50)	-0.5 (.11)	-0.5 (.12)	0.3 (.42)	0.1 (.85)	0.1 (.64)	0.4 (.24)	-0.1 (.87)

The temporal resolution of sensing is a remaining issue. Although the change in the nose temperature is an effective feature to understand a student's effort, it requires a long time to be observed (see Figure 4.15). In the application scenario, it can be used for the measurement on each learning unit or page. But it seems difficult to apply our measurements on small parts such as each paragraph, image, or sentence. We need to investigate how much the time resolution can be minimized.

4.3.4 Conclusion

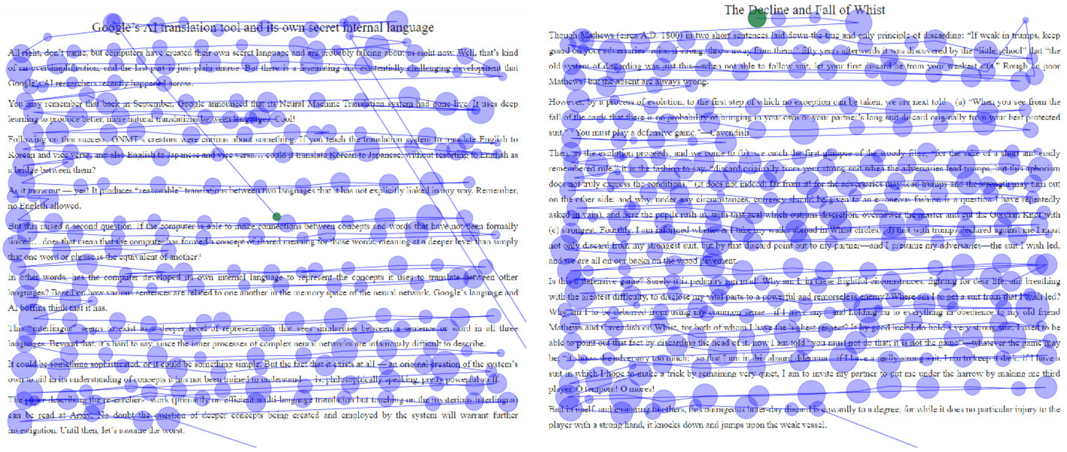
This section demonstrated the affective state analysis by using an eye tracker and an infrared thermal camera. We have developed an application to retrieve the change in the nose temperature from a commercial infrared thermal camera (FLIR ONE). We asked 12 high school students to read/solve learning materials in Physics and investigated the relation between sensor signals and surveys about their affective states. The changing of the pupil diameter was highly correlated with interest. Although the temporal resolution was not enough high for a real time application scenario, the changing of the nose temperature represented their efforts for reading/solving learning materials.

4.4 Interest Recognition on Newspaper Articles

Interest in reading can be motivated by concentration, curiosity, and demand. It may not rise out of habit but it motivates the habit and subsequently the learning process. According to Ibrahim Bafadal, "Reading is a process of capturing or acquiring the concepts intended by the author, interpret, evaluate the author's concepts, and reflect, or act as intended of those concepts". Hence it not only depends on the ability to interpret and evaluate the contents but also the will to do the same for comprehensive understanding [74, 142, 162].

This urge in reading, if recognized, can be used to improve the data made available to the reader and also help in better human-document interaction and the design. Predicting a reader's interest can help to make document more interactive or dynamic [15]. Research done on dynamically changing text shows that reading dynamic text is much smoother and faster than reading static text [177]. Eye gaze, if used to predict a user's interest, comprehension and difficulty, can influence his/her interaction with the learning environment and thereby affect the learning process [36]. This can further assist teaching techniques and promote understanding and active interest in students.

As shown in Figure 4.17, the reading behavior should reflect how much a reader is interested in the document. We propose an interest detection method by utilizing sensors. Following two research questions addressed in this section. (1) How accurately can sensing devices estimate a reader's interest? (2) Since reading behaviors are different for each reader, which measurement can be used as a common feature?



(A) on a document labeled as interesting

(B) on a document labeled as boring

FIGURE 4.17: Examples of eye gaze of one reader (circle: fixation, line: saccade)

4.4.1 Approach

Eye Tracking

Figure 4.18 shows an example of the gaze events while reading an article. Eye movements while reading are composed of three basic metrics: fixations, saccades and blinks. A fixation occurs when the gaze falls on something of interest to the screen area and usually lasts for about 100 - 150 ms. The rapid movement of the eye between fixations is called a saccade. A blink is a semi-autonomic rapid closing of the eyelid. Pupil diameters can be also obtained from an eye tracker. We detect the gaze events by following steps.

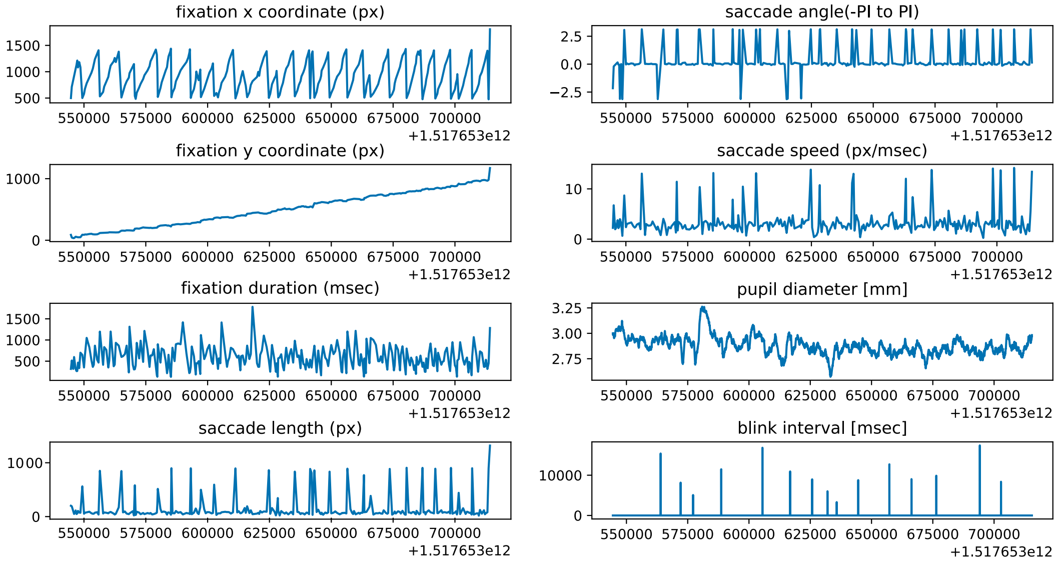


FIGURE 4.18: Gaze events calculated by a signal of an eye tracker

As preprocessing, we filter raw eye movements to get fixations and saccades on the basis of the approach proposed by Buscher *et al.* [24]. The midpoint of the left and the right gaze coordinates is taken as the gaze point, only if both values are nonzero, else the (left or right) non-zero coordinates is taken as the gaze point. A fixation typically consists of more than six successive gaze locations grouped in succession. This makes the minimum fixation duration 100 ms as mentioned earlier for data recorded at a rate of 60 Hz. The successive gaze points making a new fixation should fit inside a threshold rectangle of 30×30 pixels. All further gaze points falling inside a 50×50 pixel rectangle is considered to belong to the current fixation. This is done so that noise and small eye movements are tolerated. If the gaze point does not fall in the rectangle, it is either an outlier or the start of a new fixation, which further merges with six other points. The fixation is considered to have ended if at least six successive gaze points cannot be merged.

The movement or transition from one fixation to the other is recorded as a saccade. Saccades are further divided into forward saccades and backward saccades. The x-coordinate of successive fixations denotes the direction of the saccade. Forward saccades imply regular reading behavior, while backward saccades can either be regressions or line breaks. Regressions are backward eye movements which allow re-reading of the text [18]. Line breaks are separated from regressions by analyzing the length of the backward saccade. If the length is equal to or greater than the length of a line, then they are categorized as line breaks (observed as peaks in the saccade length in Figure 4.18).

Further, we use pupil diameter obtained from raw gaze data, which is the average of left and right pupil diameter, if both are non-zero. Another characteristic eye behavior we record is blink. The average duration for a blink of a human eye is 100 - 400 ms. Hence, 6 - 24 consecutive zeroes in the left and the right gaze coordinates are considered as one blink in our approach. The average latency of two consecutive blinks is one second and blinks detected in between are considered as noise.

TABLE 4.7: List of features from an eye tracker

No.	feature
1-2	{mean, SD} of fixation duration
3-4	{mean, SD} of forward saccade length
5-6	{mean, SD} of forward saccade speed
7-8	{mean, SD} of regression length
9-10	{mean, SD} of regression speed
11-12	count of {forward saccades, regressions}
13	regression ratio
14-15	{mean, SD} of pupil diameter
16	blink frequency
17	SD of blink interval

On the basis of the gaze events, we extracted seventeen features for further analysis, as listed in Table 4.7. *Fixation duration* is the time taken for each fixation. *Forward saccade length* is the distance between the two consecutive fixations that make the saccade. *Forward saccade speed* or *Regression speed* is the length of the saccade divided by the time taken for the saccade. *Regression ratio* describes the fraction of regressions out of the total number of saccades (i.e., (No.11 / (No.11 + No. 12)). *Regression length* is the distance between the two fixation coordinates that makes the regression. *Pupil diameter* is the diameter of the right pupil obtained from the raw gaze data, provided the x and y gaze coordinates are non-zero else taken as zero. Since pupil diameter is user and environment dependent, it was taken as a relative value compared to the pupil diameter during the questionnaire which was taken as a baseline. *Blink* *frequency* is the number of blinks divided by the total time taken by the reader (for each document). *Blink interval* is the time lag between two consecutive blinks.

Physiological Sensing

We utilize E4 wristband for to measure a user's behavior. It is used for the acquisition of real-time physiological data with the help of sensors designed to gather high-quality data. It has a *photoplethysmography (PPG) sensor* which measures the blood volume pulse (BVP), an *electrodermal activity (EDA) sensor* to measure electrical properties of the skin, an *infrared thermopile* to measure skin temperature.

We decompose a raw EDA signal to the phasic and tonic component as shown in Figure 4.19 by utilizing *cvxEDA* algorithm [67]. Then the following 6 features are calculated from the components. (1) the slope of the tonic part of the signal for which the slope of the line of best fit was used (Linear Regression), (2) EPC - sum of all positive EDA changes, (3) Minimum peak amplitude of the phasic signal, (4) Maximum peak amplitude of the phasic signal, (5) Mean amplitude of the phasic signal and (6) Number of phasic responses [118].

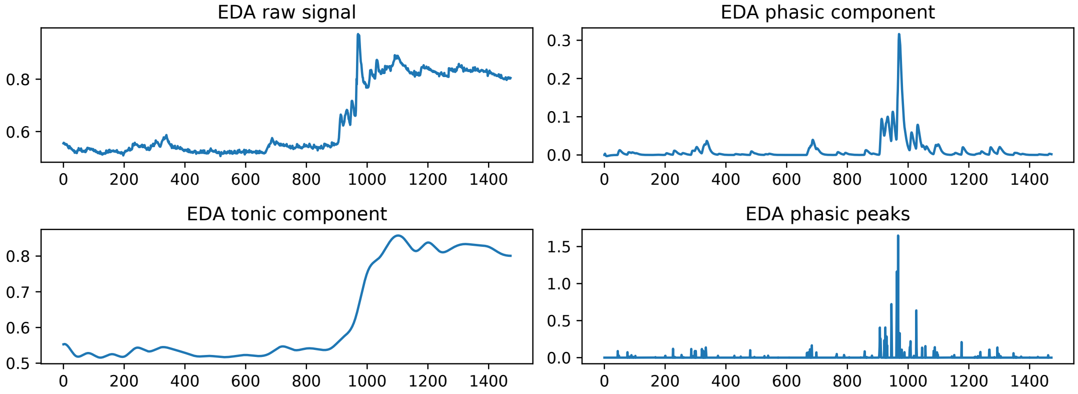


FIGURE 4.19: A decomposition of an EDA raw signal

Features relevant to heart rate (HR), inter-beat interval (IBI) and BVP are extracted from the data. The features used are - (7, 8) mean and standard deviation of BVP, (9, 10) mean and standard deviation of HR, (11, 12) difference in mean and standard deviation of HR amplitude during task and baseline, (13) standard deviation of IBI normalized by baseline (data recorded 5 seconds before start), (14) square root of the mean of the square (RMSSD) of the successive differences between IBI. Features from skin temperature recorded by the wearable are also included namely, (15, 16) mean and standard deviation of skin temperature, (17) difference in mean of temperature during task and baseline, and (18) slope of temperature.

TABLE 4.8: List of features from a wristband

No.	feature
1	slope of the tonic component
2	the number of positive EDA changes (EPC)
3-4	{min, max} peak amplitude in the phasic component
5	mean amplitude of the phasic component
6	the number of phasic responses
7-8	{mean, SD} of BVP
9-10	{mean, SD} of HR
11-12	{mean, SD} of HR amplitude normalized by baseline
13	SD of IBI normalized by baseline
14	RMSSD between IBI normalized by baseline
15-16	{mean, SD} of the arm temperature
17	mean of the arm temperature normalized by baseline
18	slope of the arm temperature

Classification

We utilize *Support Vector Machine* (SVM) to estimate the interest of a reader. Hyper parameters of the SVM classifier (*C*, *kernel* and *gamma*) are optimized by 3-fold grid search cross-validation. It searches exhaustively through a manually defined set of parameters and finds those that achieved the highest score in the validation procedure. We separate training data into training for parameter optimization and the evaluation for each classification.

4.4.2 Experimental Design

Figure 4.20 shows an overview of the experimental setup. We displayed documents on a computer desktop and recorded reading behaviors. To capture the reader's interest, we prepared newspaper articles with a wide range of topics from different platforms like technology, politics, sports, cooking etc. Thirteen university students (mean age: 25, std: 3, male: 6, female: 7, 2 of them are familiar with an eye tracker and a physiological sensing wristband) participated in the experiment where each of them was asked to read eighteen newspaper articles comprising of 403 - 649 words each (mean: 555, std: 70) as shown in Figure 4.17.

After reading each document, participants answered three questions. (1) the level of interest they had in the article, which was used as ground truth (from 1 to 4, where 1 indicated "very boring" and 4 indicated "very interesting"), (2) a self-assessment about how much of the content the reader understood (subjective comprehension, from 1 to 4, where 1 indicated "I could not understand the article" and

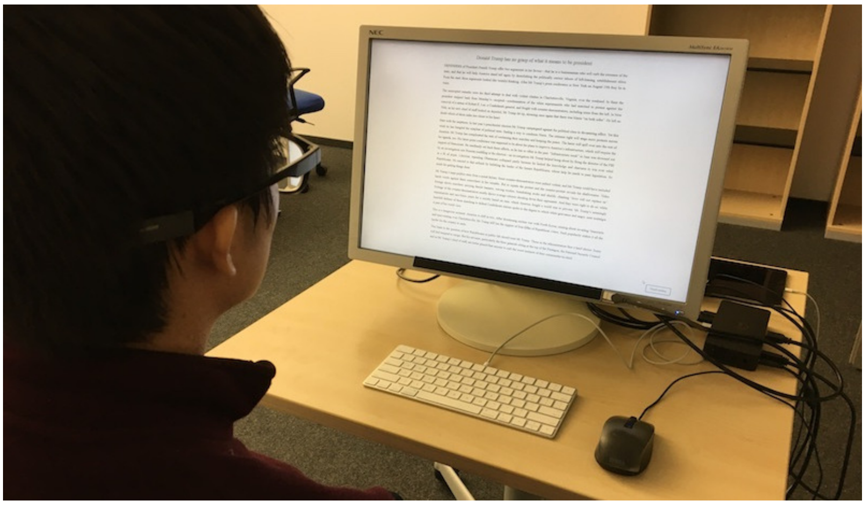


FIGURE 4.20: An overview of the experimental setup. A participant is reading a news article on a display with SMI REDn Scientific 60 Hz remote eye tracker.

4 indicated "I could understand the article"), and (3) one multiple-choice question about of the article (i.e., objective comprehension).

In order to avoid eye-fatigue, the recordings were done in two sessions of one hour each. The experimental setup was maintained in a stable position from the first until the last recording. The lighting of the room was set so as not to affect the gaze data (pupil diameter). A calibration of an eye tracker was performed before reading every document.

We followed three different approaches to separate the train-test data before classification. *Leave-one-recording-out* uses each recording (data of each participant on each document) as test data, the rest as training and the average of the accuracy in all cases together is taken as the classification accuracy. Similar to this approach, *leave-one-document-out* approach exempts the data of a document completely from the training set and uses it for testing. *Leave-one-participant-out* approach uses one the data from all participants except one as training and uses the data from the left-out participant as testing. It is quite significant as, in a realistic scenario; the system does not have training data from a new user.

4.4.3 Results and Discussion

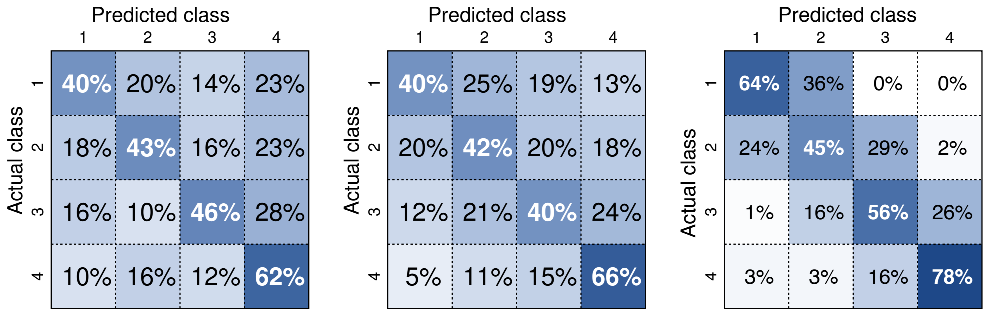
Table 4.9 represents the classification accuracies using SVM. The most frequently selected hyper parameters were C: 32, gamma: 0.125, kernel: Radial Basis Function. Note that we preliminary considered *Random Forest Classifier*. But we found that SVM performs better in our classification task. We also incorporated feature reduction techniques like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), but there was no commendable improvement in the classification.

We included non-eye related measures like reading speed and the level of subjective/objective comprehension of the user. The accuracies are summarized in Table 4.9. Figure 4.22 shows the Pearson's correlation of the features for each participant with the level of interest. The level of subjective comprehension of a person can be seen to have a very high effect on a person's level of reading interest (denoted by red-high and blue-low). We got an accuracy of 50 % when features from eye tracking or physiological sensing were used for classification (Table 4.9).

TABLE 4.9: Accuracies of four-class classifications [%]

	leave-one-participant-out	leave-one-document-out	leave-one-recording-out
1. reading speed	25	32	35
2. subjective comprehension	**59**	60	58
3. objective comprehension	34	34	34
4. eye tracking	32	47	50
5. physiological sensing	37	46	50
combination 2 and 4	50	**66**	**64**
combination 2 and 5	47	54	53

The distribution of the predicted classes are observed as confusion matrixes shown in Figure 4.21. The overall correlation between the various features used and the labels are still quite small. However, when individual participants were considered, from correlations between features and interests shown in Figure 4.22 and Figure 4.23, we found that there was (1) a negative correlation between mean/standard deviations of fixation duration with the interest labels, (2) a considerably small positive correlation existed for the standard deviation of regression speed, (3) also with the number of forward saccades. But this was observed for only half the number of participants or less, the rest having no or very slight correlations with the features.



(A) Gaze only

(B) E4 only

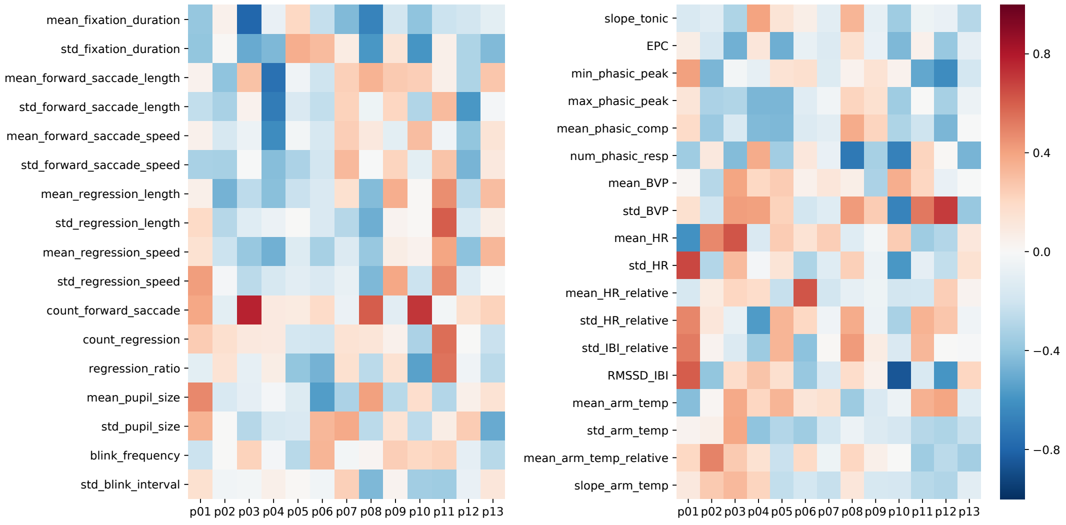
(C) Gaze and subjective comprehension

FIGURE 4.21: Confusion matrices of the leave-one-recording-out condition

Although the accuracies were not as high as expected, this research threw light on using a remote eye tracker for affective state measurement. We found that mean forward-saccade speed, mean fixation duration and mean regression speed plays a vital role in predicting a reader's interest. And that SVM with an RBF kernel is best to classify gaze-based features.

However, a higher correlation of the features to the labels was expected, though it was observed to be quite small. The correlation was quite different in the case of each participant for all features except for ones earlier mentioned. This led us to believe that cognitive predictions are user-dependent and not just document-dependent. For example, pupil diameter may not undergo the same changes in every user during the same psychological process. These features are dependent on the user and his/her affective state. We also found that the collection of ground truth related to interest and understanding are widely prone to human error and individual behavior.

Subjective comprehension of a person has a very high correlation to the level of interest while reading, which makes sense, since interest can only be realized if the person truly understands the text. But using eye measures while reading is an added advantage to understand this and should be deeply explored, since it can be realistically collected while reading without reader intervention.



(A) Eye gaze (B) Physiological signal

FIGURE 4.22: Pearson correlations between interests and features for each participant

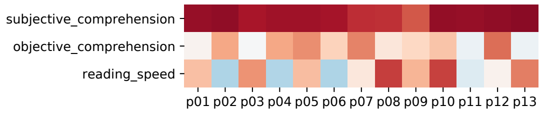


FIGURE 4.23: Pearson correlations between interests and features without sensing

4.4.4 Conclusion

This research demonstrated that eye measures from a remote eye tracker can be successfully used to predict a reader's interest. We obtained data from experiments conducted with 13 students, who read 18 newspaper articles each. We extracted seventeen features from raw gaze data obtained from the tracker and used it for further classification of the data into different levels of interest. Although the correlation of the features with the interest labels were not as high as expected, forward-saccade speed, fixation duration and regression speed were significant.

This work can be extended to include data from other sensors like an infrared thermal camera to measure nose temperature and a physiological wristband to measure heart rate and skin electrical conductance. Data acquisition could also be improved by controlling the environment to present a stress-free or at-home experience for the reader. Also, unsupervised learning could be used for data classification to avoid human error in the ground truth.

Although the accuracies were not as high as expected, this research threw light on the features that could be extracted from the physiological data from a wearable device and its role in predicting the affective state of the reader. The collection of ground truth related to interest and understanding are widely prone to human error and individual behavior, which is also a reason for the low accuracies.

4.5 Confidence Recognition on Multiple-Choice Questions

Self-confidence is a base of meta-cognitive judgments and the most common paradigm in meta-cognitive domains ranging from decision making and reasoning [2, 58] to perceptual judgments [57, 138] and memory evaluations [50, 55]. It is a manifestation of meta-cognitive assessing of own knowledge or scholastic ability, and affected by proficiency, achievement, cognitive anxiety and difficulty of a task [35]. Self-confidence can benefit a student's engagement and learning outcome [119]. Moreover, previous research has proved that self-confidence enhancement is significantly correlated to learning progress [77]. From the above, measuring self-confidence can be helpful for checking learning progress and states of students.

One of the most critical cases that self-confidence plays an important role is on Multiple-Choice Questions (MCQ). MCQ is a type of question asking to select the most appropriate choice from given ones for a question. Since the information obtained from MCQ is only the correctness of answers, it is hard to distinguish the case that students answered with high confidence or randomly without confidence. Actually, the following two cases are serious: (a) the case that students answered incorrectly with confidence and (b) the case that students answered correctly without confidence. In the case (a), they have the wrong knowledge, which may cause other misunderstandings. In the case (b), they just answered correctly by chance, and lose a chance to correct knowledge.

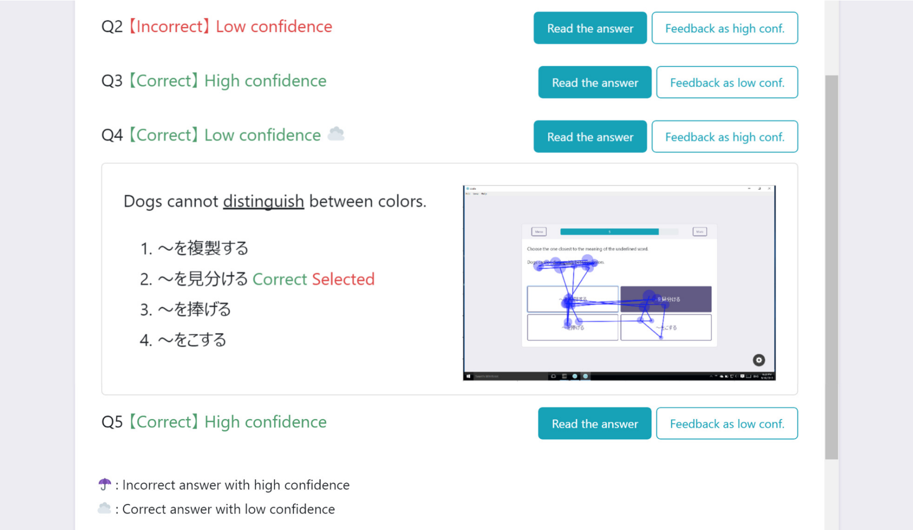
This section presents a solution to such serious cases. We propose a system that estimates self-confidence while solving MCQ. We employ eye movement data, because of our assumption that eye movement with confidence is different form that without it. On the basis of the estimation result, our system generates a report about which question should be reviewed again after solving (see Figure 4.24).

We conducted two data recordings in order to evaluate the performance. In the first recording, we created a small but well-designed dataset in the laboratory (11 Japanese university students solved 880 questions in English). On this *laboratory dataset*, we investigated effective features, the accuracy, and user-dependency. Then, we trained a classification model by the laboratory dataset and demonstrated the system in a cram school. During the five-weeks demonstration, we obtained real solving behaviors (72 Japanese high school students solved 145,489 questions in English and 14,302 of them are labeled by themselves). On this *wild dataset*, we realized the limitations and problems of our system. The number of training samples required for the self-confidence estimation is also investigated by using this large dataset. Our contributions are as follows: (1) On the laboratory dataset, we have succeeded to estimate self-confidences on MCQ in practical accuracies: 76.1 % on a user-independent training. (2) On the wild dataset, we have evaluated our proposed system on real solving behaviors and discussed limitations and potentials.

Many types of research have mentioned correlations between self-confidence or other affective states and behaviors of people in specific tasks including achievement test of learning [98, 164], cognitive test [103, 170] and cooking [140].

Roderer *et al.* have gathered participants in several ages and have found a correlation between self-confidence of participants and their age. Junior participants have tended to get higher self-confidence than senior participants [148]. In contrast to this research, we gathered participants in almost the same age so as to investigate self-confidence with only information in answering. The researches referred above, however, only have proved the correlations. On the other hand, our work is not only to find correlations but also to estimate self-confidence for practical applications.

The closest to our work is work by Yamada *et al.* They have estimated self-confidences on MCQ by utilizing an eye tracker and SVM with a user-dependent training [188]. Compared to their work, we have investigated two training approaches (an user-dependent and an user-independent) and demonstrated in the real environment. Assuming that we implement a real application, it is hard to ask all new users to record their behaviors for training the system before using it. Providing an option of a user-independent method is necessary for practical applications.



Q2【Incorrect】Low confidence

Q3【Correct】High confidence

Q4【Correct】Low confidence ☁

Dogs cannot distinguish between colors.

1. ～を複製する
2. ～を見分ける Correct Selected
3. ～を捧げる
4. ～をこする

Q5【Correct】High confidence

☂ : Incorrect answer with high confidence

☁ : Correct answer with low confidence

FIGURE 4.24: Confidence-aware learning assistant (CoaLA). After solving some Multiple-Choice Questions, it recommends questions which should be checked again on the basis of not only the correctness but also the self-confidence estimated by eye movements.

4.5.1 Approach

We hereby explain our proposed method. Our method has the following three steps. Firstly, we display an English question and record the eye gaze. Secondly, we extract features from the recorded data. Finally, we estimate the self-confidence of the answer as a classification task.

Data Recording and Pre-processing

The eye gaze of a user is recorded by a remote eye tracker attached at the bottom of a display. The output of an eye tracker includes coordinates of the gaze on a display and their timestamps.

Eye movements while solving MCQ are composed of three metrics: fixations, saccades, and blinks. A fixation indicates an event when the gaze pause in a certain position over a certain period usually minimum 100 ms. A rapid movement between fixations is called a saccade. We classify raw eye gaze into fixations and saccades by utilizing an algorithm proposed by Buscher *et al.* [24].

A blink – semi-autonomic rapid closing of the eyelid – is measured as a coordinate (0, 0) in the output of an eye tracker. But it is not analyzed in our method because a time required to solve one question (from 10 seconds to one minute) is too short to calculate statistical features. In addition, a smooth pursuit occurs when a person tracks a moving object with slow speed. But this metric is not considered in this method because all information on a display is fixed.

Feature Calculation

We define Areas of Interest (AOIs) as rectangles covering a question and each choice in order to recognize deep behaviors (e.g., a ratio of reading-times on a question and choices, a process of the decision with comparisons of choices, etc.) Fixations and Saccades are automatically associated to the corresponding AOIs in this step.

We extract features from the fixations and saccades. Table 4.10 shows the 30 extracted features. Features No. 1-14 are related to fixations and the features No. 15-28 are related to saccades. We also use the reading-time and the correctness of the answer as features.

Feature Selection

As the common behavior of users, we select effective features from 30 features to estimate self-confidence. We apply a forward stepwise for the selection. We create a subset of features. At the initial state, the subset is empty. We calculate average precision scores of estimations using each feature, and insert one with the best feature

TABLE 4.10: List of features from an eye tracker

No.	Feature
1-2	fixation {count, ratio} on Choices
3-4	fixation {count, ratio} on Question
5-8	{sum, mean, max, min} of fixation durations on Choices
9-12	{sum, mean, max, min} of fixation durations on Question
13-14	variance of {x, y} coordinate of fixations
15-16	{sum, mean} of saccade length
17-20	saccade count: {all, on Question, between Choices, between Question and Choices}
21-24	{sum, mean, max, min} of saccade durations
25-28	{sum, mean, max, min} of saccade speeds
29	reading-time
30	correctness of the answer

to the subset. Then performances of estimations with features in the subset and one new feature are calculated and keep the best combination again. These processes are repeated until the new subset performs better than the old subset. We utilize two-fold cross-validation for the each estimation. Note that this step is proceeded only while training. Preliminary selected features are used to classify an unknown sample.

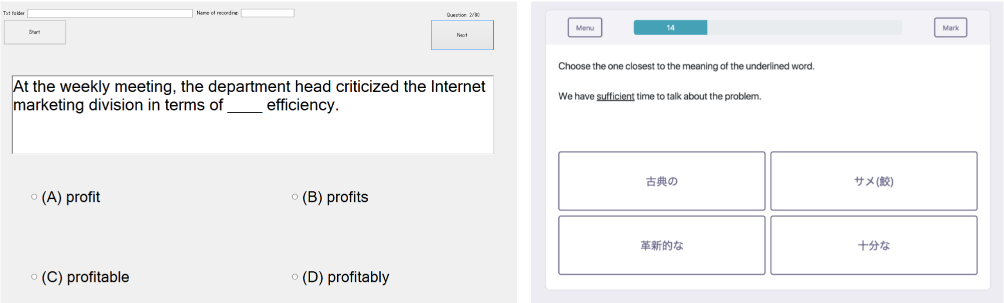
Classification

We estimate self-confidence of answers with a Support Vector Machine (SVM) by using the selected features. RBF kernel with penalty parameter: $C = 1$ and how far the influence of a single training example reaches: $gamma = 0.125$ were selected experimentally and are used for the SVM. As a preliminary experiment, we tested other machine leaning techniques including Random Forest, and it has been found that SVM performs the best overall in our classification task.

4.5.2 Experimental Design

Recording in the Laboratory

We asked 11 participants (male: 9, female: 2) to solve 80 MCQ about English vocabularies and grammars. All the participants were Japanese university students. As shown in Figure 4.25a, they answered the most appropriate word for a blank in a question from choices. After answering each question, they answered a questionnaire “Do you have a confidence in your decision?” with “Yes” or “No”. The results of questionnaires were used as ground truth labels. We utilized Tobii eyeX which is a kind of stationary eye tracker to measure eye movements of participants. The eye tracker was calibrated to a participant at the beginning of the experiment.



At the weekly meeting, the department head criticized the Internet marketing division in terms of ____ efficiency.

○ (A) profit ○ (B) profits

○ (C) profitable ○ (D) profitably

(A) Recording in the laboratory

Choose the one closest to the meaning of the underlined word.

We have sufficient time to talk about the problem.

古典の サメ(鮫)

革新的な 十分な

(B) Recording in the wild

FIGURE 4.25: Examples of questions for the recordings

One of the advantages of this controlled dataset is that we corrected data with the same number of samples from all participants. By using this balanced data, we evaluated our method with a user-dependent training and user-independent training. The performance of the model is evaluated by leave-one-document-out cross-validation. In the user-dependent training, a model was built for each participant (i.e., 79 samples of the participant were used for training to predict the remaining one sample). In the user-independent method, a model was built for each participant and each document (i.e., 79×10 samples excluding the participant and the document were used for training to predict one sample). In the user independent training, the model was built in the document independent manner.

Recording in the Wild

We collaborated with a cram school and installed our system in the school. Students solved MCQ about vocabularies in English on the system. Then they printed out a list of words involving incorrect answers and correct answers with low self-confidence. The questions were prepared by the cram school. The main purpose of this installation was not recording data but demonstrating the system in the real environment. Therefore, unlike the recording in the laboratory, we did not prevent students’ natural behaviors. A calibration of an eye tracker was performed once before a student starts using the system. We asked their self-confidence (ground truth) once every five questions. Each student has own username in order to track who solved which question with or without confidence. The number of solved questions depends on the students. We utilized Tobii 4C remote 90 Hz eye tracker for this recording. Note that an upgrade key provided by Tobii is applied to use it for the scientific purpose. The duration of this demonstration was around five weeks. 83 students used our system and we collected 145,489 solving behaviors in total.

Since real recordings included many noisy behaviors, following filterers were applied to obtain a reliable dataset. (1) We decided to analyze labeled data in this work. (2) Data with invalid usernames (e.g., *guest*) are ignored in the analysis. (3) Data with only few eye gaze (a ratio of valid gaze coordinates is less than 80 % of one recording) are also ignored. Finally, the wild dataset consists of 14,302 valid samples from 72 students. We evaluated our proposed self-confidence estimation on this dataset with 10-fold cross-validation.

TABLE 4.11: Distributions of samples in the datasets

(A) Laboratory dataset

	Confident	Unconfident
Correct	131	89
Incorrect	408	252

(B) Wild dataset

	Confident	Unconfident
Correct	10,529	2,125
Incorrect	656	992

4.5.3 Results and Discussion

Table 4.11 shows distributions of samples. Characteristics of questions in the two datasets are different. Questions in the laboratory dataset seem to be difficult for participants, and there are more incorrect answers than correct answers. On the other hand, most of the answers in the wild dataset were correct. Accuracies and average precisions (AP) of each condition are summarized in Table 4.12.

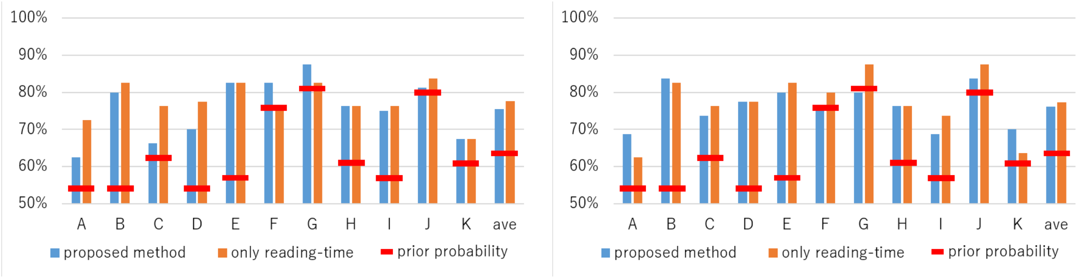
TABLE 4.12: Summary of the evaluations

Training	Testing	Classification	Detection (AUC)	
		(Accuracy)	Confident	Unconfident
Lab.	Lab.	0.76	0.89	0.78
Lab.	Wild	0.78	0.49	0.30
Wild	Wild	0.82	0.60	0.42

User-Dependency

On the laboratory dataset, we have compared the performances with two types of base lines: *reading-time only*, and *prior probability*. The former is an estimation result of the SVM using reading-time only as the feature, and the latter is a proportion of confident answers in the experiment.

Results of performances by the user-dependent training are displayed in Figure 4.26 (a). The average accuracy of the SVM is 75.6 %. According to Welch's t-test, it is significantly better than the prior probability ($p < 0.01$). Results of performances by the user-independent training are displayed in Figure 4.26 (b). The average accuracy of the SVM is 76.1 % significantly better than the prior probability ($p < 0.01$). In participant F, the accuracy was lower than a prior probability. It implies that behaviors of the participant in answering were different from others. Therefore, it was sometimes estimated wrongly due to be affected by other participants' data.



(A) user-dependent training

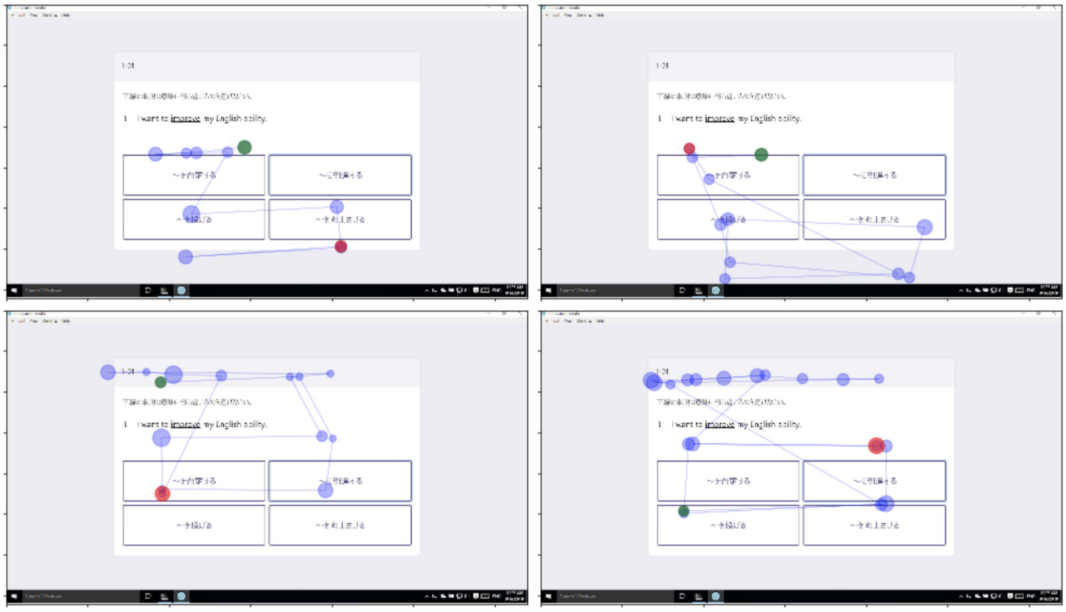
(B) user-independent training

FIGURE 4.26: Accuracies of the confidence estimation by (a) user-dependent and (b) user-independent training. The average accuracies of each method are 75.6 %, 77.6 % and 63.5 % in user-dependent and 76.1 %, 77.3 % and 63.5 % in user-independent.

In almost of all participants, we could obtain higher accuracy than prior probability. However, our proposed method could not outperform another base line (reading-time only). One assumption about this result is that the lack of the number of training samples. Especially on user-dependent training, 79 samples might be too small to find characteristics of eye movements representing confidence. The user-independent training using 790 samples performed better although each participant should have individual reading preferences. In summary, there was not significant user-dependency on our proposed system according to the evaluation of the laboratory dataset.

Limitation of Eye Tracking in the Wild

In a real learning scenario, we are not able to ask students to calibrate an eye tracker many times. They frequently move a head and change a seat position. Therefore eye gaze in the wild dataset was not precise compared to data in the laboratory. Figure 4.27 shows examples of shifted eye movements. It causes problems in our feature calculation because AOIs are predefined as absolute coordinates on a display. However, an interesting finding from scan path images is that a relative positional relationship between gazes on a question and choices is still correct even if they are shifted. In order to solve this issue, we decided to define AOIs with a new approach. From all fixations in one recording, we calculate the maximum and the minimum x and y coordinates. Then AOIs are defined on the basis of relative positions in this space. In our question format, an area of question is the 34 % top part of the space, and areas of questions are divided into a cross of the remaining 66 % bottom part.



1-21

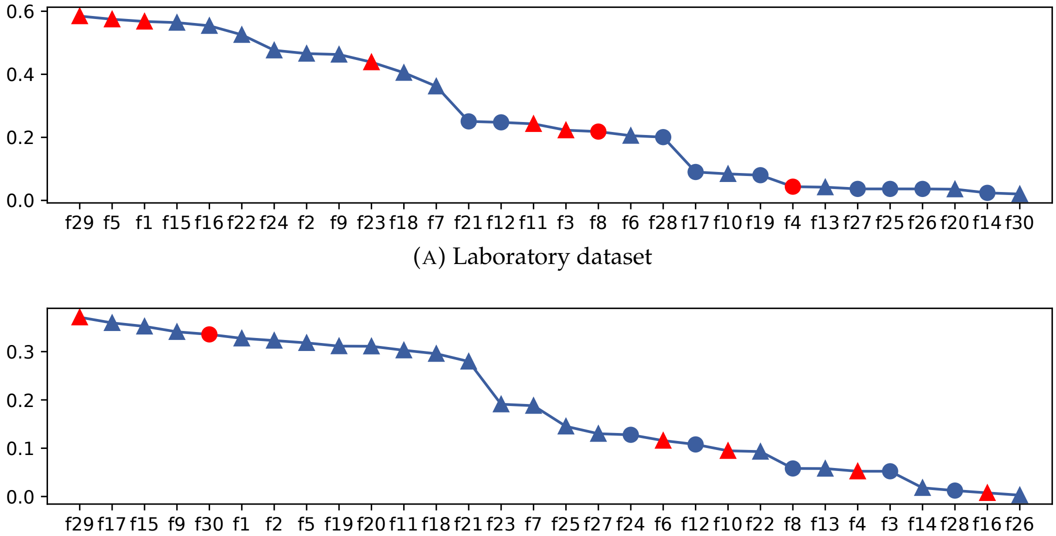
1. I want to improve my English ability.

FIGURE 4.27: Calibration problems in the wild dataset

Effective Features

Figure 4.28 shows a list of features selected on the laboratory and wild recordings. In both conditions, *f29: reading-time* has the highest correlation with self-confidence, and was selected as a feature. *f4: fixation ratio on Question* was also selected in both dataset. However, there is not more overlap between the two selections. One hypothesis about this difference is because of the gap of questions. For instance, *f30: correctness of the answer* is an effective feature in the wild dataset but it has the lowest correlation in the laboratory dataset. This is because of the difficulty of the questions (see Table 4.11 (a) and find there are many samples of high self-confidences on incorrect answers).

Most of the calculated features are negatively correlated with self-confidence. This is because the longer a student takes time to consider, the more fixations and saccades on a question and choices are observed. It is interesting that a feature which is highly correlated with self-confidence is not necessarily selected in a classifier. Furthermore, a feature which is not correlated individually (e.g., *f4: fixation ratio on Question* and *f16: mean saccade length*) can play an important role in a combination with other features.



(A) Laboratory dataset

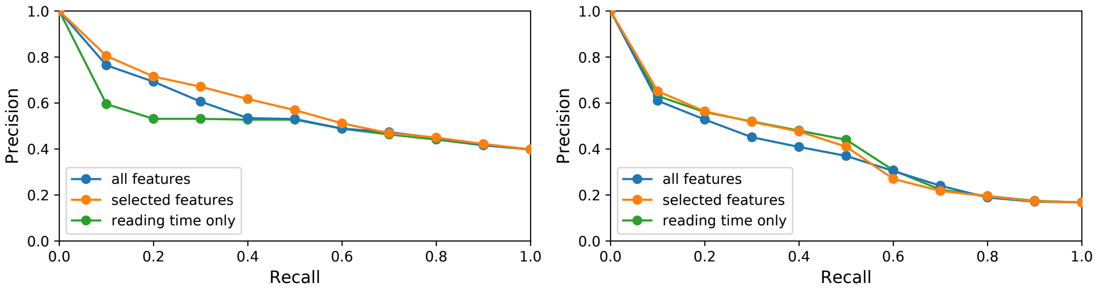
(B) Wild dataset

FIGURE 4.28: Pearson correlations between self-confidence and each feature. Features selected by the forward stepwise are highlighted as red color. (circle: positive, triangle: negative correlation; sorted by the absolute value)

Valid Metrics for the Evaluation

We are not able to evaluate our proposed method by using an accuracy on the wild dataset because the number of samples in each class is unbalanced (see Table 4.11 (b)). By considering use cases of the system, we decided to measure the performance as detection tasks of critical conditions in learning mentioned in the Introduction: high self-confidences on incorrect answers and low sef-confidences on correct answers. Figure 4.29 shows the performances of the two conditions. Correct answers and incorrect answers are used for training one classifier, but predictions were applied individually on correct answers and on incorrect answers. Users can adjust a parameter for their purpose by referring to the results.

According to Figure 4.29 (a), there is a contribution of our feature selection, and it performed better than training with all features and reading time only. This result proves the benefit of our system because at least it can detect the worst condition in learning, i.e., overconfidence or misunderstanding. On the other hand, Figure 4.29 (b) shows more improvements are required towards precise unconfidence detection.



(A) Confidence detection on incorrect answers (B) Unconfidence detection on correct answers

FIGURE 4.29: Precision-Recall Curves of the self-confidence estimation trained/tested on the wild dataset.

The Number of Training Samples

Figure 4.30 shows the relation between the number of training samples and the performance. Regarding the confidence detection, the average precision increased till the number of training samples reached to 6000. Increments more than 6000 did not contribute to the improvement, but the more training samples we had, the less standard deviation of the result was obtained. Scores of unconfidence prediction were almost the same on the all condition.

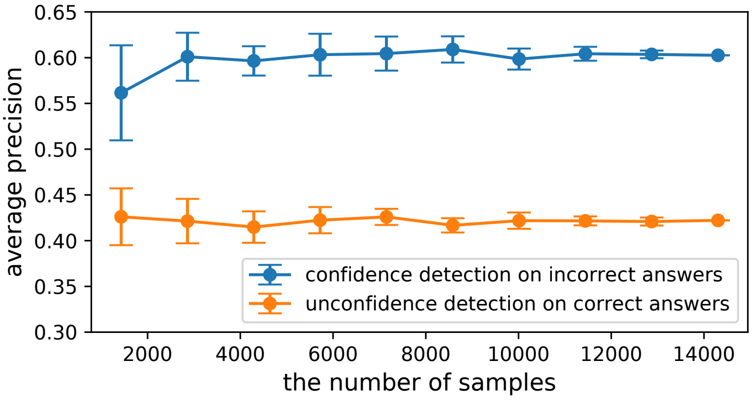
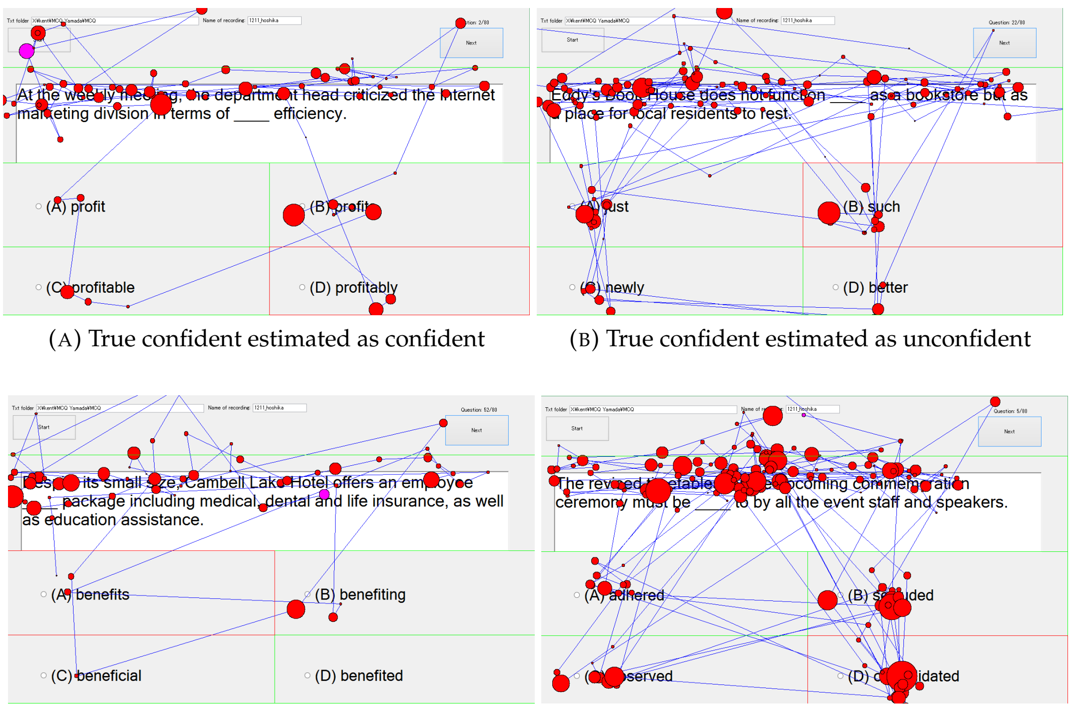


FIGURE 4.30: Average precision scores on different number of samples randomly selected from the wild dataset.

Observation of the Classified Samples

We describe differences of eye gaze between the case a participant answered with confidence and without confidence. We display on Figure 4.31 some examples of the estimation results in the laboratory dataset. The circles represent the fixations and the diameter of the circle is proportional to fixation duration. Hence the longer a participant looked at a point, the larger the diameter of the fixation is. The lines between circles represent the saccades.


At the weekly meeting, the department head criticized the Internet marketing division in terms of ____ efficiency.

○ (A) profit ○ (B) profit
○ (C) profitable ○ (D) profitably

(A) True confident estimated as confident

Eddy's Book House does not function ____ as a bookstore but as place for local residents to rest.

○ (A) just ○ (B) such
○ (C) newly ○ (D) better

(B) True confident estimated as unconfident

Despite its small size, Cambell Lake Hotel offers an employee package including medical, dental and life insurance, as well as education assistance.

○ (A) benefits ○ (B) benefiting
○ (C) beneficial ○ (D) benefited

(C) True unconfident estimated as confident

The revised timetable for the upcoming commemoration ceremony must be ____ to by all the event staff and speakers.

○ (A) adhered ○ (B) secluded
○ (C) observed ○ (D) consolidated

(D) True unconfident estimated as unconfident

FIGURE 4.31: Examples of eye gaze on each classification result

Figures 4.31 (a) and 4.31 (d) are examples of correct estimations. We can find that the confidence in answering is characterized by the fewer eye movements and smaller diameter of the fixations, on the other hand, the unconfidence is characterized by the complex eye movements and the longer fixation durations.

In Figure 4.31 (b), a participant answered without confidence, but the classifier estimated as a confident decision. We assume that he gave up to answer correctly to this question because he did not have the necessary knowledge. In such a case, we can find that the number of fixations is small and the participant took short time to answer. These characteristics are common to Figure 4.31 (a), which represents a confident decision. Therefore, the classifier estimated as a confident decision.

In Figure 4.31 (c), a participant answered with confidence, but the classifier estimated as he answered without confidence. We assume that this participant decided his answer carefully by eliminating irrelevant choice one by one. In such a case, we find more fixations and frequent transitions of eyes between rectangles. This characteristic is common to Figure 4.31 (d), which represents an unconfident answer.

4.5.4 Conclusion

We have proposed a method to estimate a self-confidence of a student in answering Multiple-Choice Questions (MCQ) by eye tracking. The method was evaluated on the laboratory dataset and the wild dataset. As results, we observed the following findings. Classifiers trained by both dataset are significantly performed better than prior probabilities. There is a contribution of eye gaze in the confidence detection on incorrect answers. Our method is not significantly user-dependent. Effective features for self-confidence estimation depends on the task, but reading-time and the fixation ratio on a question were selected on both conditions. Depending on a task, AOIs based on relative coordinates solves calibration issues. The average precision increased until the number of recorded learning behaviors reached to 6000.

As future work, we will apply our method to different kinds of subjects involving Mathematics, Science, Society and so on. We expect successful estimation of self-confidence in an MCQ which a student can answer just looking a display and thinking about a question. Moreover, we aim to apply our method to questions which do not include choices. In this work, designing AOIs for a question and each choice has been related to obtaining some effective features. We need to find new features to solve this problem.

4.6 Confidence Recognition on Spelling Tests

This section proposes a method estimating self-confidence on vocabulary spelling tests. Many learners memorize words in their non-native language by completing spelling tests repeatedly. In this approach, learners typically solve tests, check the correctness of their responses, create a list of unacquired words, and complete tests of the words in the list again. However, this approach may not be sufficiently comprehensive, and it may be helpful for learners to review words that were answered correctly to prevent the incremental forgetting of the correct spelling over time. We believe that self-confidence provide an fruitful information for selecting which words should be reviewed in addition to unacquired word.

Yu *et al.* asked participants to solve the questions which have different workload and estimate their workload by handwriting behavior. They extracted speed, pressure and stroke length of each stroke as the features. Ugurlu *et al.* analyzed the user′s emotional change by collecting the written characters in solving homework questions and exam questions [178]. Since the features extracted by written characters or the stroke lengths may depend on the questions, we use only the stroke-level features. Kishi *et al.* proposed a weak point detection algorithm by using writing interval features [101]. Asai and Yamana estimated the user′s frustration and to-be-forgotten items based on the stroke-level handwriting analysis [6, 7]. Although the feature we use is similar to theirs, the task is totally different from ours.

The closet to this research topic is work by Maruichi *et al.* [125]. They estimated self-confidence on spelling tests by analyzing typing. We follow their research and extend the sensing approach closer to natural learning behavior: hand writing.

We propose an algorithm that is question independent, specialized with self-confidence estimation of vocabulary questions, and only with the stroke-level handwriting logs. Since handwriting recognition has been already investigated by many researchers [48, 132, 155], detecting the correctness of an answer by recognizing written characters are not included in our research.

4.6.1 Approach

The proposed method consists of the following four steps: data recording, feature extraction, feature selection and classification.

Recording

We record four types of information by using a stylus pen and a computer: event type (press, release or move), timestamp of the event, the x-y coordinates of the event, and the function of the event (writing or erasing).

Feature Extraction

We extract five categories of features about answering time, writing interval, writing speed, writing pressure and erasing ratio. The detail of the features is shown in Table 4.13. In this approach, one stoke is defined as a series of action from a user touches a screen with a pen until releases the pen from the screen. The writing intervals refer to the latency between each stroke. The writing speeds are calculated by average speeds of each stroke.

TABLE 4.13: The list of extracted features

Category	No.	Features
Ans. time	f1	answering time / total stroke length
Interval	f2-f8	{ave, var, max, min, med, first, last} interval
	f9-f11	{first, last, sum} interval / answering time
Speed	f12-f16	{ave, var, max, min, med} speed
Pressure	f17-f21	{ave, var, max, min, med} pressure
Erase	f22	the number of erases / the number of strokes

Feature Selection

Using many features does not always increase classification performance. We use forward stepwise selection to select effective features. We divide training samples into five sections and conduct cross-validation by adding a feature and calculating the average accuracy. We then select the feature set associated with the highest accuracy of all average accuracy values. This step is conducted only for the training data. Preliminarily selected features are used to classify unknown samples.

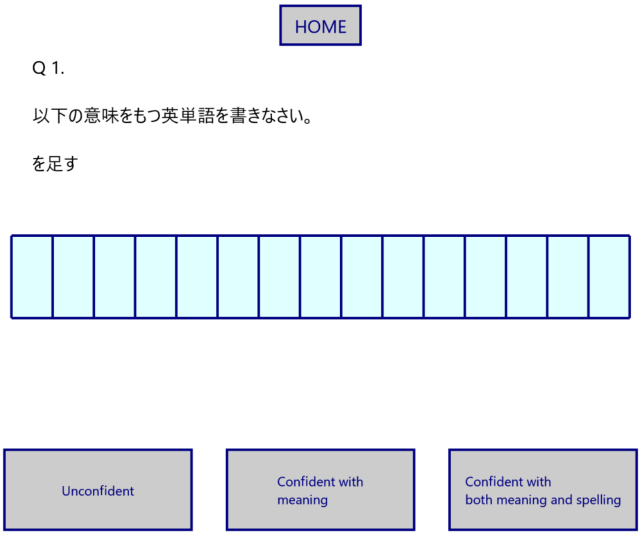
Classification

We utilize a Support Vector Machine (SVM) with a radial basis function kernel and hyper parameters: ($C = 1.0, gamma = 0.045$) to classify each sample into two classes: confident or unconfident.

4.6.2 Experimental Design

We recruited 11 university students studying in Germany as participants (three males from Japan, seven males from China, and two females from China). Participants were presented with a Japanese or Chinese word (depending on their primary language), and completed the English translation into the designated blanks, as shown in Figure 4.32 with lower case characters. We instructed participants to write one character in each blank with a block letter if possible, so that we could recognize the characters easily. Participants could skip the problem if they did not know the answer. In such cases, we were not able to record the handwriting behaviors for the question. Participants solved as many questions as possible within 1 hour. Questions were selected from a pool of vocabulary terms frequently used in Test of English for International Communication (TOEIC), as in the typing condition. The number of characters was between 3-12 for each question. Microsoft Surface Studio was used to display questions and record solving behaviors.

After solving each question, participants reported their self-confidence about their answer. They had the following three options: (1) Feeling unconfident about both the spelling and the answer, (2) Feeling unconfident about the spelling but confident about the answer, (3) Feeling confident about both the spelling and the answer. We considered both (1) and (2) as unconfident and (3) as confident for the ground truth label of the self-confidence estimation. We did not estimate self- confidence for spelling (i.e., the classification between (1) and (2)) because the number of samples labeled as (2) was too small, compared with the others, for training the model.


Q 1.

以下の意味をもつ英単語を書きなさい。

を足す

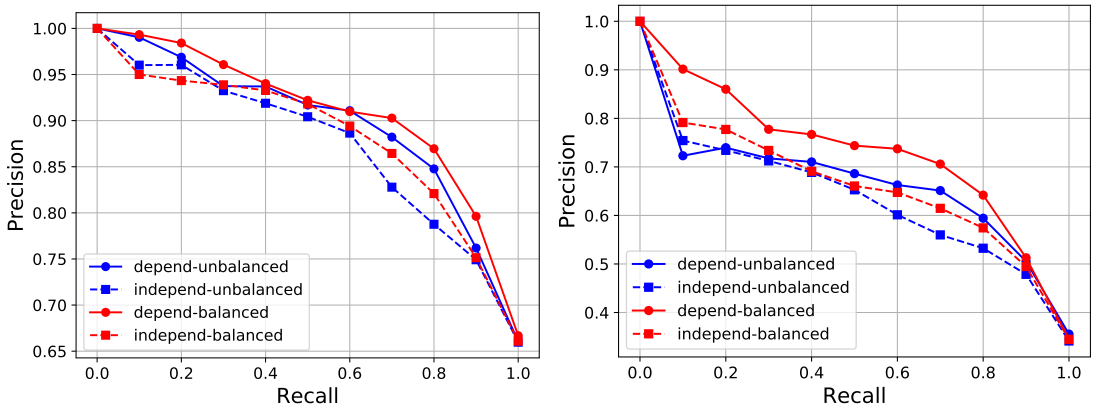
FIGURE 4.32: A screenshot of the handwriting vocabulary test

We compared the performance of the proposed method by using two different evaluation methods: user-dependent and user-independent. For the user-dependent evaluation, a part of the users' data was used for training and the other part was used for testing the performance. We selected the features in a user-dependent manner, then conducted 10-fold cross-validation. For the user-independent evaluation, data from one user were used for testing and data from all other users were used for training. We selected features user-independently, then ran a leave-one-participant-out cross-validation. Questions in the testing data were excluded from the training data to prevent over fitting. In both cases, we calculated an average accuracy among all participants, which is weighted by the number of questions. Since the dataset is not balanced, we applied to undersample with balanced bagging algorithm [75]. We report both performances with the undersampled and the non-undersampled classifier in order to evaluate the importance of the undersampling.

4.6.3 Results and Discussion

TABLE 4.14: Self-confidence estimation results

Training	Sampling	Classification	Detection (AUC)	
		(Accuracy)	Confident	Unconfident
baseline		0.66 ± 0.10	0.66	0.34
user-dependent	unbalanced	**0.80** ± **0.06**	0.90	0.64
	balanced	0.79 ± 0.04	**0.91**	**0.73**
user-independent	unbalanced	0.74 ± 0.07	0.87	0.62
	balanced	0.74 ± 0.03	0.88	0.65



(A) High confidence detection

(B) Low confidence detection

FIGURE 4.33: Precision-Recall curve

The average and standard deviation of accuracy and the Area Under the Curve (AUC) of each class is shown in Table 4.14. The baseline is defined as the ratio of the majority class. The user-dependent used unbalanced dataset achieved the highest performance of all methods. However, our dataset is too much unbalance in some participants. Thus, we also have drawn the precision-recall curve for each method and each class in Figure 4.33. Comparing the two methods, the AUC of high confident detection not so much changed. However, the AUC of low confident detection is better when using balanced training dataset. We think we had better use balanced training dataset.

For the user-independent method, the performance is not drastically changed compared with the user-dependent method. Since the dataset is basically balanced, there is not so much difference between balanced and unbalanced training.

Selected Features

Table 4.15 shows selected features in the user-independent method. The number *1* refers selected feature while *0* represent the unselected feature. Figure 4.34 refers correlation between self-confidence and each feature in the user-independent method. We find there is a weak negative correlation between writing interval and confidence. In other words, lower confidence leads longer writing interval. The number of erasing has the same tendency. The features are effective to self-confidence estimation. Speed and pressure do not have so much correlation to self-confidence. However, some of them are selected. We assume that since they are no strong correlation with interval or the number of erasing, they may work effective combined with interval or the number of erasing.

TABLE 4.15: Selected features in user-independent method

sampling	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	*f*14	f15	f16	f17	f18	f19	f20	f21	f22
unbalanced	1	1	0	1	1	1	0	1	0	1	1	1	0	0	1	0	1	0	0	1	1	1
balanced	0	0	1	1	1	1	0	0	0	1	1	1	0	0	0	0	1	1	1	0	0	0

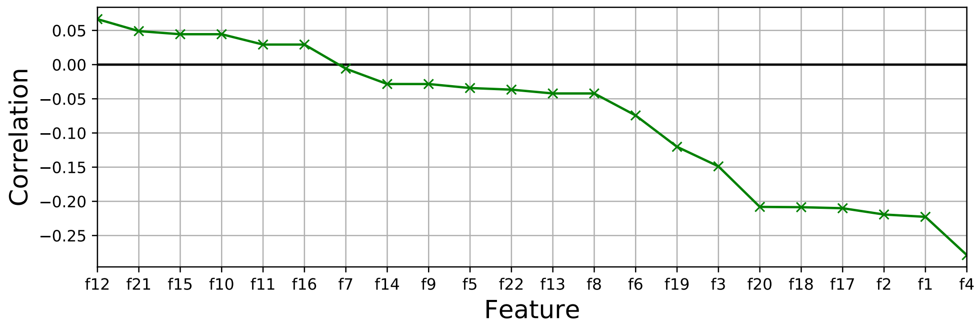


FIGURE 4.34: Pearson correlation between self-confidence and each feature in user-independent method

The Number of Training Samples

For the user-independent method, we tried to regulate the number of training samples to make it clear how much samples required for the good estimation. The result is shown in Figure 4.35. The performance of the system gets higher as the number of samples increases. The impact is smaller than we have expected. We find 200-400 will be enough since the accuracy is stable over 200-400 in both methods.

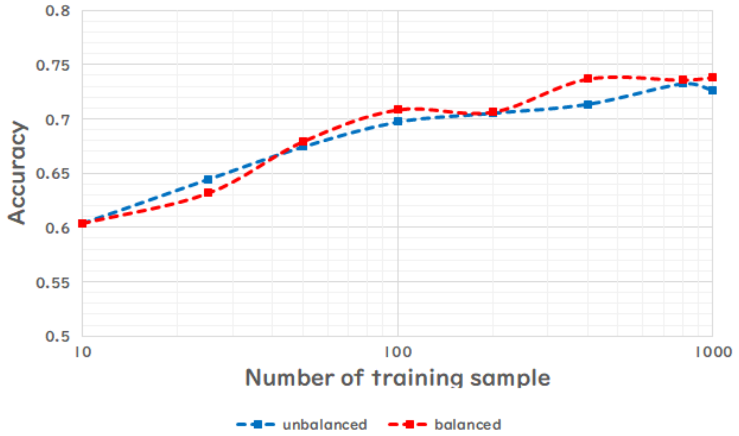


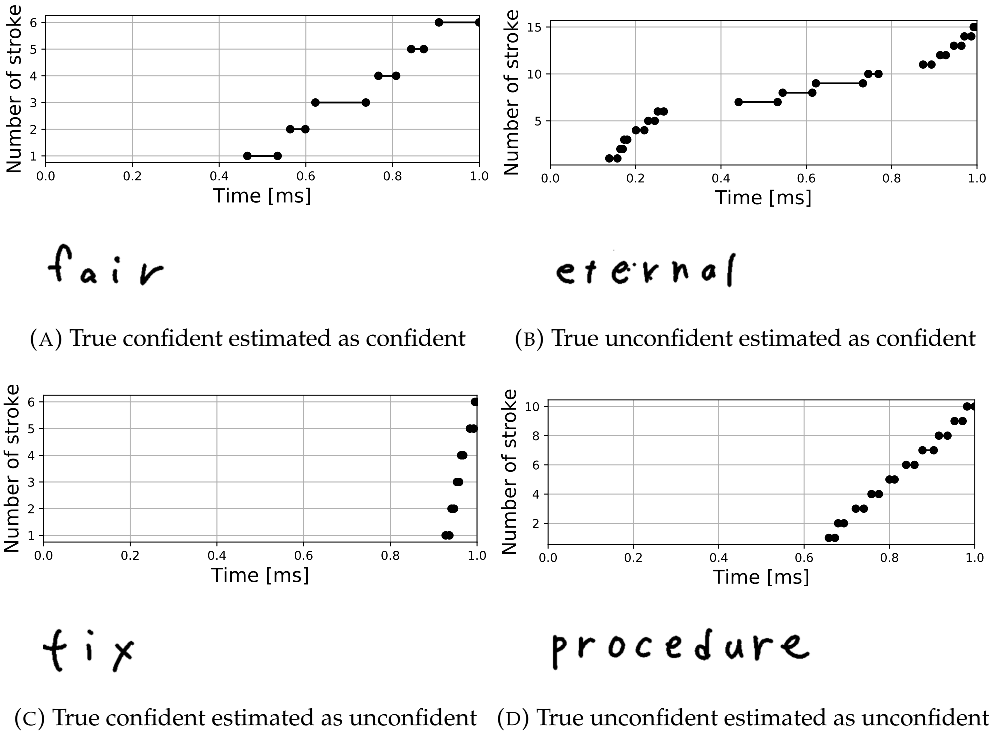
FIGURE 4.35: The number of training samples and the accuracy

Observation of the Classified Samples

We visualize some of the classified samples in Figure 4.36 to discuss the reason for the classification errors. The horizontal axis has normalized by the answering time. The string shown at the bottom of each figure is the actual word that the participant wrote. Note that the number of stroke is not equal to the number of character if the participant used an erasing function while writing.

In Figure 4.36 (b), the user felt unconfident with the answer in actual. Since the behavior became similar to confident case in Figure 4.36 (a), the classifier predicted the questions as high confident. In Figure 4.36 (c), the user was thinking or might take a short rest before solving and the interval got too long. Therefore the classifier predicted the questions as low confident.

These kinds of classification errors were caused due to our method highly depends on the writing behavior, especially on the writing interval (for instance $f4$ in Table 4.13 and Table 4.15). It is hard to solve these by current features. Therefore, we have to find another feature which is as effective as the writing interval features.



(A) True confident estimated as confident (B) True unconfident estimated as confident

(C) True confident estimated as unconfident (D) True unconfident estimated as unconfident

FIGURE 4.36: Examples of stroke intervals in each class

4.6.4 Conclusion

We have proposed a self-confidence estimation method based on the stroke-level handwriting analysis. The accuracy of our method is 80 % in an user-dependent case, and 74 % in an user-independent prediction. Our method uses only question-independent features. We found that self-confidence is negatively correlated with writing interval and the number of erases. Limitations of this study are the unbalanced dataset and the high dependency on writing interval. Future work includes increasing the number of participants, applying our method to other types of questions, and evaluating the performance of confidence-aware review feedback.

Chapter 5

Intervention in Reading and Learning

Every human has different preferences in reading. For example, some people need details about a background for further understanding while the others do not require. It depends on who reads what. However, documents have traditionally been static. We believe that reading experiences should become more immersive and interesting if documents behave differently for each individual reader. The combination of an digital document and activity recognition (e.g., recognizing interest, workload, and comprehension) makes it possible to provide dynamically-generated content individually optimized for each reader and context.

This chapter presents work towards realizing such an interactive document. Section 5.1 demonstrates *HyperMind*: the interactive digital textbook as an application utilizing affective states measured by the sensors [81]. This thesis includes an evaluation of an early prototype but it improved reading experiences of students. Section 5.2 presents a graphical user interface to create the textbooks [84]. Digitalizing reading materials should effect in both positive and negative way. Section 5.3 reports how reading activities are influenced by media: paper or screen. Comprehension tests, an eye tracker, and a physiological sensing wristband were used for the assessment [20].

The following sections are based on collaborative work with students. Section 5.2: Ko Watanabe and Section 5.3: Iuliia Brishtel.

5.1 HyperMind Reader: Intelligent Digital Textbook

Various school subjects are aligned with contents which are captured in textbooks. Although curiosity is an important factor for learning, every student has a different way of learning based on individual speed and preferences, textbooks have traditionally been found to be static and consistently dull for a variety of learners. Therefore, students sometimes avert their eyes from reading a textbook because it is boring. One of the solutions to this problem is to apply Human-Document Interaction in textbook reading, i.e., to develop a digital textbook which can make the materials for learning and instruction dynamic and anticipating on display. For instance, in Physics, it is highly efficient to show phenomenon, experiments, representations and 3D models as dynamic contents using multi medias.

The idea of making texts *dynamic* has been originally proposed by Biedert *et al.* as *Text 2.0* [15]. They have created a framework to construct gaze-responsive real time interactions to enhance the reading experience (e.g., displaying images, translations, footnote, and bookmarks). But even if the various interactions are actuated by each reader's reading behavior, the augmentations are same among all readers. The motivation of our study is to make texts not only dynamic but also *anticipating*. We measure the affective states of learners including their interests, workloads, and comprehensions, then augment the text by providing individualized information to enhance their learning abilities.

This section demonstrates the architecture and a initial study investigating positive and negative effects of the augmentation to students' learning experiences.

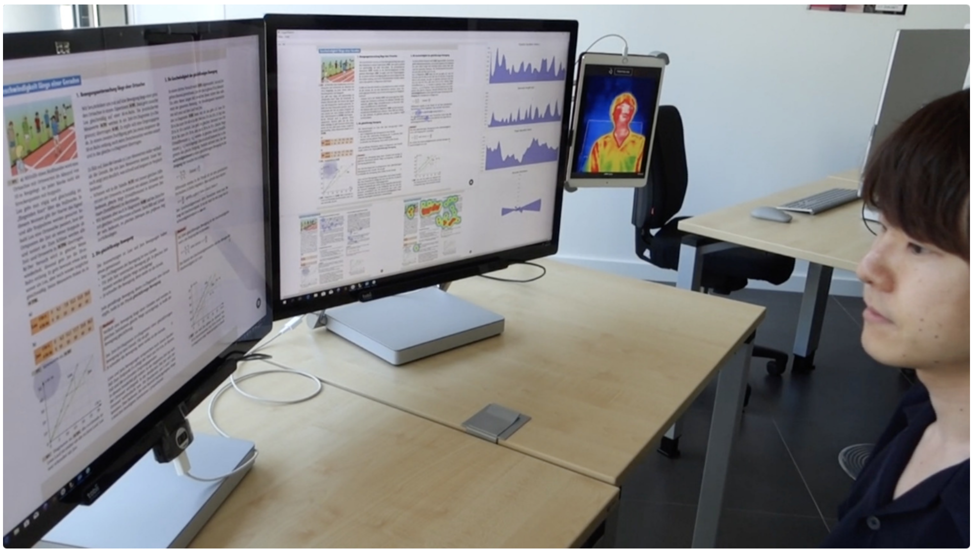


FIGURE 5.1: An overview the intelligent digital textbook: HyperMind

5.1.1 Architecture

We utilize sensors to recognize affective states and to give feedback to students (intervention) and teachers (visualization). This subsection explain the details of the technical background.

Intervention

As a prototype, we have implemented a system which displays supplemental learning materials on a textbook on the basis of eye movements. *When*, *what*, and *how* should be displayed are important problems. *When* and *what* are investigated as affective states recognition mentioned in Chapter 4. Therefore we focus on *how* the additional learning materials should be shown on the limited space of the display.

We consider a case a video will be displayed when a student has trouble to understand a figure or a description. Initially we developed the following two interfaces: *replace* and *popup*. *Replace* is an idea overwriting an original content by a supporting material. The advantage of this idea is that an additional space is not required. However, those who could or was trying to understand do not satisfy this idea. Even if the system asks students whether the content should be replaced, it disturbs natural reading behaviors. *Popup* is an approach overlaying a supporting material but not on an original content. Students can read both contents side-by-side. In this case, we need to discuss on which place the material should be displayed. Overlaying on the center of the display hides other contents. Keeping one space (e.g., the corder edge of a page) as blank and displaying all popups on the space does not disturb reading but it causes a split attention effect.

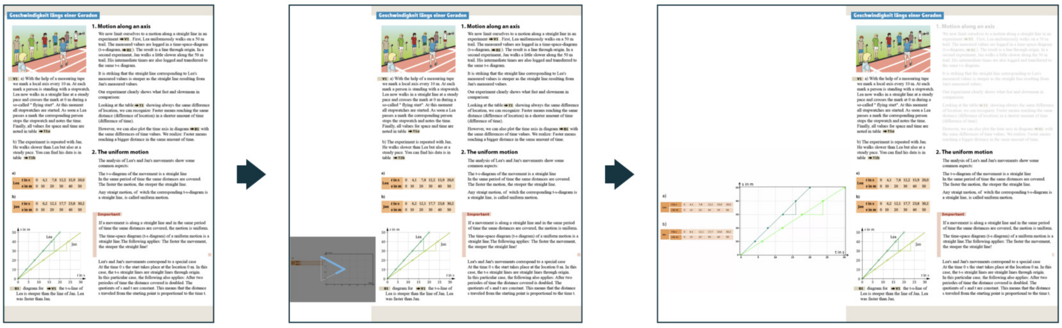


FIGURE 5.2: A gaze-oriented interaction on a textbook

Finally we designed a sliding popup approach shown in Figure 5.2. When the system finds a demand of an assistant, it shows a thumbnail image of the video on the margin of the textbook. If a student keeps watching the thumbnail, the size of the material will be increased enough to be read. If the material involves a video or sound, it starts automatically. While the material is displayed, the original contents are shifted horizontally to create a space for the augmentation. If the attention of the student moves to parts not related to the supporting material, the textbook shifts again to the original position.

Visualization

Affective states of a student is visualized in real time as shown in the right screen in Figure 5.1. Therefore teachers can monitor the process and find anomalies while reading. Figure 5.3 shows some heat maps generated by eye movements and nose temperature for examples. The fixation duration based heat map represents on which part of the textbook the student read with high attention. The regression based heat map give insights about the potion hard to understand. The nose temperatures can be visualized as not only a time-series signal but also as a heat map by synchronizing with eye gaze using the time stamp.

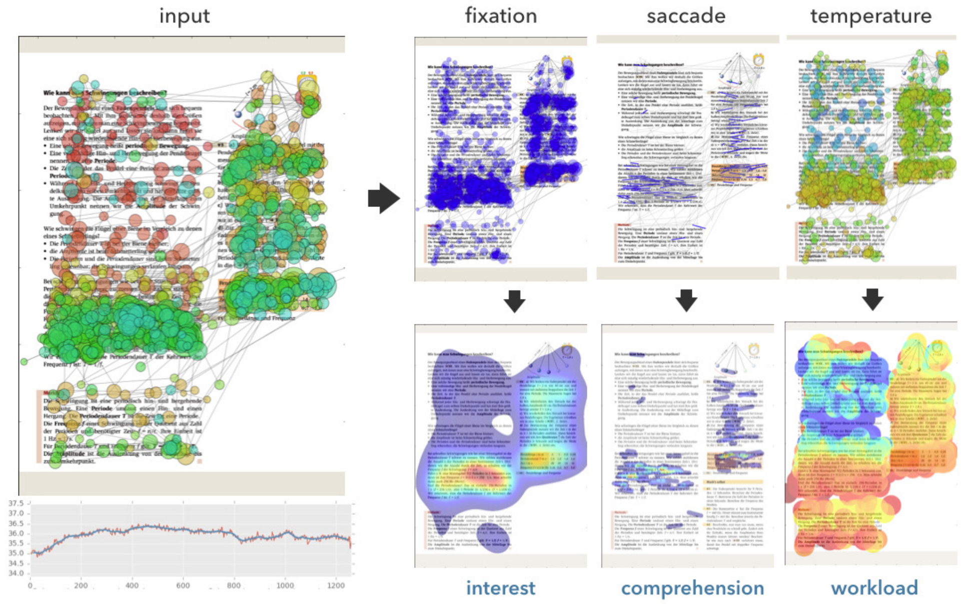


FIGURE 5.3: Combining eye tracking and thermography to recognize affective states

5.1.2 Experimental Design

We evaluated an early prototype of the proposed system. We asked 32 high school students in Germany (14±1.2 years old) to participate in an experiment. In order to distribute expertise, we recruited them from two grades: 12 of them are 8th grade and 20 students are 10th grade. They were separated into two groups reading following two textbooks and answering exercises and surveys.

We prepared two textbooks in Physics (Kinematics as an easy subject and Electromagnetism as relatively difficult subject) in two conditions (static and dynamic). The static textbooks did not have any interactions, and the dynamic textbooks had interactions mentioned above. Participants read one subject on a static textbook and another subject on a dynamic textbook. The order of subjects and conditions were randomized in all recordings. The time limit of one reading was 7 minutes. Tobii 4C 90 Hz remote eye tracker was used for interactions. We calibrated the eye tracker before each reading.

After reading a textbook, participants solved exercises related to the content and answered to surveys shown in Table 5.1. We compared their performances and feelings in the two conditions by analyzing the answers.

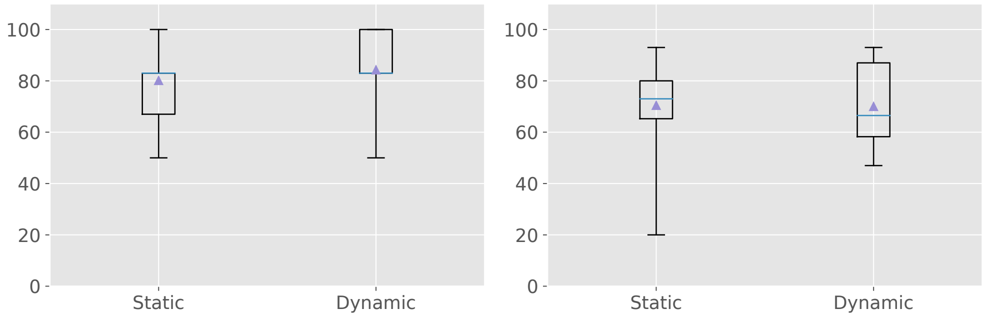
TABLE 5.1: Twelve surveys

No.	Survey
s1	I could follow the text well.
s2	I could reproduce the main contents of the text.
s3	The reading of the text ran smoothly.
s4	The illustrations were helpful to understand the text.
s5	The tables were helpful to understand the text.
s6	The design of the text page was appealing.
s7	The design of the text page made it easier to understand the text.
s8	I could well imagine what was described in the text.
s9	The textbook was difficult.
s10	Understanding the textbook has created problems for me.
s11	I needed more time to read the textbook text.
s12	I had to make an effort to understand the content of the text.

5.1.3 Results and Discussion

Figure 5.4 shows distributions of scores of exercises in the two conditions. On a Kinematics page, there is a small increase of the mean score in the dynamic condition. The distribution moved to higher score. On an Electromagnetism page, the minimum score was increase in the dynamic condition. However, the mean value was decreased and scores were distributed widely compared to the static condition. According to this result, the dynamic assistant does not always improve the performance of learning significantly.

We discuss the reason by analyzing answers of surveys shown in Figure 5.5. The most interesting finding from the answers was about s11. Some of the participants reported that they did not enough time to finish reading a textbook in the time limit.



(A) Kinematics (easy subject)

(B) Electromagnetism (difficult subject)

FIGURE 5.4: Distributions of scores on the static and dynamic textbook. The box plot represents the minimum, 25 %, 75 %, and maximum score in each condition. The median is highlighted as a blue line and the mean is displayed as a triangle marker.

This is because the dynamic textbook included videos and most of students watched the videos from begging to the end, furthermore, some of them watched several times. Therefore they might needed to hurry in reading the last minutes or might not be able to finish reading.

On the other hand, according to s2, s9 and s12 (meanings of the questions are almost same), self-confidence about the understanding increased in the dynamic condition. This improvement does not appears in a short-term evaluation but it should motivate students to read a textbook with positive feelings.

5.1.4 Conclusion

This section demonstrated the technical overview of the intelligent digital textbook and reported a pilot study using the prototype. Although it takes longer time to read than a static textbook, a dynamic textbook has a potential to improve both of the learning performance and the learning experience.

Future work includes evaluating of the next prototypes involving other sensing modalities. The investigation about learning materials, i.e., what should the system as a supporting material display is still remaining as an important problem.

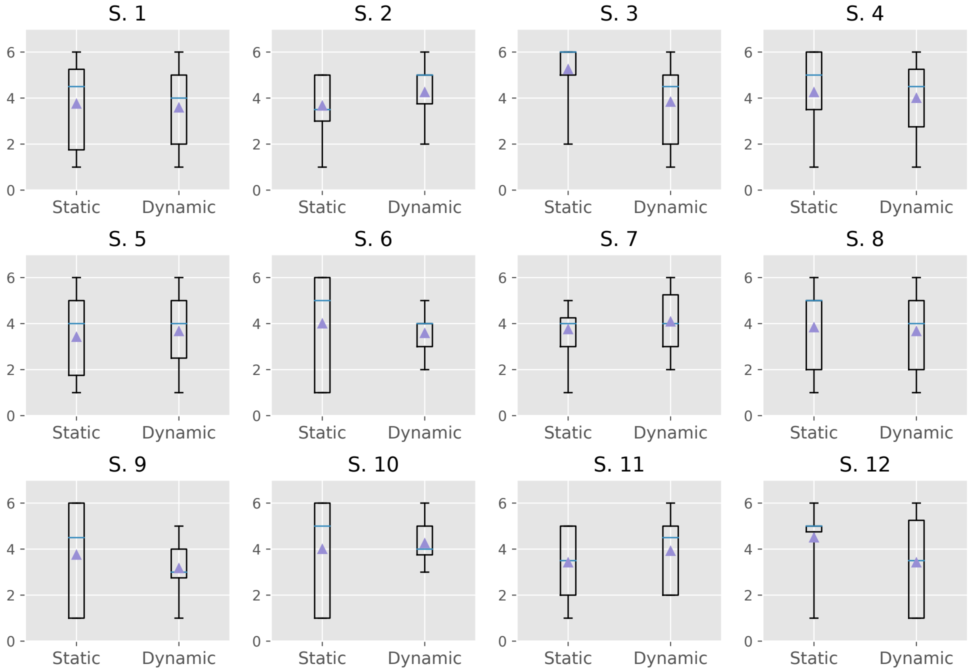


FIGURE 5.5: Distributions of survey answers asked after reading textbooks

5.2 HyperMind Builder: GUI to Create Interactive Documents

Since the appearance of human sensing technologies including eye tracking, real-time collection of reading behavior has been getting more affordable in several environments [86, 112]. These technologies have enabled researchers to design intelligent interactive documents as presented in Section 5.1. Furthermore, *Text 2.0* [15], a framework to create gaze-oriented dynamic documents in HTML and JavaScript, has helped software developers to implement interactive documents.

However, implementing interactions on documents is difficult especially for those who need it (e.g., teachers, publishers, and researchers in education). They still need helps of a person who can write programs to create interactive documents. In such a case, there is a possibility of having a discrepancy on the understanding of each other. It is difficult for teachers to explain their ideas completely, and software developers may misunderstand them.

In order to give everyone an opportunity to create intelligent interactive documents, we propose *HyperMind Builder* – Graphical User Interface (GUI) for creating intelligent interactive documents without requiring any programming skills. This section presents an overview of our proposed system, application scenario, and an initial observation to investigate further improvements. Towards our vision of the ecosystem shown in Figure 5.6, this section presents HyperMind Builder and an initial observation.

The closest concept to our system is the visual programming languages application like *Scratch* [146]. Scratch is an open-source media-rich programming environment. This application allowed many users to learn the concept of programming with an intuitive drag and drop method. It motivates many users and lowered the startup hurdle of programming. Our GUI toolkit has a similar concept, which is to lower the hurdle of creating an interactive document.

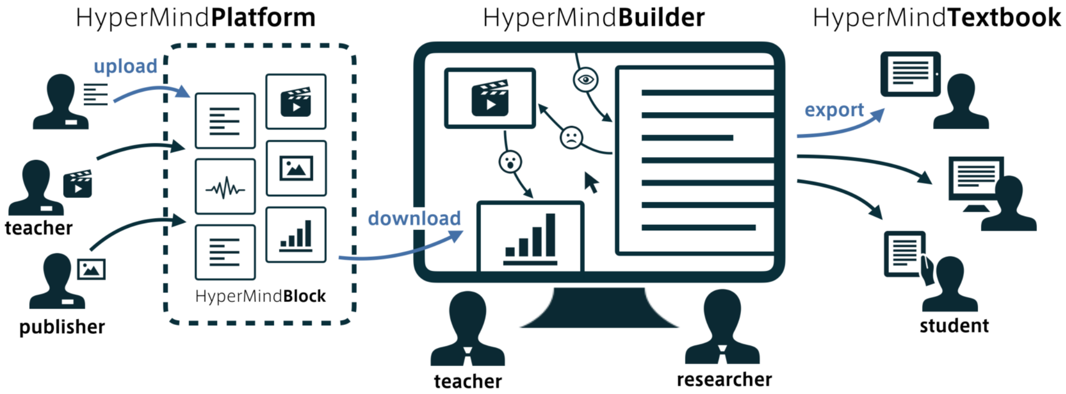
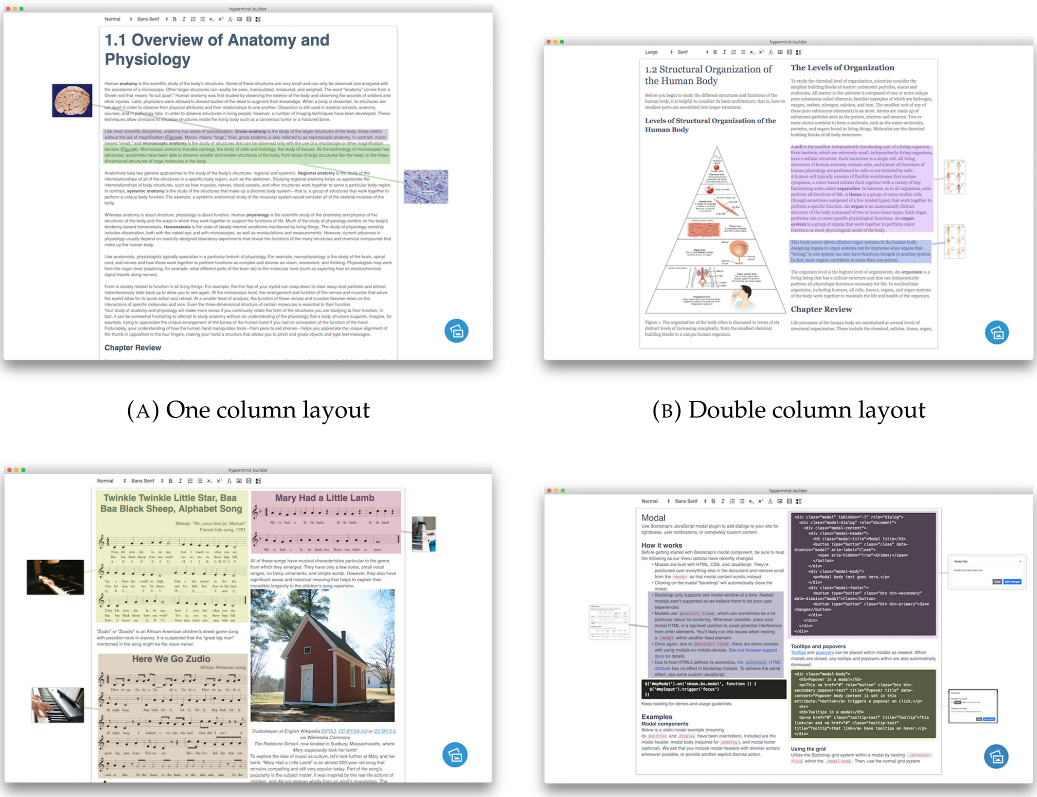


FIGURE 5.6: Overview of a work-flow in an educational scenario

5.2.1 Approach

We focus on designing a system with no programming and allowing intuitive operation for users. In our system, the screen is divided into three columns. In the central column, we provide an open source rich text editor. Since it is a WYSIWYG editor, a user can easily write texts and modify the styles or copy-and-paste texts from other shared contents. On each side of the editor, we arrange columns of material container. A user puts additional materials (e.g., images and videos related to the content) into the container by drag-and-drop. Providing columns on each side of the editor allows a user to add materials anywhere close to the context. After inserting materials, a user draws a hidden rectangle on the main content and creates relations between inserted materials and rectangles. After creating a document, a user can export and share the data with a format of *HyperMind Reader* [81]. In summary, our system requires only the writing of the content, drag-and-drop, and mouse clicks.


1.1 Overview of Anatomy and Physiology
Chapter Review
1.2 Structural Organization of the Human Body
Levels of Structural Organization of the Human Body
The Levels of Organization
Chapter Review
Twinkle Twinkle Little Star, Baa Baa Black Sheep, Alphabet Song
Mary Had a Little Lamb
Here We Go Zudio
Modal
How it works
Tooltips and popovers
Examples
Using the grid

(A) One column layout (B) Double column layout

(C) A document including music scores (D) A document including programming codes

FIGURE 5.7: Examples of intelligent interactive documents created on our system. Document sources are from OpenStax, MILINE Library, and Twitter Bootstrap (CC license).

Figure 5.7 presents examples of documents created on our system. The most promising use-case is for textbooks. Additional materials which will be displayed when a student is interested in (or has troubles to understand) the content should improve the learning experience. In addition, the gaze-oriented interaction is useful in several scenarios including reading a musical score or programming codes.

The additional materials around the main content will be displayed when they are required. The current implementation supports an activation based on a reader's attention. In other words, it utilizes an eye tracker and the related materials will be displayed if a reader's eye gaze is on a hidden rectangle longer than a threshold. We can register other activation rules (e.g., interest, comprehension, mental workload) if we utilize additional sensors.

5.2.2 Experimental design

In order to explore how simple and useful our system is for users, we conducted a small study. Following subsections describe the condition and the analysis results.

We asked 10 college students with an age between 20 - 29 to participate in our study. We provided sample texts, supporting materials (videos and images), and multiple-choice questions related to the text for measuring the comprehension. Tasks for the participants were (1) to create an interactive document, (2) to read a document created by another participant, and (3) to solve multiple-choice questions. Before starting the tasks, we gave an instruction of our system to the participants. After the task, they answered surveys of NASA-TLX [73] and two free-writing questions.

5.2.3 Results and Discussion

Participants put 5.4 ± 2.1 supporting materials on a document. Figure 5.8 shows the result of NASA-TLX. From the result, we calculated weights by pair comparison on each factors as shown in Figure 5.9. This figure infers that higher the weight of factor, it corresponds to a cause of a high work-load on each participant during the task. For instance, for Participant 9, *Performance* was the highest weight recorded. Hence, our system must improve *Performance* according to this participant. Overall result infer that *Mental Demand*, *Effort*, and *Temporal Demand* are factors that can be improved, and *Physical Demand* seems to be lower.

Regarding a free-writing question: "How was the usage of our system?", we obtained some answers like "I was a bit confused until I saw the example" or "Drag and drop were a bit confusing". We utilized the drag-and-drop because it is an intuitive function but we should consider preparing other options such as selecting from a list. But overall, we received several positive feedback including "It was really intuitive and useful" or "The usage of the system was straightforward and easy”. We also asked participants about improvements: “Do you have any idea of the additional function for the system?” but there was any feedback related to the concept of the system.

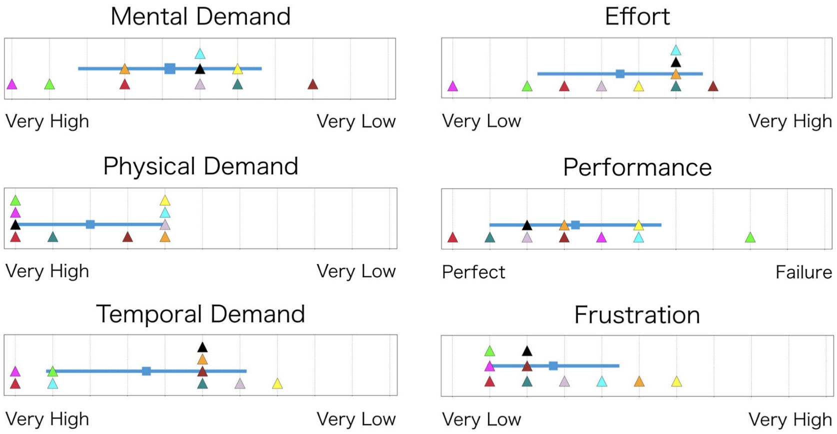


FIGURE 5.8: Results of NASA-TLX

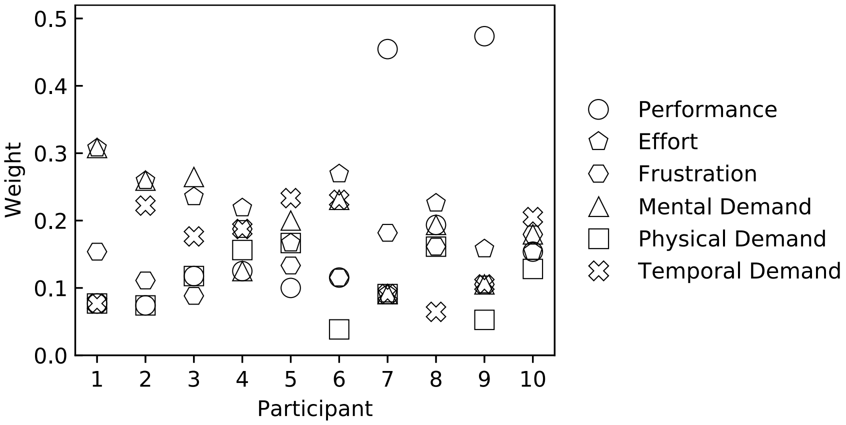


FIGURE 5.9: Weight of factors

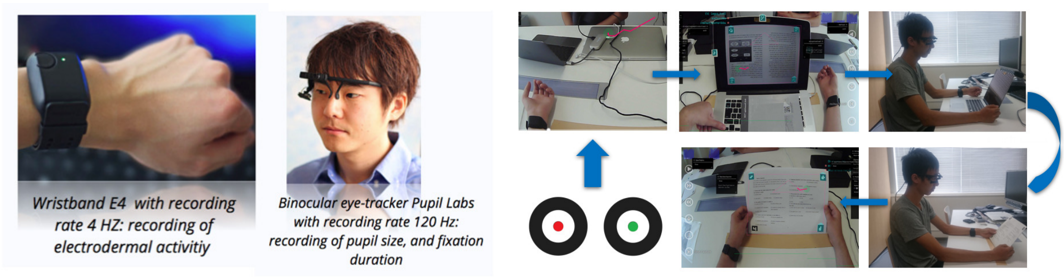
5.2.4 Conclusion

We have implemented HyperMind Builder: GUI for to create intelligent interactive documents. From the observation, we overall proof the friendliness of a toolkit. Our next aim is to identify an effective activation rules for interactions. Thereby, we can add a new function of allowing creators to design not only what but also when and how supporting materials are be displayed.

5.3 Mental Workload Assessment on Digital Media

According to *Global Market Insights*, market for Learning Management System (LMS) grows every year at a 5 % and is expected to reach approximately 240 billion USD by 2030. In order to provide appropriate means for knowledge transfer, the design of e-learning environment plays the key role in educational process [53]. However, previous research pointed on considerable differences in information processing on paper and electronic surfaces [78]. The results have shown significant advantages for learning processes on printed media comparing to its digital counterpart. For the user interface design, this fact means a demand for deeper understanding of more specific cognitive processes by users on the one hand and implementation of new evaluation methods to access these processes on the other hand.

In this section, we conduct an experiment assessing cognitive workload of participants by using four different measures: result of multiple-choice questions, average pupil size and fixation duration, which were found to be reliable indicators of cognitive workload [193] and tonic component of electrodermal activity (EDA), which is relatively new approach in this research area. In summary, we could obtain significant differences by comparing these four variables in paper and screen conditions.


Wristband E4 with recording rate 4 HZ: recording of electrodermal activitiy
Binocular eye-tracker Pupil Labs with recording rate 120 Hz: recording of pupil size, and fixation duration

(A) Sensors

(B) Procedures

FIGURE 5.10: An overview of an experimental setup.
After calibrating an eye tracker, participants read documents on screen and paper.
We avoided order effects of the two tasks by dividing participants into subgroups.

5.3.1 Experimental Design

Figure 5.10 shows an overview of the experiment. Two types of sensors were used to measure cognitive workload: E4 wristband and Pupil Labs wearable eye tracker. E4 wristband with recording rate of 4 Hz was placed on a non dominant hand of participants and switched on at the begin of each reading task. The binocular Pupil Labs eye tracker with recording rate of 120 Hz collected fixation duration and and average pupil size. For the reading task on screen, we used a 15-inch retina display. Lightness both in the room and on the screen was kept in the same state. The standardized distance between the used media and participants was 30 cm.

Eighteen computer science students from France, Japan, Ireland and Italy with intermediate to fluent English level and an age between 21 and 27 years participated in our study. Four of them used contact lenses to correct vision. For participation, they received a compensation in value of 1,000 JPY.

For the experiment, two passages with six related multiple choice questions were taken from two scientific texts with the same difficulty level and length. The order of passages was randomized between participants. We randomly divided participants into two groups (nine participants per group) to avoid any order effects of the used media as following. Participants in the paper first group started with a printed document and after reading solved a multiple-choice test presented on paper (*paper condition*). Then, the second document with subsequent multiple-choice question was presented on the screen (*screen condition*). The screen first group followed the same procedure in the reversed order as shown in Figure 5.10 (b). The instruction was to read documents as quick and as attentive as possible. The time limit for each document was 7 minutes and 30 seconds and no time limit for solving tests.

5.3.2 Data Analysis Approach

Multiple-Choice Question for Understandably Measurement

We combined all answers from paper and screen conditions into two groups. Consequently, group means were calculated and analyzed with the Wilcoxon-Mann-Whitney since the data did not satisfy requirements of t-test.

Average Pupil Size and Fixation Duration Processing

Average pupil size and fixation duration were extracted by Pupil Labs software. In the preprocessing stage, raw data was filtered by 10 Hz low-pass filter and then controlled for outliers. In the next step, we calculated individual means for both variables in paper and screen conditions and run t-test on individual level in both groups. Finally, average pupil size and fixation duration from paper conditions were taken as a baseline for calculation of relative changes in these variables in related screen conditions.

EDA Processing

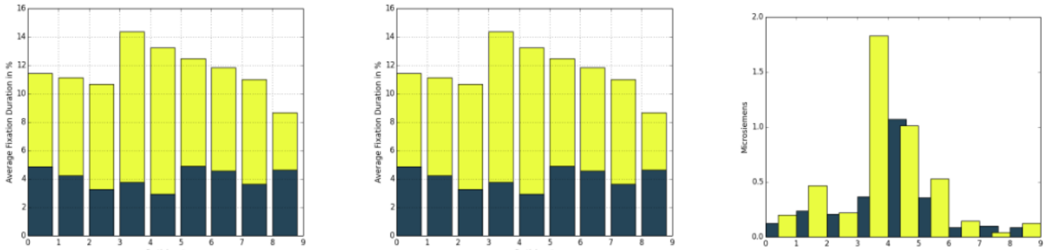
Electrodermal activity (EDA) relates to the Sympathetic Nerve System (SNS) and increase in physical, emotional or affective state can be obtained in rising of EDA signal. Tonic component is one of electrodermal measures which activity is associated with internal information processing [42]. It was processed by the method proposed by Greco *et al.* [66]. The data was filtered by 2 Hz low-pass forward-backward digital filter and then tonic component was extracted. In the next step, we inspected data for outliers. Then, individual mean of tonic component for each condition was calculated and analyzed by t-test. Finally, EDA signal in paper conditions was taken as a baseline for calculation of relative changes in EDA signal screen conditions.

5.3.3 Results and Discussion

The rate of correctly given answers in the multiple-choice question was 72.2 % while after reading documents the results were at 13.9 % worse ($p < 0.05$).

TABLE 5.2: Relative changes of variables in screen condition comparing to paper condition

Variables	The paper first group			The screen first group		
	Relative Diff.	SD	Sig.	Relative Diff.	SD	Sig.
Average Pupil Diameter	+10.71 %	0.52	.01	+10.91 %	0.60	.01
Fixation Duration	+11.64 %	1.53	.01	+11.07 %	1.60	.01
EDA (Tonic Component)	+73.38 %	23.18	.01	+74.49 / -32.90 %	7.78/18.10	.01



(A) Pupil diameter

(B) Fixatoin duration

(C) EDA tonic component

FIGURE 5.11: Differences in each index. Yellow: screen, navy: paper

Table 5.2 (and Figure 5.11 for clear visualizations for the comparison) shows the relative changes of the variables and p-values of t-test in the screen condition versus paper condition in both groups. In both groups the average fixation duration while reading on screen was significantly higher compared to reading on the paper. The same significantly increase was obtained with the pupil diameter. In the section first group, tonic component by reading on screen significantly increased in average to 73.38 % compared to the paper condition. In the screen first group, two tendencies in changes of magnitudes in tonic component were obtained: by four participants, the magnitude of tonic component by reading on screen increased in average to 74.49 %, while for five participants this magnitude decreased in average at 32.90 % comparing to the reading on paper.

The results of multiple-choice questions show significantly better performance in test solving after reading the documents on paper than on screen, which is consistent with a number of several studies [78]. The findings in pupil diameter size and fixation duration in our study are consistent with previous studies: in response to rising cognitive workload pupil diameter and fixation duration significantly increase [193]. This result is interesting in the way of natural response of pupil on the back light from computer screen: since the pupil size decreases in response on light source we obtained here an opposite effect comparing it with response on paper.

Another interesting found was done in the screen first group where some participants had significantly higher level of tonic component while reading on paper. This could be explained by the order effect: fatigue from the screen conditions could trigger an increase of EDA magnitude by reading on paper.

5.3.4 Conclusion

We have presented two contributions in the field of designing intelligent user interfaces. First, the results of this study show a significant difference in cognitive workload by treating the same information either on screen or paper. Thus, this issue should be considered by creating of user interface design, making e-learning environment less demanding for users. Secondly, our experiment shows that there are new opportunities to assess mental workload using non-expensive, simple and pervasive devices like EDA wristband.

Chapter 6
Conclusion

This chapter presents a summary of the research findings and discusses future work. Section 6.1 highlights the contributions of this thesis by recapitulating the main research problems and answers. Limitations of this work are discussed in Section 6.2 in order to design future research directions and to propose new potential questions.

6.1 Summary of the Thesis

In Chapter 1, I proposed a research hypothesis: “Meta-skills can be quantified by smart sensors" towards realizing Meta-Augmented Human Systems. In order to assist meta-skills based on the context of each user, it is necessary to understand how the activities are performed. In addition, sensing approaches should be smart, which can be utilized in daily life or in a classroom. I have proven the hypothesis by answering the following three questions, with experiments summarized in Table 6.1.

TABLE 6.1: Activities and internal states recognized in this thesis

Section	Sensor	Input	Output	Pub.
3.1	Google Glass	eye blink, head motion	activity classification	[93]
3.2	JINS MEME prototype	eye movement, head motion	activity classification	[89, 90]
3.3	JINS MEME	eye movement, head motion	reading detection	[85]
3.4	JINS MEME prototype	eye movement	word count	[83, 87]
4.1	SMI/Tobii eye tracking glasses	eye gaze	comprehension	[81, 82]
4.2	Tobii 4C, JINS MEME, Empatica E4	eye gaze, EDA, BVP, arm temp.	comprehension	[134]
4.3	SMI REDn, FLIR One for iOS	eye gaze, nose temp.	interest, mental workload	[86]
4.4	SMI REDn, Empatica E4	eye gaze, EDA, BVP, arm temp.	interest	[95, 96]
4.5	Tobii 4C	eye gaze	self-confidence on MCQ	[92]
4.6	Surface Studio	hand writing log	self-confidence on spelling	[124]

6.1.1 How can smart sensors quantify reading activities in daily life?

Chapter 3 claimed that eye movements and head movements measured by sensors on eyewear computers are able to detect daily reading activities and, furthermore, to track the amount of reading.

Section 3.1 proposed a blink detection algorithm using an infrared proximity sensor equipped on Google Glass. By combining features from eye blinks to head movements measured by an inertial measurement unit (IMU), five activities (reading, watching, solving, sowing, and talking) were classified with 82 % accuracy. The combination improved the performance. For instance, talking and watching were relatively easily distinguished by other activities by eye blinks and features from head movements helped to classify sawing, reading, and solving.

Section 3.2 proposed an activity recognition method on JINS MEME, commercial Electrooculography (EOG) glasses, which is more suitable for daily activity recognition than Google Glass. Although it was a pilot study involving only two participants, four activities (reading, typing, eating, and talking) could be classified with 70 % accuracy by using statistical features from signals from EOG and IMU.

Section 3.3 proposed natural reading detection using JINS MEME (developer version) and deep neural networks. In order to investigate the difference between *controlled reading* designed carefully in the laboratory and *natural reading* without any limitations in the wild, a large-scale experiment asking seven participants to wear sensors for more than two weeks (880 hours recording in total) were conducted. A Long-Short-Term Memory (LSTM) based network trained by the data could classify *natural reading* and *not reading* with 74 % accuracy. Since *controlled reading* and *not reading* were able to be classified with 93 % accuracy, this research highlighted the limitation of activity recognition research in a controlled environment.

Section 3.4 proposed a method estimating the number of read words by analyzing EOG signals on JINS MEME. Forward- and backward-saccades while reading were detected as peaks on the signals. With the cooperation of five participants as an experiment, the number of read words was able to be estimated with a 16 % word count error in a user-dependent training. An interesting finding from this research is that the estimation error decreases if results on some documents were summarized (to 3.0 % by summarizing results of 38 documents) because some of them were estimated more than the ground truth while the others were less.

6.1.2 How can smart sensors quantify affective states of learners?

Chapter 4 introduced the potential of eye tracking and physiological sensing in quantifying affective states while learning. The performances were not enough, but interesting features with high correlations to affective states were observed.

Section 4.1 presented comprehension recognition methods on a textbook. By analyzing the reading behaviors of high school students on a textbook in Physics, participants' scores on related exercises could be classified into three classes with 100 % accuracy using an AOI-based approach and 70 % using a subsequence-based approach (statistical features in a one-minute recording).

Section 4.2 presented an investigation on a video lecture. In the experiment, features observed by Tobii 4C, JINS MEME, and E4 wristband could not estimate whether a learner solves related exercises correctly with enough accuracy. However, this work revealed a significant correlation between the ratio of watching duration (the duration while gaze is on the screen divided by the total duration of a video) and comprehension. Learners could often answer correctly when the ratio was high.

Section 4.3 presented the results of a pilot study, finding correlations between sensor signals and affective states collected by a subjective survey after reading a textbook on physics. This work revealed a positive correlation between interest and pupil diameter ($p < 0.01$) and a negative correlation between mental workload and nose temperature measured by an infrared thermal camera ($p < 0.05$).

Section 4.4 proposed an interest recognition method while reading by utilizing SMI REDn Scientific 60 Hz remote eye tracker and E4 wristband. Features from both sensors could classify the level of interest in newspaper articles into four classes with 50 % accuracy. In this experiment, which included surveys, the most effective feature was *subjective comprehension*. Indeed, readers cannot find interest in a document that is difficult to understand. The mean value of fixation was small while reading interesting documents for most of the participants.

Section 4.5 proposed a method to estimate self-confidence when answering multiple-choice questions about English vocabulary and grammar using eye gaze. The method was evaluated through laboratory and wild datasets. On the laboratory dataset, the binary classification accuracy achieved 76 % on a user-independent training. Since samples in the wild dataset are unbalanced, the performance in the wild dataset could not be evaluated as an accuracy, but it could detect high confidence on incorrect answers (the worst case that a learner has wrong knowledge and he/she did not realize it) with 60 % average precision. The reading-time and the fixation ratio on a question were selected as effective features in both conditions.

Section 4.6 proposed a method self-confidence estimation spelling English vocabulary tests by analyzing the log of handwriting. It could classify self-confidence with 80 % accuracy on user-dependent learning and 74 % on user-independent learning. It is understandable that features related to intervals between each stroke were effective for the estimation. Furthermore, this work revealed that the pressure of a pen was negatively related to the self-confidence against my expectation.

6.1.3 How can a system augment reading/learning experiences?

This thesis demonstrated applications giving feedback to a user by using the estimation results of cognitive activities and affective states.

Wordometer 2.0 presented in Section 3.4 motivates people to improve daily reading habits by quantifying the amount of reading. Although a word-counting algorithm using mobile eye tracking glasses [110] and medical EOG sensors [109] have been proposed [9], this work was the first attempt to implement the algorithm on a smart device and give feedback on a real-time monitoring view or a dashboard view.

Confidence-Aware Learning Assistant (CoALA) presented in Section 4.5 lets learners review the results on multiple-choice questions on the basis of self-confidence in their answers. This system has been deployed in a cram school in Japan and used by more than 70 students. In this cram school, students print out a list of words involving incorrect and correct answers regarding low self-confidence.

HyperMind Reader presented in Section 5.1 is an intelligent digital textbook displaying information based on gaze, i.e., utilizing an eye tracker to measure visual attention and to employ it for vivid interaction.

HyperMind Builder presented in Section 5.1 enables everyone to create their own intelligent interactive documents without any programming skills. It is on the context pioneered by *Text 2.0* [15] and *HyperMind Builder* that have extended the range of users to those who need it, including teachers and researchers in education.

As presented in Section 5.3, paper is still comfortable media for reading compared to screen, and digitalizing learning media may not always effect reading experiences in a positive way. The fixation duration, the pupil diameter, the EDA tonic component while reading on screen were higher than on paper. The learning media should be designed carefully not to increase cognitive load.

6.2 Limitations and Future Work

There are some limitations in the presented work. In particular, there is still room for improvement or new research ideas around the following topics.

Performances of the estimations. While this thesis tried to solve challenging recognition tasks, accuracies 74 % into two classes (Section 3.3), 70 % into three classes (Section 4.1), 50 % into four classes (Section 4.4), 76 % into two classes (Section 4.5) might not be enough to claim the approaches can recognize them sufficiently.

Deep learning in learning analytics. Deep learning techniques have rapidly accelerated research in the field of image, audio and natural language processing.

However, affective state recognition are not sufficiently based on the benefits technology offers. One hypothesis of the reason is that collecting a large amount of data for training with annotations is more difficult than it is in other research fields.

Ground truth labeling. A common problem in the studies with low estimation results was the difficulty of collecting ground truth labels. Subjective labels depend on the participants. Labels answered with low self confidence may become noise while training a model. The more participants are asked to answer their states frequently, the more their activities become unnatural. Asking participants to wear a camera and labeling by others sometimes causes a privacy issue.

Evaluation of the interventions. Investigations about the performance of interventions (e.g., how many more books did people read and how much more students' academic scores did increase by proposed systems) require large-scale experiments. Due to the time constraints of the doctoral research, such long-term evaluations are out of the scope for the time being and reluctantly left for future work.

Learner as a sensor. This thesis investigated a learner by using sensors. But a learner can also be a sensor of an environment. Measuring behaviors and performances of learners for the assessment of textbooks, lectures, and classrooms should be an interesting topic. Since expensive advanced sensors cannot be utilized because of limitations of the time, the place, and the number of participants, I believe that smart sensors play important role in this field.

In this thesis, I proposed Meta-Augmented Human Systems, which amplify skills to acquire new skills. Wordometer, Confidence-Aware Learning Assistant, and HyperMind were demonstrated as examples of the systems. The common philosophy behind the applications is that understanding physical, cognitive, and affective states is the first step for giving appropriate feedback on skill acquisition. Furthermore, sensors should be designed for everyday use to track long-term improvements. By presenting several recognition algorithms utilizing devices in the market, this thesis verified the hypothesis: meta-skills can be quantified by smart sensors. My next interests are in interaction research that includes designing effective interventions.

Bibliography

[1] Y. Abdelrahman, E. Velloso, T. Dingler, A. Schmidt, and F. Vetere, "Cognitive heat: Exploring the usage of thermal imaging to unobtrusively estimate cognitive load", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 33, 2017.

[2] R. Ackerman and V. A. Thompson, "Meta-reasoning", *Reasoning as memory*, pp. 164–182, 2015.

[3] M. C. Acosta, J. Gallar, and C. Belmonte, "The influence of eye solutions on blinking and ocular comfort at rest and during work at video display terminals", *Experimental eye research*, vol. 68, no. 6, pp. 663–669, 1999.

[4] S. Ainsworth, "Deft: A conceptual framework for considering learning with multiple representations", *Learning and instruction*, vol. 16, no. 3, pp. 183–198, 2006.

[5] O. Amft, F. Wahl, S. Ishimaru, and K. Kunze, "Making regular eyeglasses smart", *Pervasive Computing, IEEE*, vol. 14, no. 3, pp. 32–43, 2015.

[6] H. Asai and H. Yamana, "Detecting student frustration based on handwriting behavior", in *Proceedings of the adjunct publication of the 26th annual ACM symposium on User interface software and technology*, ACM, 2013, pp. 77–78.

[7] ——, "Detecting learner's to-be-forgotten items using online handwritten data", in *Proceedings of the 15th New Zealand Conference on Human-Computer Interaction*, ACM, 2015, pp. 17–20.

[8] O. Augereau, H. Fujiyoshi, and K. Kise, "Towards an automated estimation of english skill via toeic score based on reading analysis", in *Pattern Recognition, 2016 23rd International Conference on*, IEEE, 2016, pp. 1285–1290.

[9] O. Augereau, C. L. Sanches, K. Kise, and K. Kunze, "Wordometer systems for everyday life", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, p. 123, 2018.

[10] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild", in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 354–361.

[11] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data", in *International Conference on Pervasive Computing*, Springer, 2004, pp. 1–17.

[12] V. M. G. Barrios, C. Gütl, A. M. Preis, K. Andrews, M. Pivec, F. Mödritscher, and C. Trummer, "Adele: A framework for adaptive e-learning through eye tracking", *Inquiring Knowledge Networks on the Web*, pp. 609–616, 2004.

[13] A. R. Bentivoglio, S. B. Bressman, E. Cassetta, D. Carretta, P. Tonali, and A. Albanese, "Analysis of blink rate patterns in normal subjects", *Movement Disorders*, vol. 12, no. 6, pp. 1028–1034, 1997.

[14] R. Biedert, G. Buscher, and A. Dengel, "The eyebook–using eye tracking to enhance the reading experience", *Informatik-Spektrum*, vol. 33, no. 3, pp. 272–281, 2010.

[15] R. Biedert, G. Buscher, S. Schwarz, M. Möller, A. Dengel, and T. Lottermann, "The text 2.0 framework: Writing web-based gaze-controlled realtime applications quickly and easily", in *Proceedings of the 2010 workshop on Eye gaze in intelligent human machine interaction*, ACM, 2010, pp. 114–117.

[16] R. Biedert, J. Hees, A. Dengel, and G. Buscher, "A robust realtime reading-skimming classifier", in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ACM, 2012, pp. 123–130.

[17] R. A. Bolt, "A gaze-responsive self-disclosing display", in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 1990, pp. 3–10.

[18] R. W. Booth and U. W. Weger, "The function of regressions in reading: Backward eye movements allow rereading", *Memory & cognition*, vol. 41, no. 1, pp. 82–97, 2013.

[19] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[20] I. Brishtel, S. Ishimaru, O. Augereau, K. Kise, and A. Dengel, "Assessing cognitive workload on printed and electronic media using eye-tracker and eda wristband", in *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, ACM, 2018, p. 45.

[21] A. Bulling, J. A. Ward, and H. Gellersen, "Multimodal recognition of reading activity in transit using body-worn sensors", *ACM Transactions on Applied Perception*, vol. 9, no. 1, 2:1–2:21, 2012.

[22] A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster, "Robust recognition of reading activity in transit using wearable electrooculography", in *International Conference on Pervasive Computing*, Springer, 2008, pp. 19–37.

[23] A. Bulling, J. A. Ward, H. Gellersen, and G. Troster, "Eye movement analysis for activity recognition using electrooculography", *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 4, pp. 741–753, 2011.

[24] G. Buscher, A. Dengel, and L. van Elst, "Eye movements as implicit relevance feedback", in *Extended abstracts of the 2008 CHI Conference on Human Factors in Computing Systems*, ACM, 2008, pp. 2991–2996.

[25] V. Bush, "As we may think", 1945.

[26] P. Casale, O. Pujol, and P. Radeva, "Human activity recognition from accelerometer data using a wearable device", in *Pattern Recognition and Image Analysis*, Springer, 2011, pp. 289–296.

[27] K. Cater, A. Chalmers, and P. Ledda, "Selective quality rendering by exploiting human inattentional blindness: Looking but not seeing", in *Proceedings of the ACM symposium on Virtual reality software and technology*, ACM, 2002, pp. 17–24.

[28] P. Chandler and J. Sweller, "Cognitive load theory and the format of instruction", *Cognition and instruction*, vol. 8, no. 4, pp. 293–332, 1991.

[29] M. Chau and M. Betke, "Real time eye tracking and blink detection with usb cameras", Boston University Computer Science Department, Tech. Rep., 2005.

[30] S.-C. Chen, H.-C. She, M.-H. Chuang, J.-Y. Wu, J.-L. Tsai, and T.-P. Jung, "Eye movements predict students' computer-based assessment performance of physics concepts in different presentation modalities", *Computers & Education*, vol. 74, pp. 61–72, 2014.

[31] S. Chen, J. Epps, N. Ruiz, and F. Chen, "Eye activity as a measure of human mental effort in hci", in *Proceedings of the 16th international conference on Intelligent user interfaces*, ACM, 2011, pp. 315–318.

[32] A. S. Chetwood, K.-W. Kwok, L.-W. Sun, G. P. Mylonas, J. Clark, A. Darzi, and G.-Z. Yang, "Collaborative eye tracking: A potential training tool in laparoscopic surgery", *Surgical endoscopy*, vol. 26, no. 7, pp. 2003–2009, 2012.

[33] A. Clark, *Natural born-cyborg*, 2003.

[34] A. C. Clarke, *Profiles of the Future*. Hachette UK, 2013.

[35] R. Clément, Z. Dörnyei, and K. A. Noels, "Motivation, self-confidence, and group cohesion in the foreign language classroom", *Language learning*, vol. 44, no. 3, pp. 417–448, 1994.

[36] L. Copeland, T. Gedeon, and S. Caldwell, "Framework for dynamic text presentation in elearning", *Procedia Computer Science*, vol. 39, pp. 150–153, 2014.

[37] D. D. Cox and R. L. Savoy, “Functional magnetic resonance imaging (fmri)“brain reading”: Detecting and classifying distributed patterns of fmri activity in human visual cortex”, *Neuroimage*, vol. 19, no. 2, pp. 261–270, 2003.

[38] H. D. Critchley, “Electrodermal responses: What happens in the brain”, *The Neuroscientist*, vol. 8, no. 2, pp. 132–142, 2002.

[39] A. E. Cunningham and K. E. Stanovich, “What reading does for the mind”, *American educator*, vol. 22, pp. 8–17, 1998.

[40] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. W. Mayol-Cuevas, “You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video.”, in *BMVC*, 2014.

[41] K. N. Daniel, E. Kamioka, *et al.*, “Detection of learners concentration in distance learning system with multiple biological information”, *Journal of Computer and Communications*, vol. 5, no. 04, p. 1, 2017.

[42] M. E. Dawson, A. M. Schell, and D. L. Filion, “The electrodermal system”, *Handbook of psychophysiology*, vol. 2, pp. 200–223, 2007.

[43] A. Dengel, “Digital co-creation and augmented learning”, in *Proceedings of the The 11th International Knowledge Management in Organizations Conference on The changing face of Knowledge Management Impacting Society*, ACM, 2016, p. 3.

[44] S. Dernbach, B. Das, N. C. Krishnan, B. L. Thomas, and D. J. Cook, “Simple and complex activity recognition through smart phones”, in *2012 8th International Conference on Intelligent Environments*, 2012, pp. 214–221.

[45] N. Dimakis, J. K. Soldatos, L. Polymenakos, P. Fleury, J. Curín, and J. Kleindienst, “Integrated development of context-aware applications in smart spaces”, *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 71–79, 2008.

[46] T. Dingler, K. Kunze, and B. Outram, “Vr reading uis: Assessing text parameters for reading in vr”, in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, ACM, 2018, LBW094.

[47] J. Dolin, “Science education standards and science assessment in denmark”, *Making it comparable: Standards in science education*, pp. 71–82, 2007.

[48] L. D’souza and M. Mascarenhas, “Offline handwritten mathematical expression recognition using convolutional neural network”, in *2018 International Conference on Information, Communication, Engineering and Technology*, IEEE, 2018, pp. 1–3.

[49] A. T. Duchowski, V. Shivashankaraiah, T. Rawls, A. K. Gramopadhye, B. J. Melloy, and B. Kanki, "Binocular eye tracking in virtual reality for inspection training", in *Proceedings of the 2000 symposium on Eye tracking research & applications*, ACM, 2000, pp. 89–96.

[50] J. Dunlosky, M. J. Serra, G. Matvey, and K. A. Rawson, "Second-order judgments about judgments of learning", *The Journal of General Psychology*, vol. 132, no. 4, pp. 335–346, 2005.

[51] S. for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures, W. Boucsein, D. C. Fowles, S. Grimnes, G. Ben-Shakhar, W. T. Roth, M. E. Dawson, and D. L. Filion, "Publication recommendations for electrodermal measurements", *Psychophysiology*, vol. 49, no. 8, pp. 1017–1034, 2012.

[52] D. C. Engelbart, "Augmenting human intellect: A conceptual framework", 1962.

[53] B. Faghih, D. Azadehfar, M. Reza, P. Katebi, *et al.*, "User interface design for e-learning software", *arXiv preprint arXiv:1401.6365*, 2014.

[54] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective", in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 1226–1233.

[55] B. Finn and J. Metcalfe, "The role of memory for past test in the underconfidence with practice effect.", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 33, no. 1, p. 238, 2007.

[56] C. Fleming, *The glass cage*, 6931. British Medical Journal Publishing Group, 1994, vol. 308, p. 797.

[57] S. M. Fleming, B. Maniscalco, Y. Ko, N. Amendi, T. Ro, and H. Lau, "Action-specific disruption of perceptual confidence", *Psychological Science*, vol. 26, no. 1, pp. 89–98, 2015.

[58] L. Fletcher and P. Carruthers, "Metacognition and reasoning", *Phil. Trans. R. Soc. B*, vol. 367, no. 1594, pp. 1366–1378, 2012.

[59] F Foerster, M Smeja, and J Fahrenberg, "Detection of posture and motion by accelerometry: A validation study in ambulatory monitoring", *Computers in Human Behavior*, vol. 15, no. 5, pp. 571–583, 1999.

[60] K. Futami, T. Terada, and M. Tsukamoto, "Success imprinter: A method for controlling mental preparedness using psychological conditioned information", in *Proceedings of the 7th Augmented Human International Conference 2016*, ACM, 2016, p. 11.

[61] U. Garain, O. Pandit, O. Augereau, A. Okoso, and K. Kise, "Identification of reader specific difficult words by analyzing eye gaze and document content", in *Document Analysis and Recognition, 2017 14th IAPR International Conference on*, IEEE, vol. 1, 2017, pp. 1346–1351.

[62] A. Gevins, M. E. Smith, H. Leong, L. McEvoy, S. Whitfield, R. Du, and G. Rush, "Monitoring working memory load during computer-based tasks with eeg pattern recognition methods", *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 40, no. 1, pp. 79–91, 1998.

[63] J. Ghosh, Y. J. Lee, and K. Grauman, "Discovering important people and objects for egocentric video summarization", in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 1346–1353.

[64] R Girwidz, T. Rubitzko, S Schaal, and F. Bogner, "Theoretical concepts for using multimedia in science education", *Science Education International*, vol. 17, no. 2, pp. 77–93, 2006.

[65] A. Greco, A. Lanata, L. Citi, N. Vanello, G. Valenza, and E. P. Scilingo, "Skin admittance measurement for emotion recognition: A study over frequency sweep", *Electronics*, vol. 5, no. 3, p. 46, 2016.

[66] A. Greco, A. Lanata, G. Valenza, E. P. Scilingo, and L. Citi, "Electrodermal activity processing: A convex optimization approach", in *Engineering in Medicine and Biology Society, 2014 36th Annual International Conference of the IEEE*, IEEE, 2014, pp. 2290–2293.

[67] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "Cvxeda: A convex optimization approach to electrodermal activity processing", *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 797–804, 2016.

[68] M Haak, S Bos, S Panic, and L. Rothkrantz, "Detecting stress using eye blinks and brain activity from eeg signals", *Proceeding of 1st Driver Car Interaction and Interface*, 2009.

[69] Z. M. Hafed and J. J. Clark, "Microsaccades as an overt measure of covert attention shifts", *Vision research*, vol. 42, no. 22, pp. 2533–2545, 2002.

[70] N. Y. Hammerla and T. Plötz, "Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition", in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, 2015, pp. 1041–1051.

[71] D. M. Hansen, "A discourse structure analysis of the comprehension of rapid readers.", 1975.

[72] C. Harrison, H. Benko, and A. D. Wilson, "Omnitouch: Wearable multitouch interaction everywhere", in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, ACM, 2011, pp. 441–450.

[73] S. G. Hart, "Nasa task load index (tlx). volume 1.0; computerized version", 1986.

[74] S. Hidi, "Interest, reading, and learning: Theoretical and practical considerations", *Educational Psychology Review*, vol. 13, no. 3, pp. 191–209, 2001.

[75] S. Hido, H. Kashima, and Y. Takahashi, "Roughly balanced bagging for imbalanced data", *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 2, no. 5-6, pp. 412–426, 2009.

[76] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[77] J.-C. Hong, M.-Y. Hwang, K.-H. Tai, and C.-R. Tsai, "An exploration of students' science learning interest related to their cognitive anxiety, cognitive load, self-confidence and learning progress using inquiry-based learning with an ipad", *Research in Science Education*, pp. 1–20, 2017.

[78] J. Hou, J. Rashid, and K. M. Lee, "Cognitive map or medium materiality? reading on paper and screen", *Computers in Human Behavior*, vol. 67, pp. 84–94, 2017.

[79] A. Hyrskykari, P. Majaranta, A. Aaltonen, and K.-J. Räihä, "Design issues of idict: A gaze-assisted translation aid", in *Proceedings of the 2000 symposium on Eye tracking research & applications*, ACM, 2000, pp. 9–14.

[80] Y. Ishiguro, A. Mujibiya, T. Miyaki, and J. Rekimoto, "Aided eyes: Eye activity sensing for daily life", in *Proceedings of 1st Augmented Human International Conference*, 2010, p. 25.

[81] S. Ishimaru, S. S. Bukhari, C. Heisel, N. Großmann, P. Klein, J. Kuhn, and A. Dengel, "Augmented learning on anticipating textbooks with eye tracking", in *Positive Learning in the Age of Information*, Springer, 2018, pp. 387–398.

[82] S. Ishimaru, S. S. Bukhari, C. Heisel, J. Kuhn, and A. Dengel, "Towards an intelligent textbook: Eye gaze based attention extraction on materials for learning and instruction in physics", in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ACM, 2016, pp. 1041–1045.

[83] S. Ishimaru, T. Dingler, K. Kunze, K. Kise, and A. Dengel, “Reading interventions: Tracking reading state and designing interventions”, in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, ACM, 2016, pp. 1759–1764.

[84] S. Ishimaru, N. Großmann, A. Dengel, K. Watanabe, Y. Arakawa, C. Heisel, P. Klein, and J. Kuhn, “Hypermind builder: Pervasive user interface to create intelligent interactive documents”, in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, ACM, 2018, pp. 357–360.

[85] S. Ishimaru, K. Hoshika, K. Kunze, K. Kise, and A. Dengel, “Towards reading trackers in the wild: Detecting reading activities by eog glasses and deep neural networks”, in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ACM, 2017, pp. 704–711.

[86] S. Ishimaru, S. Jacob, A. Roy, S. S. Bukhari, C. Heisel, N. Großmann, M. Thees, J. Kuhn, and A. Dengel, “Cognitive state measurement on learning materials by utilizing eye tracker and thermal camera”, in *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition*, IEEE, vol. 8, 2017, pp. 32–36.

[87] S. Ishimaru, K. Kunze, K. Kise, and A. Dengel, “The wordometer 2.0: Estimating the number of words you read in real life using commercial eog glasses”, in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ACM, 2016, pp. 293–296.

[88] S. Ishimaru, K. Kunze, K. Kise, and M. Inami, “Position paper: Brain teasers - toward wearable computing that engages our mind”, in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, Seattle, Washington: ACM, 2014, pp. 1405–1408.

[89] S. Ishimaru, K. Kunze, K. Tanaka, Y. Uema, K. Kise, and M. Inami, “Smart eyewear for interaction and activity recognition”, in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, Seoul, Republic of Korea: ACM, 2015, pp. 307–310.

[90] S. Ishimaru, K. Kunze, Y. Uema, K. Kise, M. Inami, and K. Tanaka, “Smarter eyewear: Using commercial eog glasses for activity recognition”, in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ser. UbiComp ’14 Adjunct, ACM, 2014, pp. 239–242.

[91] S. Ishimaru, K. Kunze, Y. Utsumi, M. Iwamura, and K. Kise, “Where are you looking at? - feature-based eye tracking on unmodified tablets”, in *Proceedings of the 2013 2Nd IAPR Asian Conference on Pattern Recognition*, IEEE Computer Society, 2013, pp. 738–739.

[92] S. Ishimaru, T. Maruichi, K. Kise, and A. Dengel, “Gaze-based self-confidence estimation on multiple-choice questions and its feedback”, in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’20 Asian CHI Symposium 2020, ACM, 2020.

[93] S. Ishimaru, J. Weppner, K. Kunze, A. Bulling, K. Kise, A. Dengel, and P. Lukowicz, “In the blink of an eye: Combining head motion and eye blink frequency for activity recognition with google glass”, in *Proceedings of the 5th Augmented Human International Conference*, Kobe, Japan: ACM, 2014, pp. 150–153.

[94] S. Ishimaru, J. Weppner, A. Poxrucker, P. Lukowicz, K. Kunze, and K. Kise, “Shiny: An activity logging platform for google glass”, in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, Seattle, Washington: ACM, 2014, pp. 283–286.

[95] S. Jacob, S. S. Bukhari, S. Ishimaru, and A. Dengel, “Gaze-based interest detection on newspaper articles”, in *Proceedings of the 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction*, ACM, 2018, p. 4.

[96] S. Jacob, S. Ishimaru, and A. Dengel, “Interest detection while reading newspaper articles by utilizing a physiological sensing wristband”, in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, ACM, 2018, pp. 78–81.

[97] S. C. Jacobs, R. Friedman, J. D. Parker, G. H. Tofler, A. H. Jimenez, J. E. Muller, H. Benson, and P. H. Stone, “Use of skin conductance changes during mental stress testing as an index of autonomic arousal in cardiovascular research”, *American heart journal*, vol. 128, no. 6, pp. 1170–1177, 1994.

[98] R. Jersakova, R. J. Allen, J. Booth, C. Souchay, and A. R. O’Connor, “Understanding metacognitive confidence: Insights from judgment-of-learning justifications”, *Journal of Memory and Language*, vol. 97, pp. 187–207, 2017.

[99] M. A. Just and P. A. Carpenter, “A theory of reading: From eye fixations to comprehension.”, *Psychological review*, vol. 87, no. 4, p. 329, 1980.

[100] D. Kahneman and J. Beatty, “Pupil diameter and load on memory”, *Science*, vol. 154, no. 3756, pp. 1583–1585, 1966.

[101] K. Kishi and M. Miura, "Detecting learners' weak points utilizing time intervals of pen strokes", *International Journal of Learning Technologies and Learning Environments*, vol. 1, no. 1, pp. 61–77, 2018.

[102] P. Klein, A. Dengel, and J. Kuhn, "Students' visual attention while solving multiple representation problems in upper-division physics", in *Positive Learning in the Age of Information*, Springer, 2018, pp. 67–87.

[103] S. Kleitman and J. Gibson, "Metacognitive beliefs, self-confidence and primary learning environment of sixth grade students", *Learning and Individual Differences*, vol. 21, no. 6, pp. 728–735, 2011.

[104] R. Kliegl, A. Nuthmann, and R. Engbert, "Tracking the mind during reading: The influence of past, present, and future words on fixation durations.", *Journal of Experimental Psychology: General; Journal of Experimental Psychology: General*, vol. 135, no. 1, p. 12, 2006.

[105] K. Kojima, K. Muramatsu, and T. Matsui, "Experimental study toward estimation of a learner mental state from processes of solving multiple choice problems based on eye movements", in *Proceedings of 20th International Conference on Computers in Education*, 2012, pp. 81–85.

[106] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2176–2184.

[107] M. T. Kucewicz, J. Dolezal, V. Kremen, B. M. Berry, L. R. Miller, A. L. Magee, V. Fabian, and G. A. Worrell, "Pupil size reflects successful encoding and recall of memory in humans", *Scientific reports*, vol. 8, no. 1, p. 4949, 2018.

[108] K. Kunze, S. Ishimaru, Y. Utsumi, and K. Kise, "My reading life: Towards utilizing eyetracking on unmodified tablets and phones", in *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, Zurich, Switzerland: ACM, 2013, pp. 283–286.

[109] K. Kunze, M. Katsutoshi, Y. Uema, and M. Inami, "How much do you read?: Counting the number of words a user reads using electrooculography", in *Proceedings of the 6th Augmented Human International Conference*, ACM, 2015, pp. 125–128.

[110] K. Kunze, H. Kawaichi, K. Yoshimura, and K. Kise, "The wordometer–estimating the number of words read using document image retrieval and mobile eye tracking", in *12th International Conference on Document Analysis and Recognition*, 2013, pp. 25–29.

[111] K. Kunze, K. Masai, M. Inami, Ö. Sacakli, M. Liwicki, A. Dengel, S. Ishimaru, and K. Kise, "Quantifying reading habits: Counting how many words you read", in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, 2015, pp. 87–96.

[112] K. Kunze, S. Sanchez, T. Dingler, O. Augereau, K. Kise, M. Inami, and T. Tsutomu, "The augmented narrative: Toward estimating reader engagement", in *Proceedings of the 6th Augmented Human International Conference*, ACM, 2015, pp. 163–164.

[113] A. Kushki, J. Fairley, S. Merja, G. King, and T. Chau, "Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites", *Physiological measurement*, vol. 32, no. 10, p. 1529, 2011.

[114] C. Lander, M. Speicher, D. Paradowski, N. Coenen, S. Biewer, and A. Krüger, "Collaborative newspaper: Exploring an adaptive scrolling algorithm in a multi-user reading scenario", in *Proceedings of the 4th International Symposium on Pervasive Displays*, ACM, 2015, pp. 163–169.

[115] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors.", *IEEE Communications Surveys and Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[116] R. S. Lazarus and E. M. Opton Jr, "The study of psychological stress: A summary of theoretical formulations and experimental findings", *Anxiety and behavior*, vol. 1, 1966.

[117] H. Lee, Y. Kanakogi, and K. Hiraki, "Building a responsive teacher: How temporal contingency of gaze interaction influences word learning with virtual tutors", *Royal Society open science*, vol. 2, no. 1, p. 140 361, 2015.

[118] D. Leiner, A. Fahr, and H. Früh, "Eda positive change: A simple algorithm for electrodermal activity to measure general audience arousal during media exposure", *Communication Methods and Measures*, vol. 6, no. 4, pp. 237–250, 2012.

[119] E. A. Linnenbrink and P. R. Pintrich, "The role of self-efficacy beliefs instudent engagement and learning intheclassroom", *Reading &Writing Quarterly*, vol. 19, no. 2, pp. 119–137, 2003.

[120] D. G. Lowe, "Object recognition from local scale-invariant features", in *Proceedings of the seventh IEEE international conference on Computer vision*, Ieee, vol. 2, 1999, pp. 1150–1157.

[121] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition", in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1894–1903.

[122] S. Martinez-Conde, S. Macknik, X. Troncoso, and T. Dyar, "Microsaccades counteract visual fading during fixation", *Journal of Vision*, vol. 5, no. 12, pp. 72–72, 2005.

[123] P. Martínez-Gómez and A. Aizawa, "Recognition of understanding level and language skill using measurements of reading behavior", in *Proceedings of the 19th international conference on Intelligent User Interfaces*, ACM, 2014, pp. 95–104.

[124] T. Maruichi, S. Ishimaru, and K. Kise, "Self-confidence estimation on vocabulary tests with stroke-level handwriting logs", in *Proceedings of the 15th IAPR International Conference on Document Analysis and Recognition*, ser. ICDAR HDI'19, 2019, pp. 18–22.

[125] T. Maruichi, K. Kise, O. Augereau, and M. Iwata, "Keystrokes tell you how confident you are: An application to vocabulary acquisition", in *Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2018 ACM International Symposium on Wearable Computers*, ACM, 2018.

[126] M. Matsubara, O. Augereau, C. L. Sanches, and K. Kise, "Emotional arousal estimation while reading comics based on physiological signal analysis", in *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, ACM, 2016, p. 7.

[127] D. E. Meltzer, "Relation between students' problem-solving performance and representational format", *American Journal of Physics*, vol. 73, no. 5, pp. 463–478, 2005.

[128] S. Michie, C. Abraham, C. Whittington, J. McAteer, and S. Gupta, "Effective techniques in healthy eating and physical activity interventions: A meta-regression.", *Health Psychology*, vol. 28, no. 6, pp. 690–701, 2009.

[129] P. Mistry, P. Maes, and L. Chang, "Wuw-wear ur world: A wearable gestural interface", in *CHI'09 extended abstracts on Human factors in computing systems*, ACM, 2009, pp. 4111–4116.

[130] S. Mozaffari, S. Bukhari, A. Dengel, P. Klein, and J. Kuhn, "A study on representational competence in physics using mobile remote eye tracking systems, proceedings", in *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, accepted, 2016.

[131] O. J. Muensterer, M. Lacher, C. Zoeller, M. Bronstein, and J. Kübler, "Google glass in pediatric surgery: An exploratory study", *International Journal of Surgery*, vol. 12, no. 4, pp. 281–289, 2014.

[132] M. H. H. Nashif, M. B. A. Miah, A. Habib, A. C. Moulik, M. S. Islam, M. Zakareya, A. Ullah, M. A. Rahman, and M. Al Hasan, "Handwritten numeric and alphabetic character recognition and signature verification using neural network", *Journal of Information Security*, vol. 9, no. 03, p. 209, 2018.

[133] M. S. Nomikos, E. Opton Jr, and J. R. Averill, "Surprise versus suspense in the production of stress reaction.", *Journal of Personality and Social Psychology*, vol. 8, no. 2p1, p. 204, 1968.

[134] Y. Ohbayashi, S. Ishimaru, D. Andreas, and K. Kise, "Investigating gaze and physiological features to estimate comprehension on e-learning video lectures", in *Proceedings of the first international interdisciplinary Symposium on Reading Experience and Analysis of Documents*, 2018.

[135] A. Okoso, T. Toyama, K. Kunze, J. Folz, M. Liwicki, and K. Kise, "Towards extraction of subjective reading incomprehension: Analysis of eye gaze features", in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, ACM, 2015, pp. 1325–1330.

[136] L. N. Orchard and J. A. Stern, "Blinks as an index of cognitive activity during reading", *Integrative Physiological and Behavioral Science*, vol. 26, no. 2, pp. 108–116, 1991.

[137] K. E. Patterson, N. Graham, and J. R. Hodges, "Reading in dementia of the alzheimer type: A preserved ability?", *Neuropsychology*, vol. 8, no. 3, p. 395, 1994.

[138] M. A. Peters and H. Lau, "Human observers have optimal introspective access to perceptual processes even for visually masked stimuli", *Elife*, vol. 4, e09651, 2015.

[139] A. Poole and L. J. Ball, "Eye tracking in hci and usability research", in *Encyclopedia of human computer interaction*, IGI Global, 2006, pp. 211–219.

[140] J. A. Pooler, R. E. Morgan, K. Wong, M. K. Wilkin, and J. L. Blitstein, "Cooking matters for adults improves food resource management skills and self-confidence among low-income participants", *Journal of nutrition education and behavior*, vol. 49, no. 7, pp. 545–553, 2017.

[141] M. Porta, S. Ricotti, and C. J. Perez, "Emotional e-learning through eye tracking", in *Global Engineering Education Conference, 2012 IEEE*, IEEE, 2012, pp. 1–6.

[142] G. E. Raney, S. J. Campbell, and J. C. Bovee, “Using eye movements to evaluate the cognitive processes involved in text comprehension”, *Journal of visualized experiments: JoVE*, no. 83, 2014.

[143] C. A. Rashotte and J. K. Torgesen, “Repeated reading and reading fluency in learning disabled children”, *Reading Research Quarterly*, pp. 180–188, 1985.

[144] K. Rayner, “Eye movements in reading and information processing: 20 years of research.”, *Psychological bulletin*, vol. 124, no. 3, p. 372, 1998.

[145] K. Rayner, K. H. Chace, T. J. Slattery, and J. Ashby, “Eye movements as reflections of comprehension processes in reading”, *Scientific Studies of Reading*, vol. 10, no. 3, pp. 241–255, 2006.

[146] M. Resnick, J. Maloney, A. Monroy-Hernández, N. Rusk, E. Eastmond, K. Brennan, A. Millner, E. Rosenbaum, J. Silver, B. Silverman, *et al.*, “Scratch: Programming for all”, *Communications of the ACM*, vol. 52, no. 11, pp. 60–67, 2009.

[147] D. A. Robinson, “A method of measuring eye movemnent using a scieral search coil in a magnetic field”, *IEEE Transactions on bio-medical electronics*, vol. 10, no. 4, pp. 137–145, 1963.

[148] T. Roderer and C. M. Roebers, “Can you see me thinking (about my answers)? using eye-tracking to illuminate developmental differences in monitoring and control skills and their relation to performance”, *Metacognition and learning*, vol. 9, no. 1, pp. 1–23, 2014.

[149] D. S. Rudmann, G. W. McConkie, and X. S. Zheng, “Eyetracking in cognitive state detection for hci”, in *Proceedings of the 5th international conference on Multimodal interfaces*, ACM, 2003, pp. 159–163.

[150] R. M. Ryan and E. L. Deci, “Intrinsic and extrinsic motivations: Classic definitions and new directions”, *Contemporary educational psychology*, vol. 25, no. 1, pp. 54–67, 2000.

[151] ——, “Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being”, *American psychologist*, vol. 55, no. 1, p. 68, 2000.

[152] C. L. Sanches, O. Augereau, and K. Kise, “Manga content analysis using physiological signals”, in *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, ACM, 2016, p. 6.

[153] R. M. Sapolsky, *Why zebras don’t get ulcers*. WH Freeman New York, 1994.

[154] T. Sasaki, M. Saraiji, C. L. Fernando, K. Minamizawa, and M. Inami, "Metalimbs: Multiple arms interaction metamorphism", in *ACM SIGGRAPH 2017 Emerging Technologies*, ACM, 2017, p. 16.

[155] M. Schall, D. Sacha, M. Stein, M. O. Franz, and D. A. Keim, "Visualization-assisted development of deep learning models in offline handwriting recognition", 2018.

[156] F. F. Schmitt and R. Lahroodi, "The epistemic value of curiosity", *Educational Theory*, vol. 58, no. 2, pp. 125–148, 2008.

[157] W. Schnotz, "The cambridge handbook of multimedia learning: An integrated model of text and picture comprehension", 2005.

[158] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable eda device", *IEEE Transactions on information technology in biomedicine*, vol. 14, no. 2, pp. 410–417, 2010.

[159] Y. Shiga, T. Toyama, Y. Utsumi, K. Kise, and A. Dengel, "Daily activity recognition combining gaze motion and visual features", in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ACM, 2014, pp. 1103–1111.

[160] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge", *Nature*, vol. 550, no. 7676, p. 354, 2017.

[161] G. D. Spache, "Is this a breakthrough in reading?", *The Reading Teacher*, vol. 15, no. 4, pp. 258–263, 1962.

[162] S. Squires, *The effects of reading interest, reading purpose, and reading maturity on reading comprehension of high school students*. Baker University, 2014.

[163] M. Stager, P. Lukowicz, and G. Troster, "Implementation and evaluation of a low-power sound-based user activity recognition system", in *Wearable Computers, 2004. ISWC 2004. Eighth International Symposium on*, IEEE, vol. 1, 2004, pp. 138–141.

[164] L. Stankov, J. Lee, W. Luo, and D. J. Hogan, "Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety?", *Learning and Individual Differences*, vol. 22, no. 6, pp. 747–758, 2012.

[165] J. Steil and A. Bulling, "Discovery of everyday human activities from long-term visual behaviour using topic models", in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, 2015, pp. 75–85.

[166] Y. Sugano, Y. Matsushita, and Y. Sato, “Appearance-based gaze estimation using visual saliency”, *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 2, pp. 329–341, 2013.

[167] ——, “Learning-by-synthesis for appearance-based 3d gaze estimation”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1821–1828.

[168] K. Tagawa, K. Watanabe, and H. Ogata, *Takram Design Engineering; Pendulum of Design Innovation*. LIXIL Publishing, 2014.

[169] T. Takahashi, K. Shiro, A. Matsuda, R. Komiyama, H. Nishioka, K. Hori, Y. Ishiguro, T. Miyaki, and J. Rekimoto, “Augmented jump: A backpack multirotor system for jumping ability augmentation”, in *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, ACM, 2018, pp. 230–231.

[170] J. R. Themanson and P. J. Rosen, “Examining the relationships between self-efficacy, task-relevant attentional control, and task performance: Evidence from event-related brain potentials”, *British Journal of Psychology*, vol. 106, no. 2, pp. 253–271, 2015.

[171] B Thomas, K. Grimmer, J Zucco, and S. Milanese, “Where does the mouse go? an investigation into the placement of a body-attached touchpad mouse for wearable computers”, *Personal and Ubiquitous computing*, vol. 6, no. 2, pp. 97–112, 2002.

[172] T. Toyama, A. Dengel, W. Suzuki, and K. Kise, “Wearable reading assist system: Augmented reality document combining document retrieval and eye tracking”, in *Document Analysis and Recognition, 2013 12th International Conference on*, IEEE, 2013, pp. 30–34.

[173] T. Toyama, T. Kieninger, F. Shafait, and A. Dengel, “Museum guide 2.0-an eye-tracking based personal assistant for museums and exhibits”, in *Proc. of Int. Conf. on Re-Thinking Technology in Museums*, vol. 1, 2011.

[174] ——, “Gaze guided object recognition using a head-mounted eye tracker”, in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ACM, 2012, pp. 91–98.

[175] M.-J. Tsai, H.-T. Hou, M.-L. Lai, W.-Y. Liu, and F.-Y. Yang, “Visual attention for solving multiple-choice science problem: An eye-tracking analysis”, *Computers & Education*, vol. 58, no. 1, pp. 375–385, 2012.

[176] M. J. Tyre and E. Von Hippel, “The situated nature of adaptive learning in organizations”, *Organization science*, vol. 8, no. 1, pp. 71–83, 1997.

[177] M. Uetsuki, J. Watanabe, H. Ando, and K. Maruya, "Reading traits for dynamically presented texts: Comparison of the optimum reading rates of dynamic text presentation and the reading rates of static text presentation", *Frontiers in Psychology*, vol. 8, p. 1390, 2017.

[178] B. Ugurlu, R. Kandemir, A. Carus, and E. Abay, "An expert system for determining the emotional change on a critical event using handwriting features", *TEM Journal*, vol. 5, no. 4, p. 480, 2016.

[179] A. Van Heuvelen, "Learning to think like a physicist: A review of research-based instructional strategies", *American Journal of Physics*, vol. 59, no. 10, pp. 891–897, 1991.

[180] S. Von Stumm, B. Hell, and T. Chamorro-Premuzic, "The hungry mind intellectual curiosity is the third pillar of academic performance", *Perspectives on Psychological Science*, vol. 6, no. 6, pp. 574–588, 2011.

[181] N. J. Wade and B. W. Tatler, "Did javal measure eye movements during reading?", *Journal of Eye Movement Research*, vol. 2, no. 5, 2009.

[182] M. Weiser, "Some computer science issues in ubiquitous computing", *Communications of the ACM*, vol. 36, no. 7, pp. 75–84, 1993.

[183] M. Weiser and J. S. Brown, "The coming age of calm technology", in *Beyond calculation*, Springer, 1997, pp. 75–85.

[184] G. M. Weiss and J. W. Lockhart, "The impact of personalization on smartphone-based activity recognition", in *AAAI Workshop on Activity Context Representation: Techniques and Languages*, 2012.

[185] J. Weppner, M. Hirth, J. Kuhn, and P. Lukowicz, "Physics education with google glass gphysics experiment app", in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ACM, 2014, pp. 279–282.

[186] E. Whitmire, L. Trutoiu, R. Cavin, D. Perek, B. Scally, J. Phillips, and S. Patel, "Eyecontact: Scleral coil eye tracking for virtual reality", in *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, ACM, 2016, pp. 184–191.

[187] M. Wolf and C. J. Stoodley, *Proust and the squid: The story and science of the reading brain*. Harper Perennial New York, 2008.

[188] K. Yamada, K. Kise, and O. Augereau, "Estimation of confidence based on eye gaze: An application to multiple-choice questions", in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing International Symposium on Wearable Computers*, ACM, 2017, pp. 217–220.

[189] N. Yannier, A. Israr, J. F. Lehman, and R. L. Klatzky, “Feelsleeve: Haptic feedback to enhance early reading”, in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM, 2015, pp. 1015–1024.

[190] H. Yasufuku, T. Terada, and M. Tsukamoto, “A lifelog system for detecting psychological stress with glass-equipped temperature sensors”, in *Proceedings of the 7th Augmented Human International Conference 2016*, ACM, 2016, p. 8.

[191] K. Yoshimura, K. Kunze, and K. Kise, “The eye as the window of the language ability: Estimation of english skills by analyzing eye movement while reading documents”, in *Proceedings of the 13th International Conference on Document Analysis and Recognition*, 2015, pp. 251–255.

[192] J. Zagermann, U. Pfeil, and H. Reiterer, “Measuring cognitive load using eye tracking technology in visual computing”, in *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, 2016, pp. 78–85.

[193] Z. Zhan, L. Zhang, H. Mei, and P. S. Fong, “Online learners’ reading ability detection based on eye-tracking sensors”, *Sensors*, vol. 16, no. 9, p. 1457, 2016.

[194] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, “Appearance-based gaze estimation in the wild”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4511–4520.

[195] ——, “It’s written all over your face: Full-face appearance-based gaze estimation”, in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 2017, pp. 2299–2308.

Own Publications

Journal Paper

1. Oliver Amft, Florian Wahl, Shoya Ishimaru and Kai Kunze. Making Regular Eyeglasses Smart. Pervasive Computing, IEEE, vol. 14, no. 3, pp. 32-43, 2015.

Book Chapter

1. Shoya Ishimaru, Syed Saqib Bukhari, Carina Heisel, Nicolas Großmann, Pascal Klein, Jochen Kuhn and Andreas Dengel. Augmented Learning on Anticipating Textbooks with Eye Tracking. Positive Learning and Transformation in the Information Age (PLATO) - A blessing or a curse?, Springer, pp. 387-398, December 2017.

International Conference Papers

1. Shoya Ishimaru, Ko Watanabe, Nicolas Großmann, Carina Heisel, Pascal Klein, Yutaka Arakawa, Jochen Kuhn and Andreas Dengel. HyperMind Builder - Pervasive User Interface to Create Intelligent Interactive Documents. Proc. UbiComp 2018 Adjunct, pp. 357-360, October 2018.

2. Soumy Jacob, Shoya Ishimaru and Andreas Dengel. Interest Detection while Reading Newspaper Articles by Utilizing a Physiological Sensing Wristband. Proc. UbiComp 2018 Adjunct, pp. 78-81, October 2018. **Honorable Mention**

3. Dayananda Herurkar, Shoya Ishimaru and Andreas Dengel. Combining Software-Based Eye Tracking and a Wide-Angle Lens for Sneaking Detection. Proc. UbiComp 2018 Adjunct, pp. 54-57, October 2018.

4. Jayasankar Santhosh, Shoya Ishimaru and Andreas Dengel. Estimating Fixation Durations for Each Word in Documents towards Readability Measurement. Proc. READ2018, October 2018.

5. Yuya Ohbayashi, Shoya Ishimaru, Andreas Dengel and Koichi Kise Investigating Gaze and Physiological Features to Estimate Comprehension on E-learning Video Lectures. Proc. READ2018, October 2018.

6. Nicolas Großmann, Iuliia Brishtel, Shoya Ishimaru, Andreas Dengel, Carina Heisel, Pascal Klein and Jochen Kuhn. iQL - Immersive Quantified Learning Lab. Proc. READ2018, October 2018.

7. Soumy Jacob, Shoya Ishimaru, Syed Saqib Bukhari and Andreas Dengel. Gaze-based Interest Detection on Newspaper Articles. Proc. ETRA 2018 (PETMEI 2018), June 2018.

8. Iuliia Brishtel, Shoya Ishimaru, Olivier Augereau, Koichi Kis, and Andreas Dengel. Assessing Cognitive Workload on Printed and Electronic Media using Eye-Tracker and EDA Wristband. Proc. IUI 2018, March 2018.

9. Shoya Ishimaru, Soumy Jacob, Apurba Roy, Syed Saqib Bukhari, Carina Heisel, Nicolas Großmann, Michael Thees, Jochen Kuhn and Andreas Dengel. Cognitive State Measurement on Learning Materials by Utilizing Eye Tracker and Thermal Camera. Proc. ICDAR 2017 (HDI 2017), pp. 32-36, November 2017.

10. Shoya Ishimaru, Kensuke Hoshika, Kai Kunze, Koichi Kise and Andreas Dengel. Towards Reading Trackers in the Wild: Detecting Reading Activities by EOG Glasses and Deep Neural Networks. Proc. UbiComp 2017 Adjunct (WAHM 2017), pp. 704-711, September 2017.

11. Shoya Ishimaru and Andreas Dengel. ARFLED: Ability Recognition Framework for Learning and Education. Proc. UbiComp 2017 Adjunct, pp. 339-343, September 2017.

12. Sabine Hoffmann, Helga Tauscher, Andreas Dengel, Shoya Ishimaru, Sheraz Ahmed, Jochen Kuhn, Carina Heisel and Yutaka Arakawa. Sensing thermal stress at office workplaces. Proc. ICHES 2016, October 2016.

13. Shoya Ishimaru, Cognitive State Recognition for Developing Anticipating Textbook. Proc. ICMU 2016, No. 20, October 2016.

14. Shoya Ishimaru, Syed Saqib Bukhari, Carina Heisel, Jochen Kuhn and Andreas Dengel. Towards an Intelligent Textbook: Eye Gaze Based Attention Extraction on Materials for Learning and Instruction in Physics. Proc. UbiComp 2016 Adjunct (WAHM 2016), pp. 1041-1045, September 2016.

15. Shoya Ishimaru, Kai Kunze, Koichi Kise and Andreas Dengel. The Wordometer 2.0 - Estimating the Number of Words You Read in Real Life using Commercial EOG Glasses. Proc. UbiComp 2016 Adjunct, pp. 1217-1220, September 2016.

16. Shoya Ishimaru, Tilman Dingler, Kai Kunze, Koichi Kise and Andreas Dengel. Reading Interventions - Tracking Reading State and Designing Interventions. Proc. UbiComp 2016 Adjunct (EyeWear 2016), pp. 1759-1764, September 2016.

17. Shoya Ishimaru and Koichi Kise. Quantifying the Mental State on the Basis of Physical and Social Activities. Proc. UbiComp 2015 Adjunct (WAHM 2015), pp. 1217-1220, September 2015.

18. Kai Kunze, Yuji Uema, Katsuma Tanaka, Shoya Ishimaru, Koichi Kise and Masahiko Inami MEME – Eye Wear Computing to Explore Human Behavior. Proc. UbiComp 2015 Adjunct, pp. 361-363, September 2015.

19. Shoya Ishimaru, Kai Kunze, Katsuma Tanaka, Yuji Uema, Koichi Kise and Masahiko Inami. Smart Eyewear for Interaction and Activity Recognition. Proc. CHI 2015 Extended Abstracts, pp. 307-310, April 2015.

20. Shoya Ishimaru, Kai Kunze, Katsuma Tanaka, Yuji Uema, Koichi Kise and Masahiko Inami. Smarter Eyewear – Using Commercial EOG Glasses for Activity Recognition. Proc. UbiComp 2014 Adjunct, pp. 239-242, September 2014.

21. Shoya Ishimaru, Jens Weppner, Andreas Poxrucker, Kai Kunze, Paul Lukowicz and Koichi Kise. Shiny - An Activity Logging Platform for Google Glass. Proc. UbiComp 2014 Adjunct, pp. 283-286, September 2014.

22. Shoya Ishimaru, Jens Weppner, Kai Kunze, Andreas Bulling, Koichi Kise, Andreas Dengel and Paul Lukowicz. In the Blink of an Eye - Combining Head Motion and Eye Blink Frequency for Activity Recognition with Google Glass. Proc. AH 2014, pp. 150-153, March 2014.

Papers Not Related to the Thesis or with My Small Contribution

1. Koichi Kise, Olivier Augereau, Yuzuko Utsumi, Masakazu Iwamura, Kai Kunze, Shoya Ishimaru and Andreas Dengel. Quantified Reading and Learning for Sharing Experiences. Proc. UbiComp 2017 Adjunct (WAHM 2017), pp. 724-731, September 2017.

2. Christophe Rigaud, Thanh-Nam Le, Shoya Ishimaru, J.-C. Burie, J.-M. Ogier, Motoi Iwata and Koichi Kise. Semi-automatic text and graphics extraction of manga using eye tracking information. Proc. DAS 2016, pp. 120-125, April 2016.

3. Kai Kunze, Ömer Sacakli, Shoya Ishimaru, Andreas Dengel, Marcus Liwicki, Koichi Kise and Masahiko Inami. Quantifying Reading Habits – Counting How Many Words You Read. Proc. UbiComp 2015, pp. 87-96, September 2015.

4. Kai Kunze, Kazutaka Inoue, Katsutoshi Masai, Yuji Uema, Sean Shao-An Tsai, Shoya Ishimaru, Katsuma Tanaka, Koichi Kise and Masahiko Inami. MEME - Smart Glasses to Promote Healthy Habits for Knowledge Workers. Proc. SIGGRAPH 2015 Emerging Technologies, p. 17, August 2015.

5. Kai Kunze, Katsuma Tanaka, Shoya Ishimaru, Koichi Kise and Masahiko Inami. Nekoze! – Monitoring and Detecting Head Posture while Working with Laptop and Mobile Phone. Proc. PervasiveHealth 2015, pp. 237-240, May 2015.

6. Shoya Ishimaru, Kai Kunze, Koichi Kise and Masahiko Inami. Position Paper: Brain Teasers - Toward Wearable Computing that Engages Our Mind. Proc. UBiComp 2014 Adjunct (WAHM 2014), pp. 1405-1408, September 2014.

7. Shoya Ishimaru, Kai Kunze, Yuzuko Utsumi, Masakazu Iwamura and Koichi Kise. Where Are You Looking At? - Feature-Based Eye Tracking on Unmodified Tablets. Proc. ACPR 2013, pp. 738-739, November 2013.

8. Kai Kunze, Shoya Ishimaru, Yuzuko Utsumi and Koichi Kise. My reading life: towards utilizing eyetracking on unmodified tablets and phones. Proc. UbiComp 2013 Adjunct, pp. 283-286, September 2013.

9. Kai Kunze, Yuki Shiga, Shoya Ishimaru and Koichi Kise. Reading Activity Recognition Using an Off-the-Shelf EEG–Detecting Reading Activities and Distinguishing Genres of Documents. Proc. ICDAR 2013, pp. 96-100, August 2013.

Curriculum Vitae

Summary

Shoya Ishimaru is a Senior Researcher at the German Research Center for Artificial Intelligence (DFKI). His research interests are to invent new technologies augmenting abilities of human begins and investigating influences of the augmentations to human minds. After receiving B.E. and M.E. degree in Electrical Engineering and Information Science from Osaka Prefecture University in 2014 and 2016, he defended his Ph.D. in Engineering at the University of Kaiserslautern in 2019. Shoya is also a software developer. He likes not only programming but designing beautiful software. As an internship/part-time engineer, he worked for many IT companies including paperboy&co., Hatena, and Recruit. In 2016, he received a title of MITOU Super Creator by Ministry of Economy, Trade, and Industry in Japan by developing a system for visualizing the mental state. He gave a number of invited talks at universities, high schools, and TEDx. (more info: https://shoya.io)

Background

Human-Computer Interaction, Learning Analytics, Wearable Computing, and Pattern Recognition

Work Experience

Senior Researcher, German Research Center for Artificial Intelligence (DFKI)	*2019 – current*
Researcher, German Research Center for Artificial Intelligence (DFKI)	*2016 – 2019*

Education

Doctor of Engineering, University of Kaiserslautern	*2016 – 2019*
Master of Engineering, Graduate School of Engineering, Osaka Prefecture University	*2014 – 2016*
Bachelor of Engineering, Osaka Prefecture University	*2010 – 2014*

Funds

Co-Investigator, JSPS Grant-in-Aid for Scientific Research (C), 4.68 M JPY	*2017-2019*
Representative, JSPS Grant-in-Aid for Young Scientists (B), 4.16 M JPY	*2017-2019*
Creator, IPA MITOU Exploratory IT Human Resources Project, 2.30 M JPY	*2015*

Selected Part-Time Job and Service

Visiting Researcher, Osaka Prefecture University	*2016 – current*
Researcher, Keio Media Design Research Institute	*2014 – current*
Research Associate, University of Kaiserslautern	*2016 – 2019*

Selected Publications

Soumy Jacob, Shoya Ishimaru and Andreas Dengel. "Interest Detection While Reading Newspaper Articles by Utilizing a Physiological Sensing Wristband". In Proc. UbiComp '18 Adjunct, pp. 78–81, 2018. *Poster Track Honorable Mention*

Shoya Ishimaru, Syed Saqib Bukhari, Carina Heisel, Nicolas Großmann, Pascal Klein, Jochen Kuhn and Andreas Dengel. "Augmented Learning on Anticipating Textbooks with Eye Tracking". In Positive Learning in the Age of Information (PLATO), pp. 387–398, 2018. *Book Chapter*

Shoya Ishimaru, Jens Weppner, Kai Kunze, Andreas Bulling, Koichi Kise, Andreas Dengel and Paul Lukowicz. "In the Blink of an Eye: Combining Head Motion and Eye Blink Frequency for Activity Recognition with Google Glass". In Proc. AH '14, pp. 150–153, 2014. *Cited more than 100 times*