A Theisis for the Degree of Ph.D in Engineering

Facial Expression Classification Using Photo-reflective Sensors on Smart Eyewear

August 2018

Graduate School of Science and Technology

Keio University

Katsutoshi Masai

Abstract

This dissertation explores how to classify various facial expressions of a person using an eyewear device where an array of optical sensors are embedded. The sensors can measure the movement of facial muscles through the reflected intensity since the movement causes the skin deformation around eyes.

Facial expression recognition can be done automatically using computer-vision based approaches. However, there are the limitations of using a physical camera in the environment such as occlusion and trackability. These are critical in daily life scenarios.

The first contribution of the dissertation is the development of the eyewear device that is capable of classifying the posed facial expressions in a supervised manner. This approach uses photo-reflective sensors and machine learning methods. The eyewear system can recognize eight basic emotions from the posed facial expressions with the accuracy of more than 90 percent regardless of facial direction and removal/remount.

The second contribution is the classification of spontaneous facial expressions. The initial field study was undertaken to see how the user changes the facial expressions in daily life settings. Then, the mapping of the spontaneous facial expressions in daily conversations was done. The method visually summarized five user's facial expressions in an unsupervised manner. It revealed how similar the expressions of each user are. Then, the analysis of reading jokes with the device has been investigated. The sensor data could show the user's blinks, the line breaks, and the facial response of the user.

Finally, further usage of the eyewear device has been demonstrated. The input techniques by eye gestures (7 kinds, nine users, average 92.9%) in addition to facial expressions and by rubbing the face (10 areas, five users, average 88.7%) are developed.

It can expand the possibility of the interaction in the era of wearable computing.

In summary, this work aims to make the grounds of eyewear computing for classifying the information from the faces in daily life. To this end, the device is developed, and the usability is validated by the classification of prototypic expressions, spontaneous expressions of the users and the detection of eye movements and the intentional gestures.

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Associate Professor Maki Sugimoto for the support of my Ph.D. study and research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I am thankful for him to respect my ideas and let me execute it.

Besides my advisor, I would like to thank the rest of my thesis committee: Professor Hideo Saito, Professor Bruce Thomas, Assistant Professor Yuta Sugiura, and Associate Professor Kai Kunze. Without the committee, it is not possible to complete my thesis.

My sincere thanks go to Associate Professor Kai Kunze for giving me the excellent opportunity to meet great researchers overseas. I appreciate his continuous advice for my research. I would like to thank Assistant Professor. Yuta Itoh and Assistant Professor. Yuta Sugiura. Their advice is fruitful. I thank labmates doing the collaborative research in Sugimoto Lab. I appreciate all the members who took part in the study without hesitance. Last but not the least, I would like to thank my family: my parents for giving birth to me in the first place and supporting me spiritually throughout my life. I would like to thank my brother to give me the advice about my research life. I could not have done the thesis without their help.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

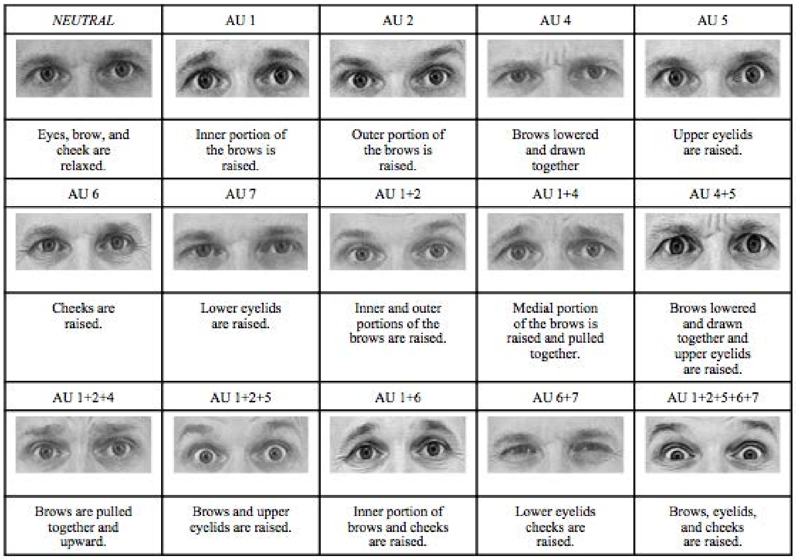
1.1 Nonverbal Communication, Faces, Facial Expressions

We communicate every day. We express ourselves, understand others, and help each other to lead a life. Even sometimes we find ourselves among others who speak a language that we do not understand, and we can still communicate. We use not only languages but also boy gestures, audio tones, head movements and facial expressions to understand each other. We, as humans, are made given the existence of other human beings.

The nonverbal communication is one of the essential elements in daily life [47,65]. When we are born, we do not speak languages. We learn the nonverbal channels faster than langauges to understand the intention and emotion. Among the nonverbal channels, a face is regarded as one of the most important channels. All of us can recognize other people's faces with little effort as we have the special mechanism to process faces [78]. Face conveys biological information such as age, sex, and nationalities. Besides, face conveys cognitive states. McDuff et al. could predict the stress condition from the face in the video with far better than random chance [64]. Facial expressions convey various complex information by activating more than 20 kinds of facial muscles. According to Fasel and Luettin, the sources of facial expressions are mental states, non-verbal communication, physiological activities and vital communication [20]. People attach various meanings to faces [72], and facial expressions are vital in communicating a person's intention, agreements and emotional states [42].

Faial expression is a highly sophisticated signal shown as a result and process of dealing with meta-information about our behavior such as thoughts, emotion, perception, and cognition. It is influenced by not only our inner state but also by our body movement and environment. It is related to our mind since Andy Clark asserts the mind as the interaction among our brain, body and the environment [8]. As such, interrupting appropriate meanings from facial expressions is a challenging problem. According to the diagram of Fussel et al., facial expression changes for various reasons [20]. Parkinson argued that the primary role of facial expressions is communicating social motives rather than emotional expressions [72]. Still, the recognizing affect from facial expressions is still active research domains. The concept of affect is firstly defined by Tomkins [96] as a biological aspect of emotions. There are two streams of the emotional theories. First one is the Cannon-Bard theory which argues that bodily response and emotion comes together [5]. Another one is James-Lange Theory [37]. The theory describes the emotion in secondary as physiological response comes first. Facial feedback hypothesis follows the theory [96]. It proposes that facial movement activates a related emotional experience. Overall, the facial expressions have various roles and meanings in our life.

The technology of facial expression recognition automatically leads to practical applications in various fields. First, recognizing facial expressions would be an essential step towards improving the user experience for Human-Computer Interaction (HCI). As nonverbal clues play an essential role in our everyday interpersonal interactions, it seems natural to incorporate them in the field of HCI. Computing systems become increasingly ubiquitous and support us in everyday situations, and they need to be able to process more contextual information such as user's intention and emotion to improve the quality of HCI such as social robots. This application is also related to the area of Affective Computing [75]. Second, the technology can be used for marketing such as analyzing the contents based on facial expressions and for online tutoring the learners by noticing when the user gets confused. Third, logging facial expression for long-term makes possible to manage user's daily satisfaction and mental health. Keeping the users in good health is useful. From the doctor's perspective, the data can be useful for understanding patients, such as pain detection and the treatment of patients' depression. This application has the potential to impact the society, yet this needs to keep track of facial expression activity for long, or in daily life.



NEUTRAL	AU 1	AU 2	AU 4	AU 5
Eyes, brow, and cheek are relaxed.	Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together	Upper eyelids are raised.
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5
Cheeks are raised.	Lower eyelids are raised.	Inner and outer portions of the brows are raised.	Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.
AU 1+2+4	AU 1+2+5	AU 1+6	AU 6+7	AU 1+2+5+6+7
Brows are pulled together and upward.	Brows and upper eyelids are raised.	Inner portion of brows and cheeks are raised.	Lower eyelids cheeks are raised.	Brows, eyelids, and cheeks are raised.

Figure 1-1: The examples of action units of the upper face in the Facial Action Coding System. The image is retrieved from [95]

1.2 Method to Recognize Facial Expression Automatically

The study of measuring facial expressions started with the work of Darwin [12]. He showed the universality of emotional facial expressions. To objectively describe the facial behavior, Ekman and Friesen developed the Facial Action Coding System (FACS) [18]. FACS uses 44 action units(see Figure 1-1) for the description. It does not directly convey mental activities, but the scheme has been widely used for the measurement because of its objectivity. However, the manual coding of facial behavior is costly. To tackle the issue, automatic facial expression recognition has been advanced mostly in the area of computer vision [95]. According to Fasel et al. [20], the general approach to automatic facial expression analysis (AFEA) in early stages consists of three steps: Face acquisition, facial data extraction and representation, and facial expression classification. The classification was based on basic emotion categories, FACS, and the dimensional description of affect [80]. Even more recently, the standard pipeline is similar: pre-processing, feature extraction, and machine learning [57].

The computer vision method has still difficulties in real life setting, and there is still relatively little research that tries to detect facial expressions in a daily life scene. The challenges of real-life detection using computer vision include real-time and robust detection, the non-frontal face pose handling, occlusion handling, illusion changes, context awareness, subtle expressions and individual differences [9]. Most of the researches focus on basic emotion classification, and recently classification of more diverse expressions such as pain detection and subtle expression is tackled.

Besides camera-based methods, many researchers used Electromyography(EMG)-based sensors to measure facial expression. EMG can measure a subtle movement of facial muscles. It has been used in the field of psychology [13].

In recent years, due to the development of wearable technology, detection of a subtle smile and long-term recording in a real environment [34, 73] are carried out. Wearables have benefits to track the user's behavior regardless of the occlusions and head poses that were problematic to computer-vision based approach.These methods both extracted the features using Independent Component Analysis(ICA) and classify smiles with either model-based or machine learning-based approach. However, the EMG signal is measured through the electrodes attached to the skin that needs to be cleaned beforehand. The process requires setup time. Besides, it may be uncomfortable for some people in daily life. Also, the measurement of each facial muscles requires one electrode. This limitation leads to usually ad-hoc application. For example, wearable EMG devices as mentioned above can only detect smile-related facial behavior.

1.3 Goal

The goal of the dissertation is to classify people's diverse facial expressions in everyday life by making an unobtrusive and truly wearable device. Concretely, the dissertation aims at classifying the emotion-related facial expression, communication-related facial expression, and cognitive aspect of facial expressions (i.e., eye movement) regardless of the daily body motion such as head-pose change, walking, and hand gestures. By

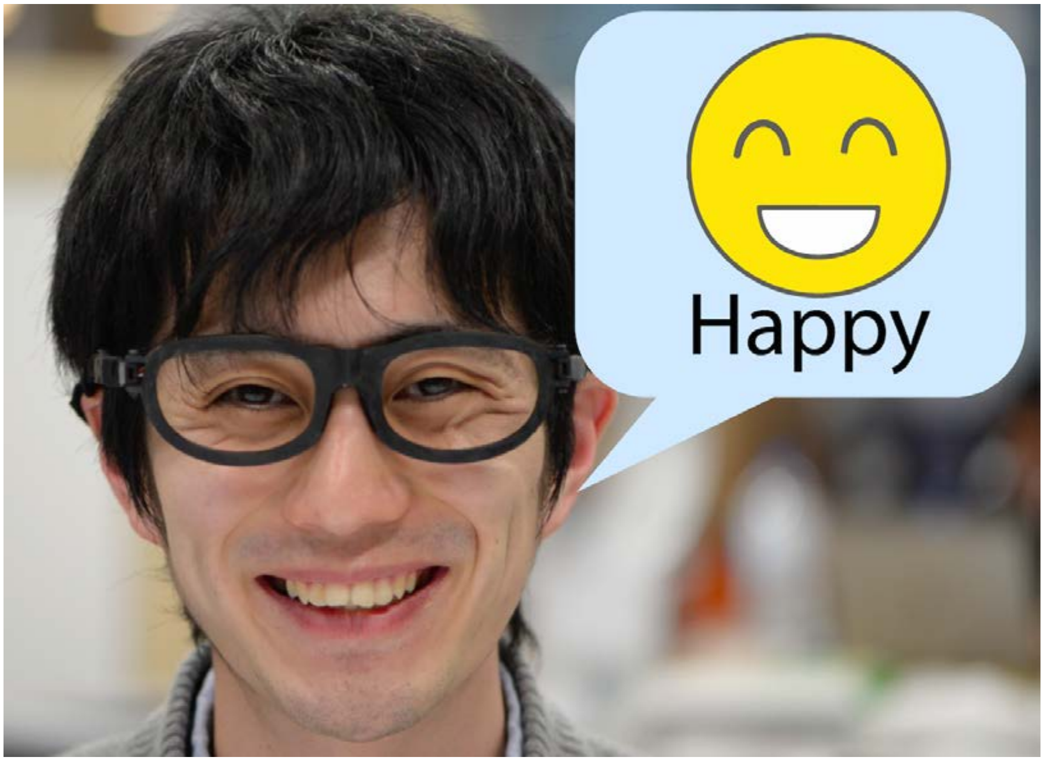


Figure 1-2: The proposed method can classify facial expressions in daily scenes.

considering the various aspects of facial expressions, it is possible to understand the user's mind deeply. It leads to the practical applications that enrich the daily life by extracting various meanings of minds. To this end, the wearable device has been developed. It can classify facial expressions, specifically in the form of smart eyewear (see Figure 1-2). The dissertation uses the word "smart" in the sense that shows the function to support human by information processing with the form that fits to use in daily life. "Wearability" is important for tracking users' facial expressions for a long-term as the context of facial expression matters. The dissertation focuses on skin deformations around the eyes caused by the movement of facial muscles in order to detect facial expressions in an efficient and minimally obtrusive way. The expression of eyes is an important part of the user's emotional expression [52]. The system uses the arrays of photo-reflective sensors that are integrated into the front frame of the glasses to detect the skin deformation around the eyes. The sensors used for the prototype are small enough that they can be potentially integrated into glasses for everyday usage. The approach improves the usability in a daily life setting, compared to the camera-based systems. The device is also "wearable" in terms of social acceptability as the design follows that of conventional eyewear. The high wearability can, in turn, translate into higher tracking ability. It is the kind of wearable device that can provide an excellent vehicle for understanding the user's affective patterns in day-to-day scenarios, as discussed by Picard et al. [76].

To summarize, the dissertation investigates the following research questions(RQs):

- RQ1: What is an effective and appropriate technology to recognize facial expressions in daily life? It leads to the following subquestions. How can the wearable device recognize facial expressions? How can a wearable device be designed in a socially acceptable and comfortable fashion? How can the technology be designed for an extended usage time?

- RQ2: What is a subset of facial expressions that can be robustly recognized by the new device?

- RQ3: How can spontaneous expressions be classified with the new device? This question leads to the following sub-questions.

    - 3-1: How can spontaneous facial expressions be captured using posed emotion labels?
    - 3-2: How can diverse expressions aside from basic emotion labels be captured by the new device?
    - 3-3: How can the new device capture the information aside from facial expressions?

- RQ4: How can this technology be used for interactions with a computer-application using intentional gestures?

Accordingly, the contributions of the dissertation are

1. Designing a truly wearable device that is socially acceptable and can be used in various situaions[RQ1]. The device is designed and implemented in the form

of fully-packaged, conventional-looking eyewear, so the device comfortably fits the context of daily usage. The arrays of photo-reflective sensors on the device cover most movements of the facial muscles. To this end, the device used small photo reflective sensors that are small and light enough to be integrated into everyday glasses. The method is 10,000 times efficient for classifying basic facial expressions than the camera-based methods.

2. Recognizing the user's expressions in the eight universal categories (neutral, happiness, disgust, anger, surprise, fear, sadness, and contempt) with the wearable device in a reliable and robust manner[RQ2]. The acquired data are applied to SVM for robust classification. It evaluated the robustness towards usage in a daily life setting: when the user changes the head direction; when the user uses the device on different days; when the user walks; and when the device slips down the nose. The averaged accuracy of classifying facial expressions is 92.8%. Among basic emotions, neutral, happiness, anger, contempt are relatively classified with high accuracy, and the system achieved more than 80% accuracy for all the expressions.

3. Classifying spontaneous facial expressions[RQ3]. The long-term distribution of the user's spontaneous facial expressions was captured in a daily life setting using the posed emotion labels. Also, facial expressions in daily conversations were mapped and summarized in an unsupervised manner. Moreover, the clusters of smiles were made across individuals. In the reading activities, the device could capture the facial response and the eye movement(blinks and line breaks) using independent component analysis.

4. Developing gesture detection algorithms[RQ4]. The system can detect eye gestures using a Dynamic Time Warping(DTW) and 1-Nearest Neighbor approach. The averaged accuracy of classifying eye gestures is 92.9% per user-dependent training with 9 participants. The system can also detect hand-over-face gestures of rubbing to 10 areas with a Random Forest approach. The averaged accuracy is 91.1% per user-dependent training with 5 participants.

1.4 Organization

This dissertation investigates how to classify various facial gestures in everyday life using smart eyewear. The second chapter makes clear the position of this dissertation by comparing other measuring technology such as camera and EMG. The chapter also describes the meaning of the facial expressions the dissertation aims to detect by introducing its application and further possibilities. Chapter 3 introduces the core technology of this dissertation: the eyewear device with embedded photo reflective sensor arrays. Chapter 4 validates the eyewear device in terms of the basic emotions of facial expressions. This chapter is mainly based on the work [58,59,60]. In Chapter 5, the detection of spontaneous expressions is discussed. The chapter includes spontaneous emotion-related facial expressions classification, facial expression classification in daily communication, and reading activity analysis that includes cognition-related behavior such as blink and eye movement. Chapter 6 focused on the further possibility of the device for interaction related to face. This chapter is extended from the work [62]. Chapter 7 concludes the dissertation.

Chapter 2

Related Work

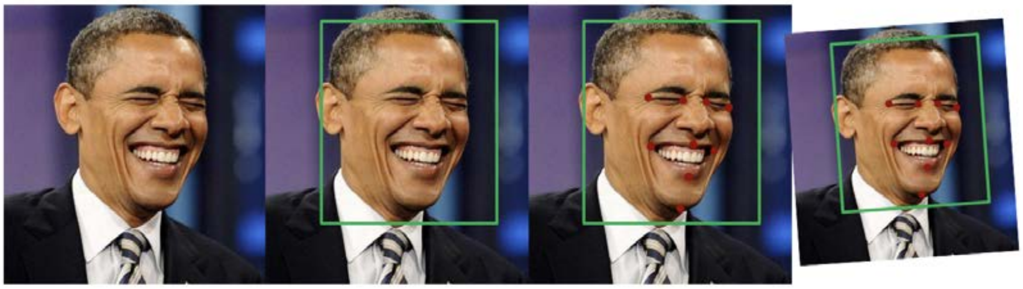


Figure 2-1: The preprocessing consists of face tracking, the facial landmark detection and normalization. The image is retrieved from [57]

This chapter summarizes the computer vision based approach. Then, the following subsection introduces the other methods to recognize facial expressions and its application. In the last section, the chapter explains the position of the dissertation.

2.1 Computer Vision Based Facial Expression Analysis

There are three primary approaches to coding facial expressions. Sign-based approach tries to associate Action Units of the face with the emotions. Message-based approach directly interprets emotion from faces. Other is a dimensional approach. Overall, these approaches focus on emotion interruption from facial expressions. According to the survey papers [1, 9, 57, 83], the basic pipeline of facial expression recognition consists of three steps. The first step is preprocessing(face tracking and detection, see Figure 2-1). The second step is feature detection, and the third step is expression classification. The first step of face detection can be done by detecting permanent facial features such as eyes or the dense set of features such eye contour areas. For the detection, active appearance models are the standard choice [11]. It detects the face by matching the pre-trained statical model of face appearances. The current state of the art algorithm uses Deep Neural Network (DeepFace) [92] or a Faster R-CNN based approach [77]. The CNN-based model is available online [46]. This step includes the registration of the faces, which deal with the normalization, resizing and rotation of the head-poses to the reference model. For the feature extraction, geometric features(eyes, mouth and so on) or appearance features such as skin texture are mainly used. The geometric features can be referred to the facial landmark, which is mostly detected by the Supervised Descent Method (SDM) nowadays. The appearance features are extracted by such as Local Binary Patterns (LBP) features, Histogram of Oriented Gradients (HoG) features, Gabor features, Non-Negative Matrix Factorization (NMF) methods, Deep Convolutional Neural Networks-based features and autoencoders-based features. The Gabor features are popular as it is similar to a human visual system. The features are the combination of applying various Gabor filters that detect the specific frequency patterns using convolution. To the extracted features, machine learning is applied to interrupt facial expressions. Machine learning methods are mostly supervised methods. These approaches use extracted facial features to associated categories such as emotions or action units. The methods are divided into static modeling and Temporal modeling. The temporal model usually considers the four-time stages, neutral, onset(start), apex(peak) and offset(end). These are done by the algorithms such as Artificial Neural Network, Hidden Markov Models, Support Vector Machines, Boosting and so on.

The datasets are essential to develop and evaluate the algorithm to detect facial expressions. In other words, the computer vision approach is advanced with the datasets. There is a database that manually labeled the posed facial expression with basic emotions such as Cohn-Kanade (CK) datasets [40, 55] and MMI dataset [99]. They label basic emotions because it was said to universal [16]. Those datasets made the directions of many researchers. They have dedicated to raising the accuracy of the classification in such datasets. However, the datasets differ from a naturalistic environment where the illumination condition and head-pose changes. Besides, in real life, people do not make an exaggerated posed expressions [83]. More recently, researchers explored to gain the meaningful information from facial expressions rather than basic emotions. To detect spontaneous facial expressions, Wan and Aggarwal developed the robust metric learning approach [102]. Girard et al. measured spontaneous facial expressions in unscripted communications [25]. Aside from spontaneous facial expressions, Researchers investigated the estimation of the expression intensity [108], the detection of micro-expression that suggests real human feeling [49].

The current state of the art achieves high performance in the experimental condition. For example, Shan et al. [87] achieved 95.1% with the LBP features and SVMs using the CK datasets. Jeni et al. [38] used the SVMs with the NMF methods and achieved 99.0% in recognizing basic emotions using the CK datasets.

2.2 Sensing Facial Expressions with Wearable Devices

One of the first attempts to detect facial expressions with a wearable device was Expression Glasses [84], which can recognize specific facial expressions (confusion/interest) by measuring facial muscle movement with piezoelectric sensors. Gruebler and Suzuki designed a wearable device that can read positive facial expressions using facial EMG signals [30]. Their device has to be attached to the side of a face, but it can record the user's affective state for more than four hours with high accuracy. These prior works used contact-base sensors. Inzelberg et al. used dry, soft electrodes array attached to the cheek area to detect various kinds of smiles [34]. While they performed well, the measurement processes require continuous physical contact, meaning that the sensors/electrodes need to be attached to the user the whole time. This need for physical contact can make the user experience rather uncomfortable, especially over a more extended period of time.

Kimura et al. presented an eyeglass-based hands-free video-phone. The glasses have multiple fish-eye cameras to capture a wearer's face and can yield his/her self-portrait facial expression image [45]. Although it can reconstruct facial images of the wearer, the powerful processor to process the images were required. It made the device bulky and not suitable for daily use.

Fukumoto et al. used photo-reflective sensors attached to the glasses to capture skin deformations at the corners of eyes and cheeks that occur with happy facial expressions [23]. They then used threshold-based clustering to distinguish smiles from laughs. While efficient, this method does not scale well for multiple users because of the individual variations in determining the appropriate threshold (i.e., How much the skin around the eyes moves while smiling or laughing varies from person to person). Besides, due to the limited number of sensors, it can miscategorize other facial expressions as the target ones. There are some other exciting applications involving a limited number of photo-reflective sensors. For instance, Nakamura et al. proposed a device with one photo reflective sensor to detect the natural movement of eyebrows when users try to focus and stare at something [66].

2.3 HCI Application Using Facial Expression Recogniton

The two main roles of facial expressions are affect and communication. First, the section focuses on the application related to the affect. Then, the communication-related application was described.

A device that recognizes facial expressions may open up new opportunities for more naturalistic user experience in human-computer interactions since facial expressions provide rich information about our emotional states [42]. It is related to the area of Affective Computing that explores the possibility of incorporating human affection in computing [4, 75]. The applications of the affective computing can be divided into three areas: 1)Affect detection, 2)Affect expression, 3) Emotional computers. This section mainly focuses the works of the affect detection from human face and interaction techniques based on it. Many of the works can be described from the perspective of Cannon-Bard theory. First one is marketing. It tries to improve the contents based on the associated affect. McDuff et al. applied automated facial coding for media measurement [94]. By using the web camera, they analyzed the facial responses to the media at scale. These data can be used in market research such as testing video advertising contents or tagging emotions to the contents. The second application is information retrieval based on facial expressions. Fukumoto et al. 's work enhanced the interesting memory by retrieving the picture from the life-logging camera only when the wearer of the device smiled [23]. The third application is therapeutic interventions and support for medical professionals. Gruebler and Suzuki developed wearable EMG device that can be attached to the sides of the face. The device could quantify the smiles and frown faces of the wearer so that the human coders can objectively measure the affect.

On the other hand, some applications are based on James-Lange theory and the facial-feedback hypothesis [37, 96]. The main idea is to regulate or enhance the emotional experience by activating facial expressions. Tsujita and Rekimoto outlined a primary example: We can increase people's happiness by making them smile more [98]. They made the controlling system for home appliances by smiles. For example, the user had to smile to unlock the refrigerator. Yoshida et al. developed an emotion-evoking system [106]. This mirror system changes facial expressions artificially. Their experiment revealed that the artificial facial expression could change the user's emotion. Their work shows the high-level concept of the hypothesis: by perceiving the user's smile, the user can feel the associated emotions.

The other applications based on the communicative roles of facial expressions. One of the major application is the facial performance capture system. It is popular for animating CG avatars in video games and movies. The system used the faces as the communicative medium to the audiences or the players. The marker-based system is commercially available such as Expression by OptiTrack Ltd and CARA by VICON. The markerless approach is also explored with an RGB monocular camera [6]. Recently, automatic face reenactment method became possible with realistic images in real time [94]. Li et al. used a depth camera to capture expressions on the lower half of the face, and eight strain gauges to capture expressions on the upper half of the face inside a head-mounted display [54]. They also mapped the input signals to a 3D face model. This method can make social interaction in virtual world smooth. Another application is FaceShare [91] and Smart Face [67]. The systems artificially can modify the facial appearance of the people in videos to smiles. With the system, the users during a video conference or chatting can communicate smoothly and enhance creativity.

It is worth noting that most of the applications described above rely on smiles among various facial expressions. It is because smiles are connected to positive emotions, and varies in different contexts, and shows different meanings.

The applications based on the cognitive aspect of facial behavior were explored. Nakamura developed a photo-reflective sensor-attached glass [66]. The attached photo reflective sensor detects the natural movement of the eyebrow for controlling amount of augmented reality information. It enables natural interaction because we narrow eyebrows when we focus and stare at an object. The pain estimation from facial expressions is another example [79]. They estimated the pain of patients using neural networks. Fatigue detection gains researchers' interest. Yao et al. developed the automatic driver's fatigue/drowsiness detection system from the facial features and head motion from input videos [105]. The other application is a tutoring system. The facial behavior from videos could be used to estimate the difficulty level of a video lecture [103].

2.4 Position

Based on the discussions in the previous sections, the section describes the dissertation position. Figure 2-2 shows the mapping of the dissertation and its related works from the viewpoint of (horizontal-axis) "The number of facial expressions that can be classified (information amount of facial expression)," and (vertical-axis) "The number of facial expressions divided by the amount of data (information density of facial expressions)." The former notion describes how diverse facial expressions can be detected. For this axis, the computer vision based method is superior as it can capture the geometry of facial expressions [6]. Facial performance capture has more information about the user's face than just recognizing facial action units from the same video inputs. On the other hand, the previous wearable sensing based methods can only specific facial expressions. For example, Fukumoto et al [23]'s method can

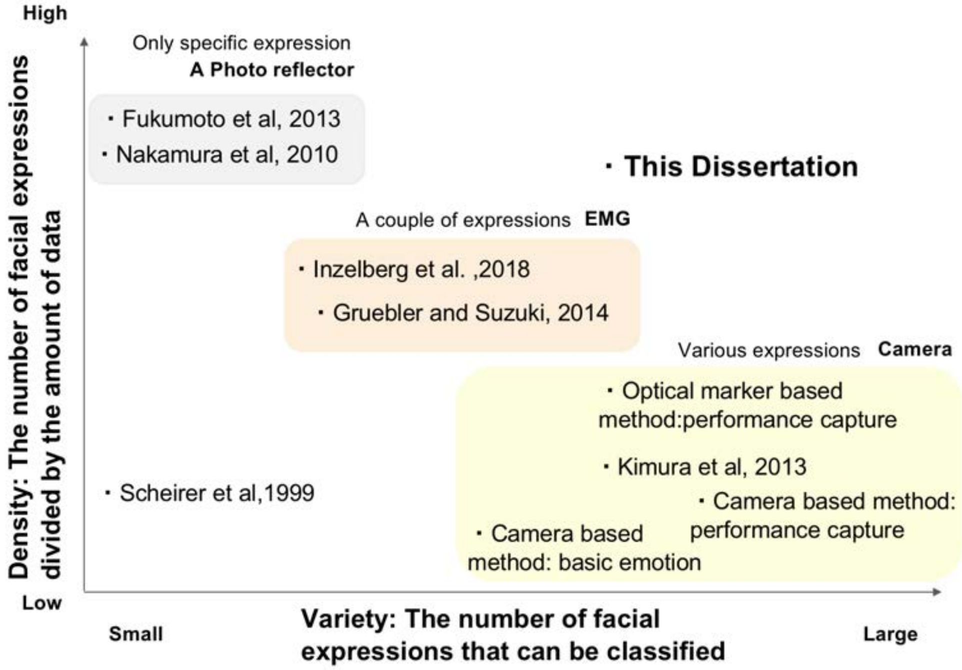
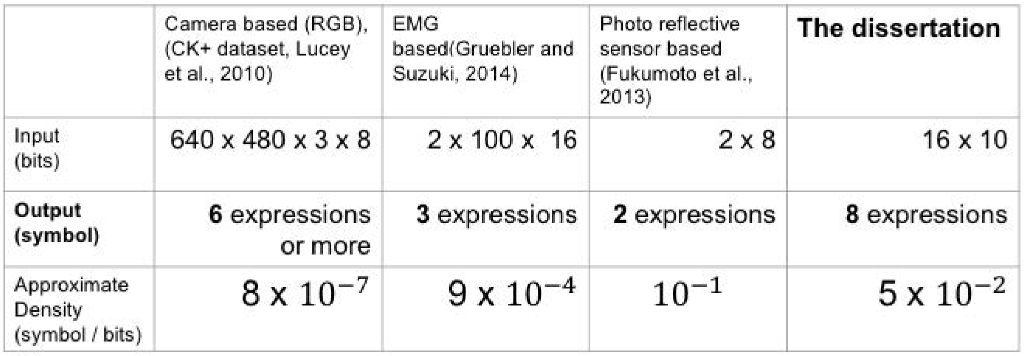

High
Only specific expression
A Photo reflector
・Fukumoto et al, 2013
・Nakamura et al, 2010
・This Dissertation
A couple of expressions EMG
・Inzelberg et al. ,2018
・Gruebler and Suzuki, 2014
Various expressions Camera
・Optical marker based method:performance capture
・Kimura et al, 2013
・Camera based method: performance capture
・Scheirer et al,1999
・Camera based method: basic emotion
Density: The number of facial expressions divided by the amount of data
Low
Small
Variety: The number of facial expressions that can be classified
Large

Figure 2-2: The positioning of the dissertation comparing to its related work.



	Camera based (RGB), (CK+ dataset, Lucey et al., 2010)	EMG based(Gruebler and Suzuki, 2014)	Photo reflective sensor based (Fukumoto et al., 2013)	**The dissertation**
Input (bits)	640 x 480 x 3 x 8	2 x 100 x 16	2 x 8	16 x 10
Output (symbol)	**6** expressions or more	**3** expressions	**2** expressions	**8** expressions
Approximate Density (symbol / bits)	8×10^{-7}	9×10^{-4}	10^{-1}	5×10^{-2}

Figure 2-3: The comparision of the computer vision-based method, EMG-based method, photo-reflective sensor-based method and the dissertation.

only detect smile and laughter. The wearable EMG-based method can capture a couple of expressions [30,34]. The latter axis suggests how much processing is required to get the information of facial expression. It can refer to the density (see Figure 2-3). It is the vital aspect of wearable devices because mobile devices cannot process a significant amount of data like PCs. The high density of information can make the wearable devices compact, small,lightweight and comfortable to wear. The comparison of the three methods can be seen in Figure 2-3. For the camera based methods, they require the preprocessing and feature extraction as they usually have high-dimensional data. For example, the camera based method for classifying six basic emotions requires 640 x 480 x 3 pixels x 8-bit inputs (7372800). EMG-based methods require less processing cost than camera-based method. However, the methods still require high sampling data and feature extraction using mainly ICA. Gruebler and Suzuki's method used 2 x 100 x 16 bits (3200) for classifying positive expressions. Photo-reflective sensor-based methods such as Nakamura et al. and Fukumoto et al. works did not require feature selection, and the processing cost is low [23, 66]. Fukumoto et al.'s method uses 2 x 8 bits (16) for classifying smiles. As such, the photo reflective sensor method requires the amount of data 10,000 times less than the camera-based methods.

Considering the two perspectives of the facial expression recognition technology, the proposed method in the dissertation has novel positioning as the method can classify various facial expressions using low-processing sensor array that have abundant information about facial expressions. On the other hand, the dissertation's method has other advantages such as contact-less and small energy consumption.

Chapter 3

Smart Eyewear for Facial Expression Classification

The dissertation aims at classifying various facial expressions in daily life. As such, the technology needs to measure various facial expressions for a long time and with little demands. The demand includes the physical demand, the social weight, and the energy demand. Therefore, the technology should always be available to the user, and comfortable to wear, socially acceptable, and consume small energy.

The dissertation designs the technology in the form of eyewear in the research. Eyewear computing is a promising technology for facial expression and affects recognition in real life. Since the head is the primary location for most of the human senses, eyewear computing can gain access to a variety of physiological signals by placing sensors in the head area. There remains a design problem as anything worn on the head is quite noticeable, yet with increasingly smaller Printed Circuit Boards (PCB), sensors, and actuators, it is possible to now build smart glasses that are similar in appearance from ordinary eyewear, making them socially acceptable concerning both appearance and comfort. The form of the eyewear is essential criteria as its shape is already standard in daily life.

As a proof of the concept, the dissertation has developed the two prototypes in the form of eyewear. For sensing modalities, the prototype devices made use of the arrays of photo-reflective sensors. The second prototype was designed based on the first prototype, so the basic structure is the same.

3.1 Photo Reflective Sensor

Photo-reflective sensors are sometimes used in the field of Human-Computer Interaction to measure human skin deformations [23,66,70,71]. The prototypes use infrared (IR) reflective sensors. The sensors are composed of an IR LED and IR phototransistor.

To establish the fundamental characteristics of skin surface reflection captured by an IR photo reflective sensor, the experimenter measured the voltage from the sensor (SG-105 by Kodenshi) (see Figure 3-1). The sensor value changes by the distance between the sensor and the skin surface. The experimenter collected 30 samples at

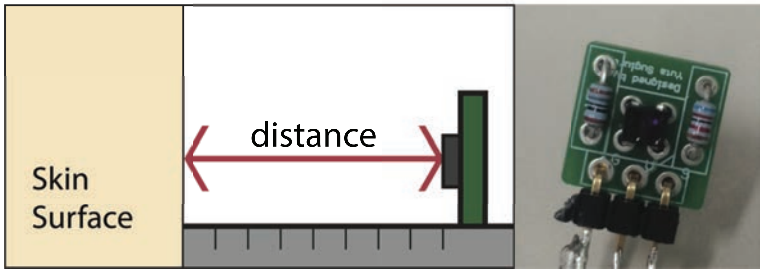


Figure 3-1: (a)The experimenter measured the distance between a photo reflective sensor and skin surface. (b)PCB used for this experiment.

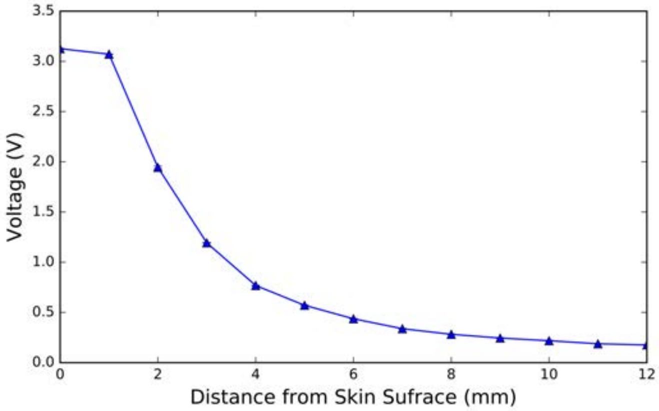


Figure 3-2: The voltage change of the photo reflective sensors related to skin surface distance.

each position. Figure 3-2 shows the average and standard deviation at each distance. The standard deviation is quite small (at most 0.014 V). The correspondence is not linear. The sensor can obtain the proximity to the skin. For closer distances, the photo-reflective sensor has a higher resolution.

3.2 Principles

The optical sensor measures the depth change between the sensor and the skin surface of the wearer.

To capture facial expressions, the system leverage skin deformations caused by the movement of facial muscles (see Figure 3-3). When users move their facial muscles, three-dimensional deformations occur on the skin surface. With the depth informa-

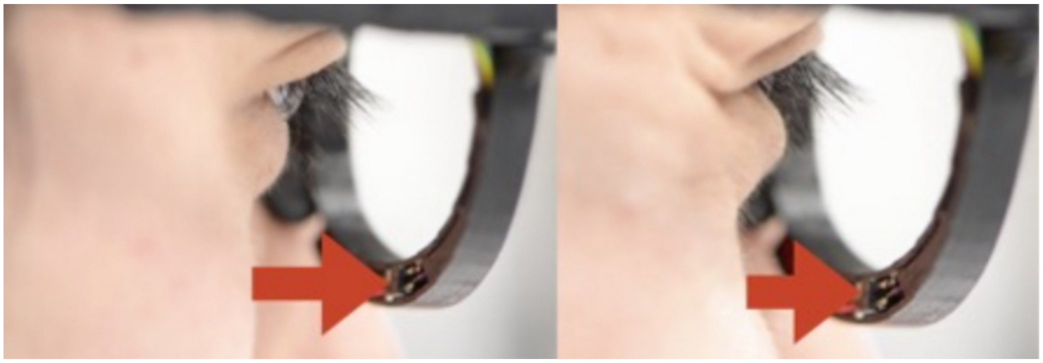


Figure 3-3: Skin deformations change the distances between the sensors and the skin surface. The deformations occur when the facial expression changes.

tion, the sensor can measure the movement of facial muscles since the movement of facial muscles causes the skin deformation around eyes. Since the photo reflective sensors are embedded in the various spots of the eyewear device, it is possible to detect various movements of the facial muscles of the wearer. Each facial expression involves different movements of facial muscles. The movements of the eyelids, the eyebrows, the nose, and the cheeks all cause three-dimensional skin deformations around the eyes. The movement of the mouth also causes the skin deformation under the eyes because the muscle movement around the mouth causes a cheek deformation that extends to the area below the eyes. According to [18], these movements are the greater parts of Action Units (AUs) with which the Facial Action Coding System codes human facial expressions. Therefore, placing sensors to capture the skin deformations around the eyes makes it possible to detect most muscle movements related to the target facial expressions.

3.3 Prototypes

Through the dissertation, two prototypes were developped.

3.3.1 First Prototype

Figure 3-4 shows the components of the first prototype. The prototype incorporates 17 photo reflective sensors (SG-105 by Kodenshi, the placement can be seen

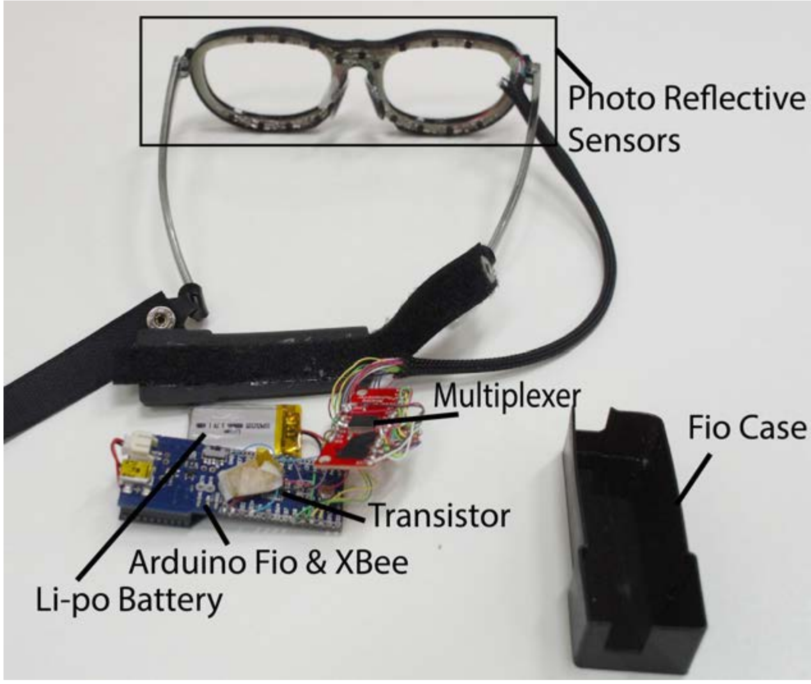


Figure 3-4: System components of the first prototype

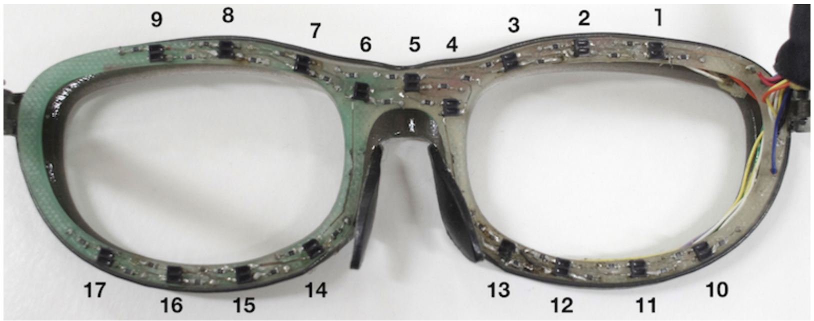


Figure 3-5: The placement of photo-reflective sensors for the first prototype

in Figure 3-5), a 16-channel multiplexer (CD74HC4067 by Sparkfun), a transistor (IRLU3410PBF by International Rectifier), Arduino Fio, Xbee, and lithium polymer battery. The weight of the prototype is around 60g. The front frame is 3D printed, and the temple tips are taken from regular commercial eyewear. An eyewear band is added to stabilize the position of the eyewear. The transistor is used to modulate the LED of the photo reflective sensors because the sensors are easily influenced by ambient light such as the fluorescent lighting in the environment. The measure the difference between the values with LED on and off. The switching frequency is around 80 Hz. With this method, it is possible to reduce the influence of ambient light. Xbee enables serial communication via ZigBee at 57600 bits per second. This device was used for recognizing basic emotions in Chapter 4.

3.3.2 Second Prototype

Figure 3-6 shows the second prototype. PCBs for sensor units (the front frame) and microcomputer units (temples) were custom-made. Temple tips were chosen from the commercially available goods. The other parts such as nose pad, hinges between PCBs were printed with a 3D printer (Form 2 from Form Lab). The nose pad can be replaced to fit the shapes of users' noses. A strap fastening tool was used to stabilize the position of the devices.

The 16 photo reflective sensors (NJL5901AR-1-TE1 produced by New Japan Radio Co., Ltd.) were placed on the front frame of the eyewear prototype. Figure 3-7 shows the sensor layout. The system used different resistant values to the phototransistors because the curvature of a face changes the distance range measured by the sensors. The system used the lower register values for the phototransistors of the sensors that measure close distance, i.e., the sensors close to the center of the front frame than the ones in the end.

The prototype measures the skin deformation around eyes. The difference from the first prototype is that the prototype can measure the eye movements in addition to facial expressions. It measures the area close to eyelids instead of eyebrows because the position of the nose pad is different. Since the movements of eyeballs cause the

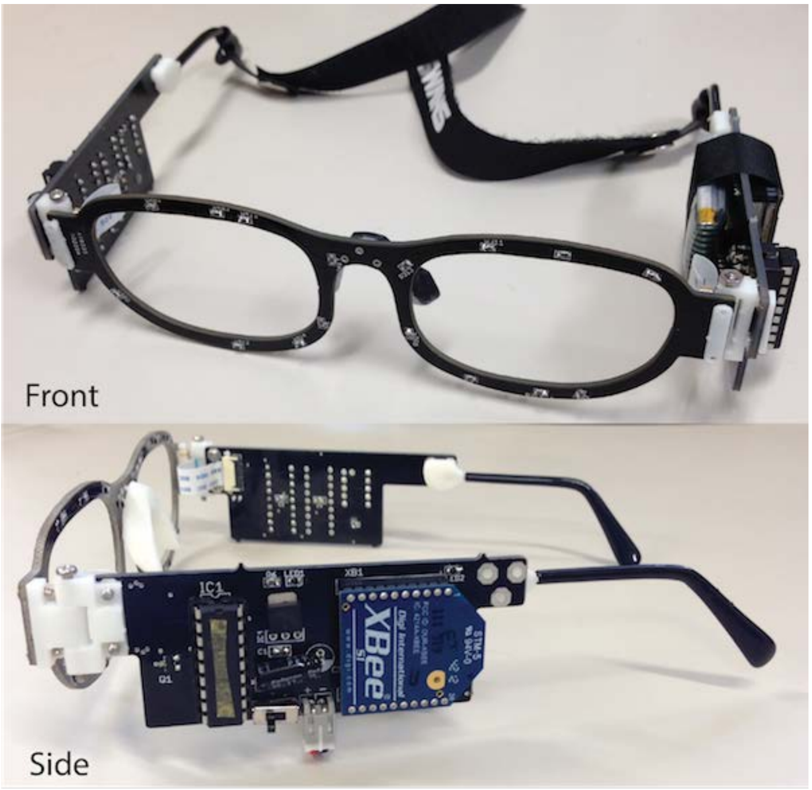


Figure 3-6: The appearance of the second prototype. It includes 16 photo reflective sensors on the front frame. On each side, microcontrollers are placed.

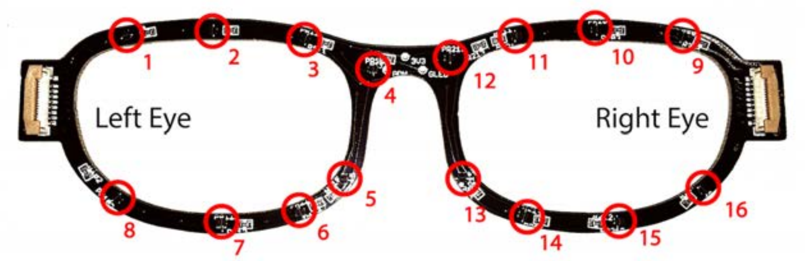


Figure 3-7: The layout of the sensors on the second prototype. The sensors are distributed into the area all around eyes.

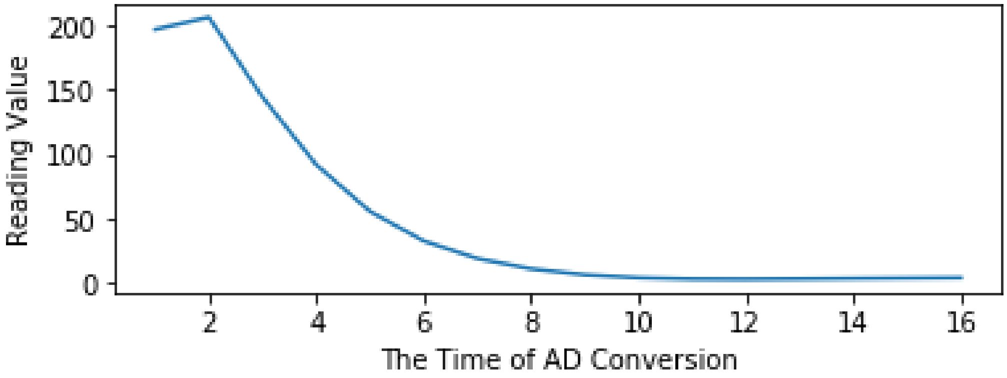


Figure 3-8: The standard deviation of AD conversion for each time

deformation around the eyes and eyelids, the values from the sensors change according to the eye movements. Like the first prototype, the change of facial expressions causes skin deformation to the area measured by the sensors on the device.

One PIC microcontroller(16F1827) was placed on each temple. Each PIC converts the voltage from each of the eight sensors to 10-bit digital value. For every PIC, one transistor is placed to turn the infrared LED of the sensors on and off (it reduces the influence of the ambient light). XBee transmits a data sample wirelessly to a laptop. The 3.7v lipo battery powers the microcontroller after the regulator controls the voltage to 3.3v. The weight of the device is around 70 g. This device was used for the Chapter 5 and the Chapter 6.

There are two gimmicks to reduce the noise of AD conversion: reduction of ambient light effect based on the characteristics of a phototransistor and faster AD conversion using two PIC microcomputers (16f1827).

The transistors turn off the LEDs first and convert the signals from all the sensors to 10-bit values. Then, they turn on the LEDs and perform AD conversion for all the sensors. Since AD conversion is executed in parallel using two microcomputers, the time required for AD conversion of all the sensors became a half compared to the use of only one microcomputer. After turning off the LEDs, the system calculates the subtraction between the data when LEDs are on and off. If the value got negative, the system changed the value to zero. Then, it sends the data to the laptop where the data is stored using serial communication (XBee).

This process creates enough time for leaving off the effect of the storage time when the influence of the reflected light remains. It takes longer if the resistance of the phototransistor is large. Even if the LEDs of the sensors are turned off, it cannot measure the effect of the ambient light immediately. The second prototype used 39k - 68K resistors. To investigate how long it takes to get rid of the effect, the experimenter measured the intensity to phototransistor while the experimenter randomly changed the distance between the sensor and a small object. The experimenter kept the condition of ambient light during the experiment. Figure 3-8 shows the AD-converted values when the system turned on the LEDs, then the system made two AD conversions, turned off the LEDs and made 14 AD conversions. One AD conversion took approximately 40 microseconds, and the system took 20 microseconds interval between each AD conversion. This figure shows that 11th AD conversion does not get influenced by the storage time. At that time, the standard deviation was less than 4. The second reading has a more relevant to the distance than first. For acquiring 16 sensor data, it takes (30 + 40) x (16/2) = 560 microseconds since it takes 30 microseconds to determine the sensor for AD conversion. For when LEDs is on, it takes (30 + 40 + 20 + 40) x (16/2) = 1040 microseconds because the system made two AD conversion for each sensor signal. Overall, it takes 1600 microseconds to obtain sensor data. Since serial communication takes more than 1.0 milliseconds, the system can cut the time to wait for a storage time by obtaining the data when LEDs are off first. Also, faster conversion using two microcomputers reduces any temporal noise from ambient light.

3.4 The Advantages

The system uses a large number of photo-reflective sensors and applies a machine learning method for measuring facial expressions. This approach has the following advantages: (1) An adaptive and robust facial expression classification that works with a variety of users, enabled by the richness of the sensor information and the machine learning framework. (2) Non-contact measurement: The sensors are unobtrusive and do not require physical contact, which improves the wearability of the device. (3) Smart appearance: The sensors are small enough to be integrated into everyday glasses (e.g., NJL5908AR by New Japan Radio Co., Ltd: 1.06 x 1.46 x 0.5 mm), making the device suitable for everyday usage. (4) Simplicity: The processing required to interpret the sensor readings is minimal as face recognition, facial features extraction are not required, so no elaborate feature extraction is necessary. The system can, therefore, keep the processing cost and energy consumption little, which is crucial for practical real-time classification for long-term use. (5) Affordability: The device only uses photo-reflective sensors with a microprocessing unit and can be manufactured at a low cost.

3.5 Usage Scenarios

This section introduces various scenarios using the device in daily life. There are five scenarios to implicate the future direction. The first scenario is relatively easy to implement with the current system. Next two scenarios assume everyday use of the system. Those scenarios require further evaluations in more realistic situations for a long time. Last two scenarios show the potential of the devices assuming the system could capture subtle expressions and used by many people. The first scenario is "Collaborative Media Tagging." Facial expressions or emotions can be used to evaluate contents. If many people read a book or watch a video with their facial expressions recorded, the content can be indexed and made searchable.Also, content creators could obtain a better grasp on their techniques, making their storytelling more effective (e.g., what surprises a reader/watcher and what does not).

The second is "Care System for Older Adults." The system can provide how a user changes facial expressions in daily life. In other words, the system can quantify facial activities. This information can be used for a care system for older adults. They can get an overview of their situation (and potential deterioration of the mood). Since more and more seniors live away from their children, they tend to feel lonely more easily than before. The system can help them and their children to take appropriate and productive care of both (e.g., the system notifies the children of the appropriate times to give their parents a call and talk to them when the system detects they are feeling sad and smiling less).

The third is "Supporting System for People with Autism Spectrum Disorders." People with autism have difficulty in creating facial expressions of emotion. As facial expression is an important source of social interaction, the system can be useful for them. The system can help them to create facial expressions by giving them a motivation to intentionally express their emotion with feedback if they could successfully make the expressions or not. Also, According to [107], "impairment in overt facial mimicry in response to others' dynamic facial expressions may underlie difficulties in reciprocal social interaction among individuals with ASD." Suppose that encouraging them the facial mimicry could improve their social interaction skill, the system used by multiple users can notify them when they can successfully achieve the mimicry and so help them to understand and encourage the mimicry.

The fourth is "Happiness Map." Given the system can achieve facial expression classification in daily life, a straightforward way to apply it is to combine facial expressions with location and demography. People can search for the place where the users smile and laugh frequently. It can be used to choose where to live or work. They can also obtain an overview of how events and policies change the affect state of a region or country. It can be a measure of the wellbeing of the places.

Finally, the device can be used for emotion regulation. The emotional impact is a very delicate problem, as people should feel empowered by technology not influenced or even oppressed. Positive emotion can have a positive effect on decision making and facilitate social interactions while negative emotion can do harm [29]. We regulate emotion to accomplish some specific goals because emotion sometimes induces an unwanted behavior. One of the emotion regulation strategies is to alter the situation people are in [28], but to do so, people have to monitor their behavior. This monitoring skill reduces when people are in negative affect(e.g., [32]). Therefore, in negative affective states, they might not regulate emotion effectively. In these cases, monitoring facial expressions with the device can help them to look back and alter their situation as emotion, and facial expression is linked to each other. Given the rise of depression and other mental illnesses related to emotional imbalances, people can benefit from devices that can help them monitor and motivate to alter their emotional state. For example, a depressed person usually talks little with others and stays in bed longer (i.e., is less physically active). If a system successfully encourages this person to be more social and active with the assistance of caring people who got his/her information of facial expressions, the work might help medical doctors to deal with his/her mental illnesses. These influences should be aligned with the user's long-term goals (not influencing them in "unwanted ways"). Of course, the proposed solution may not work in every case (e.g., when personalization is needed) and depend on the severeness.

Chapter 4

Classification of Posed Facial Expressions into Basic Emotion Categories

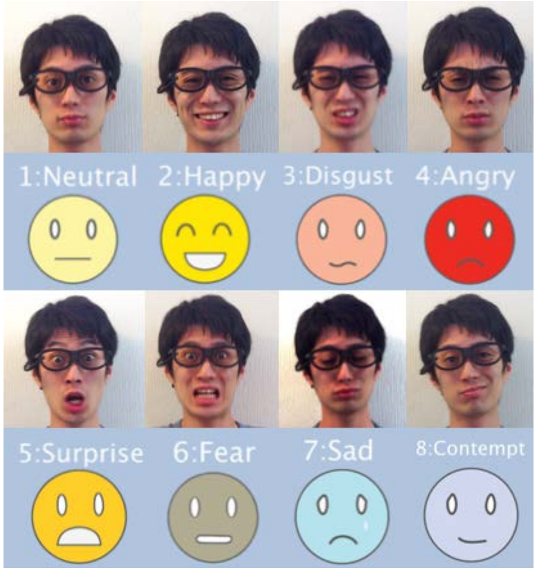


Figure 4-1: The smart eyewear classifies eight universal facial expressions

This chapter focuses on classifying eight universal facial expressions (see Figure 4-1). In addition to the universal six facial expressions (happiness, disgust, anger, surprise, fear, and sadness) defined by Ekman [16], the chapter includes contempt, which is sometimes considered as a universal facial expression [63] and "neutral" as a baseline for detection. The universality of the basic emotions of facial expressions is discussed by Ekman et al. [16]. Although some research argues the reliability of basic emotions, it is still the common framework for classifying facial expressions. For computer-vision based methods, there are various datasets that label facial expressions with the basic emotions [40, 55]. Using those datasets, many researchers improved the accuracy of classifying basic emotions. As a first step to validate the eyewear device, this chapter runs through the user study of classifying eight expressions in various situations. More complex facial expressions such as a fake smile or a "happy surprise" are out of the scope of this section. This chapter is based on the work [58, 59, 60].

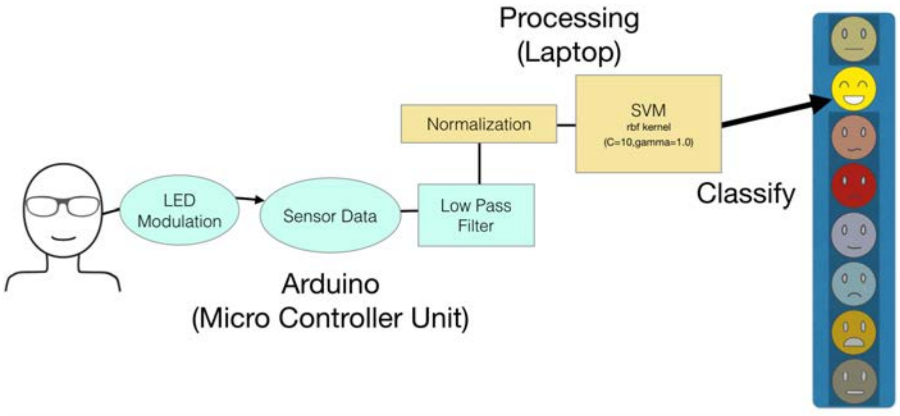
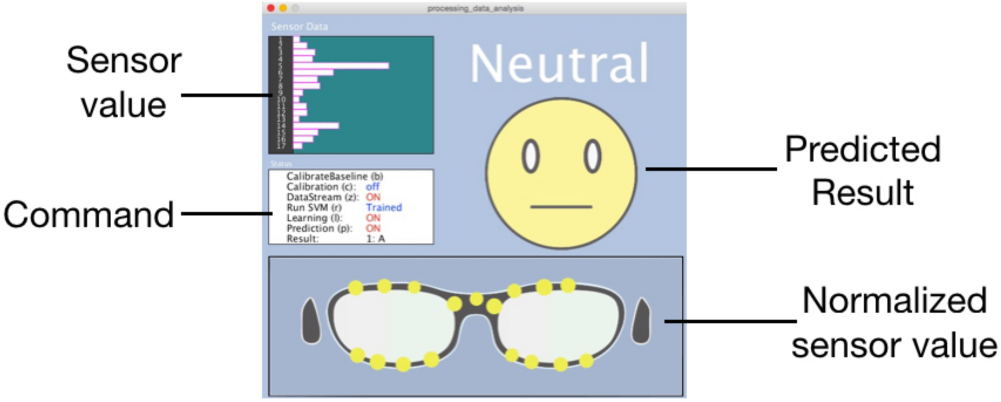


Figure 4-2: The data processing pipeline for classifying basic emotions.

4.1 Software Implementation

Figure 4-2 shows the data processing pipeline of the system. This section describes the Software details of using the first prototype.

In an Arduino environment, input from each sensor is converted into a 10-bit value. To take the difference of sensor values between LED is on and off, it multiplies minus one to the value when LED is off. The value is smoothed out by applying a moving average to five pairs(LED is on/off) of each sensor values in a row to reduce noises. One data sample is a collection of 17 elements from the sensors. Each element of the data sample is the average of 10 sensor values. The data sample is then sent to Java/Processing. In the Processing environment, the system normalized the data sample and recorded the normalized data sample with a desired facial expression label as a training set. With the training set, a Support Vector Machine algorithm (SVM) with a radial basis function (rbf) kernel (C = 10, gamma = 1.0) is applied to classify facial expressions in real time. SVM is commonly used for classification task as it learns effectively. Figure 4-3 shows the user interface of the system. The emoticon shows the classification result. The bar graph represents the raw data. The yellow dots correspond to the positions of the sensors, and the size of the dot is correlated with the normalized value of each sensor. The SVM use only one data sample for the prediction of the basic emotions. For later experiments, recorded data samples were normalized in the way described in this section.



CalibrateBaseline (b)
Calibration (c): off
DataStream (z): ON
Run SVM (r) Trained
Learning (l): ON
Prediction (p): ON
Result: 1: A

Figure 4-3: The user interface

Preprocessing for Machine Learning

1. When the user makes a neutral expression with the device on, the baseline of the elements was set in the data sample (BLV) as 0.5.

2. In this stage, the user dynamically moves his/her facial muscles. For each sensor, the range of each sensor value ($Range$) in the data sample is determined during this calibration process. Based on these values, the normalized sensor value (NSV) for each sensor is calculated as follows. An element of the data sample, which is the moving average in each time frame, is defined as $SensorInput$. $Tolerance$ is chosen as 40 experimentally.

$$Range = (Max - Min) + Tolerance \tag{4.1}$$

$$\begin{cases} NSV = 0.5 + (SensorInput - BLV)/Range \\ if\ NSV > 1,\ then\ NSV = 1 \\ if\ NSV < 0,\ then\ NSV = 0 \end{cases} \tag{4.2}$$

$Range$ includes $Tolerance$ since there is a trade-off with relying only on the acquired data.

- Advantage: Normalization can set the appropriate range for each sensor by measuring the range of $SensorInput$ that varies depending on the geometry of each user's face and the position of the sensor. The normalization can improve robustness because it can accommodate the weight of each $SensorInput$.

- Disadvantage: The position of the device during the calibration phase is not always stable. The user may move his/her facial muscles too dynamically and cause the eyewear to dislocate, resulting in inaccurate measurement of the Max and Min values. On the other hand, if the facial movement during calibration is not dynamic enough, it may reduce the amount of information that can be obtained. Therefore, it is not always possible to normalize the data sample in an optimal manner.

3. During the learning phase, normalized data samples are stored with facial expression labels that are selected as the desired outputs. A label is attached to each data sample.

4. For real-time classification, the normalized data sample is applied to SVM that is trained with the normalized data samples with the labels.

Algorithm

In addition to the training set that includes the output labels and the normalized data samples, the calculated values (CV) from two different sensors are also used for SVM. The calculation formula for (S_i,S_j | $1 \leq i, j \leq 17$) is shown below

$$\begin{cases} CV = (S_i - S_j)/2 + 0.5 \\ \\ if\, CV > 1,\ then\, CV = 1 \\ if\, CV < 0,\ then\, CV = 0 \end{cases} \tag{4.3}$$

Sensor placement is shown in Figure 3-5. The values from adjacent sensors located in the bottom part of the front frame as well as the values from the sensors that have a vertical relationship are calculated vertically. The calculated values from adjacent sensors in the bottom part are important because the surface is smooth and the skin is considered as an elastic model, thus they correlate with each other as discussed by Ogata et al. [70]. The data from the sensors that have a vertical relationship can partially inform the position of the eyewear and the face. In total, 33 dimensions were used (normalized input: 17 + adjacent data: 7 [(10,11), (11,12), (12,13), (13,14), (14,15), (15,16), (16,17)] + vertical data: 9 [(1,10), (2,11), (3,12), (4,13), (5,14), (6,15), (7,16), (8,17), (9,17)]).

4.2 System Evaluation

For the system evaluation, four experiments were conducted. The first experiment evaluated the recorded data of the test dataset immediately after user dependent training. The results suggested the user dependency of the system. Also, the dataset was evaluated for the trade-off between the number of sensors and accuracy, the robustness to changes in head direction as well as to the removal and remount of the device. The second experiment had three participants tried the first experiment for multiple days to assess the possibility of long-term usage. The third experiment tested the robustness of the system to the movement of a wearer by taking measurements while walking. The final experiment evaluated the influence of the vertical displacement of the device on recognition accuracy. The observations from a demonstration at SIGGRAPH Emerging Technologies 2015 [61] were described.

4.2.1 Evaluation 1: Basic Setup

Eight users (four Japanese, one French, one Chinese, one Taiwanese, and one Sri Lankan. Two of them female. Average age: 27.3) participated in the experiment. They were asked to sit in a chair and mimic the pictures of an American male (re-

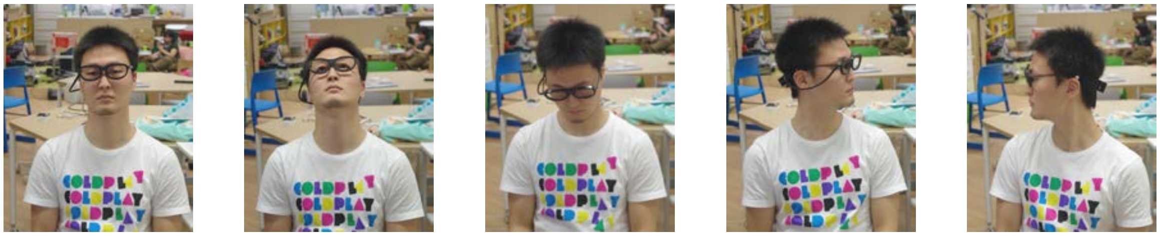


Figure 4-4: Evaluation with different face directions

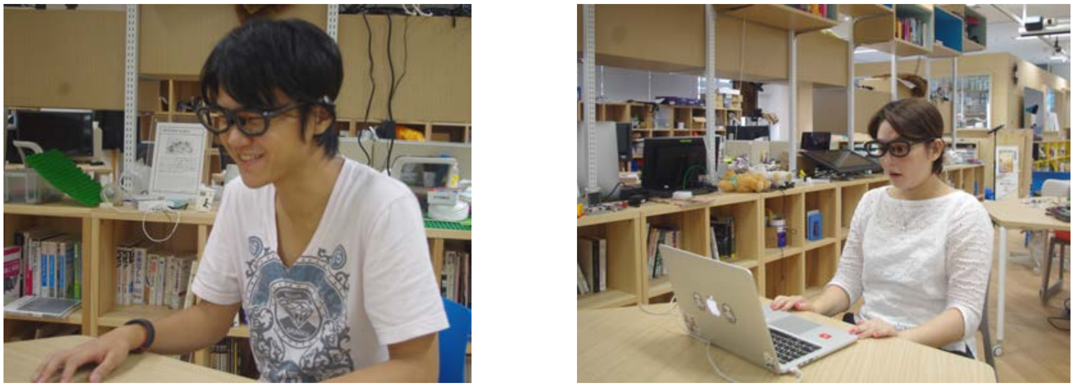


Figure 4-5: Experimental setting for basic emotion classification

trieved from the images of a man from a TV show "Lie to Me") making the universal facial expressions based on Ekman et al. [16]. First, the users looked straight ahead with a neutral facial expression, setting the baseline for the sensors. Then, they moved the facial muscles for the calibration. Then, the system collected the data samples of eight facial expressions in different poses: Looking straight ahead (three times), looking up (three times), looking down (three times), looking left (two times), looking right (two times), and taking off the device and putting it back on (two times) (see Figure 4-4). The experimenter collected data samples with different head directions because the movement of the head alone causes skin deformations as a result of the effects of gravity and the joint coupling of muscles. The experimenter manually recorded 10 data samples with facial expression labels at regular 50-millisecond intervals in the middle of the time the user kept their maximum pose of each facial expression. In other words, In one recording of each facial expression, it takes 10 data samples for 1/2 seconds. Overall, each user's dataset includes 1200 data samples (10 samples per expression per time x 8 facial expressions x 15 times in different poses. All recordings were conducted indoors. (see Figure 4-5).

Accuracy with User-Dependent Training

Figure 4-6 shows the distribution of data samples of each facial expression from User A. The sensor numbers (1 through 17) corresponds with the numbers shown in Figure 3-5. It shows that the sensor distribution is different depending of each expression. It suggests the validity of the sensor data for a classifier.

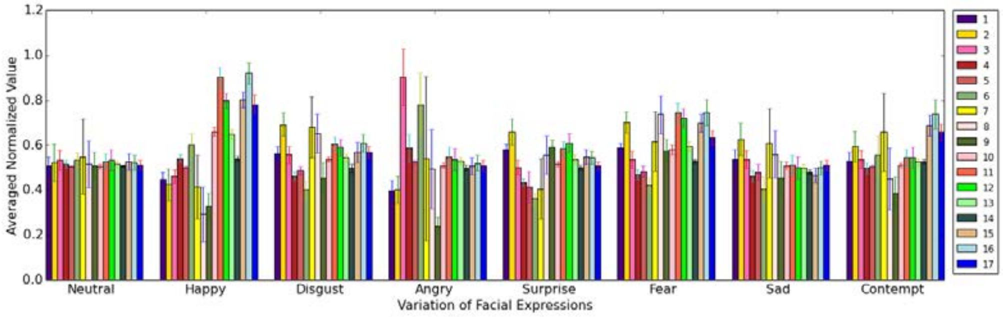


Figure 4-6: Distribution of sensor values changes for each facial expression.

First evaluation tested the accuracy in measuring each user's performance with user-dependent training. For each dataset, one recording (10 data samples) of each facial expression was divided into two sets, with the former half as a training set (600 data samples: 5 data samples per expression per time 8 facial expressions 15 times in different poses) and the latter half as a test set (600 data samples). The system applied SVM with an rbf kernel (C = 10, gamma = 1.0) to the training set. SVM was chosen among basic supervised learning algorithms such as AdaBoost, Random Forest and Naive Bayes.considering the robust accuracy in the preliminary experiment. The evaluation was done per individual. In other words, the system trained with user A's training set and tested with user A's test set. It achieved 92.8% accuracy on average (Facial-Expression-based result was 84.3% - 97.8%. User-based result: 84.8% - 99.2%.) of classifying basic 8 expressions. By learning from the data samples obtained with different head directions, the eyewear device was able to classify the facial expressions correctly, regardless of where the head was directed at the time of the measurement. Table 4.1 shows the confusion matrix of the results. As shown in the matrix, disgust can be similar to anger or fear. Besides, surprise and fear are close facial expressions with a 4.7% error to each other.

		Classified Results							
		N	H	D	A	Su	F	Sa	C
Actual Value	Neutral(N)	96.8%	0.7%	0.2%	0%	1.2%	1.0%	0.2%	0%
	Happiness(H)	0.3%	98.3%	0.5%	0%	0%	0%	0%	0.8%
	Disgust(D)	1.7%	0.2%	84.3%	3.2%	0%	5.5%	3.7%	1.5%
	Anger(A)	0.3%	0%	0.7%	96.5%	0.8%	0.3%	1.3%	0%
	Surprise(Su)	2.0%	0%	1.2%	0.2%	91.2%	4.7%	0.8%	0%
	Fear(F)	1.5%	0%	4.2%	0.8%	4.7%	87.5%	0%	1.3%
	Sadness (Sa)	4.5%	0%	2.5%	1.8%	0.8%	4.7%	85.7%	0%
	contempt(C)	1.5%	0.3%	0%	0%	0%	0.3%	0%	97.8%

Table 4.1: Confusion matrix (within subjects)

User Dependency

Second evaluation investigated user dependency by training with each user's training set and testing with all other users' test sets. For example, the system trained with User A's training set and tested with User B, C,..., F's test sets respectively. Table 4.2 shows the result. When the training set and the test set are taken from different users, the accuracy scores drop significantly: Accuracy is 48.0% at best (using the training set from User A and the test set from User G). This result indicates that, with the current system, users need to calibrate the device and train individually for accurate classification of their facial expressions.

		Test Data							
		A	B	C	D	E	F	G	H
Training Data	User A	99.2%	36.8%	31.8%	34.5%	28.8%	40.5%	48.0%	28.3%
	User B	39.2%	98.7%	23.3%	13.5%	30.0%	33.8%	32.8%	26.7%
	User C	42.5%	22.3%	84.8%	37.7%	21.0%	37.3%	40.7%	22.2%
	User D	19.8%	29.8%	33.5%	85.0%	32.0%	25.3%	30.5%	18.0%
	User E	23.3%	27.3%	18.7%	20.8%	89.3%	17.3%	39.8%	39.5%
	User F	44.3%	34.2%	28.2%	27.8%	27.7%	97.3%	35.0%	34.2%
	User G	43.8%	31.0%	24.7%	20.0%	27.7%	28.8%	88.7%	28.5%
	User H	12.5%	25.2%	15.7%	14.8%	26.2%	17.8%	29.8%	95.2%

Table 4.2: User dependency matrix

Table 4.3 shows the confusion matrix of the result using the training set and test set that are merged from all user's datasets. Though there is user dependency, happy expressions can be classified with relatively high accuracy (71.9%). The other facial expressions were harder to be classified(e.g., sadness: 17.2%, contempt: 22.7%).

		Classified Results							
		N	H	D	A	Su	F	Sa	C
Actual Value	Neutral(N)	74.9%	1.1%	6.3%	1.3%	5.2%	3.2%	2.6%	5.3%
	Happy(H)	5.3%	71.9%	2.6%	5.1%	0.4%	7.7%	2.8%	4.2%
	Disgust(D)	15.8%	7.5%	23.5%	12.2%	14.0%	17.5%	4.1%	5.2%
	Angry(A)	14.9%	7.8%	13.4%	26.6%	13.8%	15.5%	4.9%	3.1%
	Surprise(Su)	27.1%	3.2%	10.1%	4.1%	30.9%	11.7%	8.9%	4.2%
	Fear(F)	14.5%	6.5%	11.4%	8.8%	17.3%	27.7%	5.5%	8.2%
	Sad (Sa)	27.8%	8.6%	7.4%	6.4%	15.9%	13.6%	17.2%	3.1%
	contempt(C)	29.5%	7.8%	8.6%	5.0%	11.7%	11.2%	3.7%	22.7%

Table 4.3: Confusion matrix(between subjects)

Number of Sensors

Next, the evaluation was done to see the trade-off between accuracy and the number of sensors. This evaluation used the datasets from all eight users for this purpose. The training datasets are made of eight users' training sets, and the test set consists of 4800 data samples merged respectively. As in 4.2.1, SVM was applied in the same way. Using the training sets, the forwarding algorithm to pick out the sensor to be used for SVM. The algorithm began by choosing the values from only one sensor that had the best accuracy based on the result of SVM. Next, it added the values from another sensor with the second-best accuracy. It repeated the process until all values of 17 sensors were included. For this analysis, the subtraction of the adjacent data cannot be considered. Only the normalized 17-dimensional data was applied to SVM.

As shown in Figure 4-7, the experiment yielded 84.1% accuracy with 17 sensor values. The accuracy improved with the addition of values from more sensors. With the values from 13 sensors, the system achieved more than 80% (81.0%) accuracy. Sensors 1, 9, 12, and 15 were left out. Sensors 1 and 9, as well as 12 and 15, are symmetrically located at the opposite ends of the frame. The further the sensors are from the center of the face, the greater the distance between the eyewear frame and the skin surface on the face becomes. The greater distance results in the sensors having less information because they are more vulnerable to the noise of ambient light. Photo-reflective sensors work well when the distance between the sensor and the target is less than 10.00 mm.

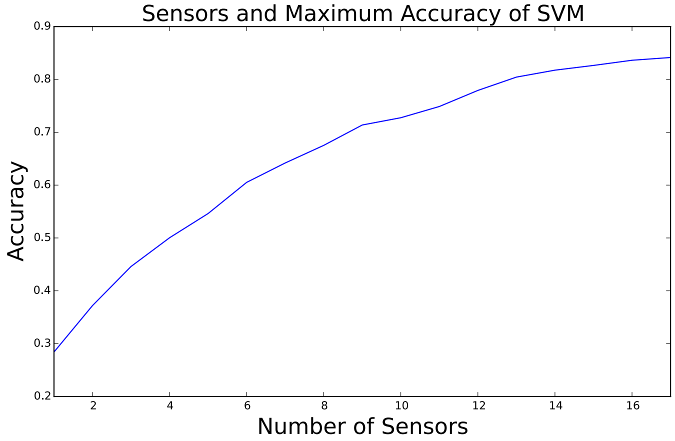


Figure 4-7: Trade-off between recognition accuracy and number of sensors

Accuracy when User Changes their Head Direction and Removes/Remounts the Device

The previous evaluations have already shown that the system can classify facial expressions even when there are changes in head direction and when the user removes and remounts the device by obtaining data samples at those conditions in the training phase. This evaluation shows how those conditions influence accuracy using the same dataset as before. The system was trained with the data samples obtained when the user looked straight ahead and tested with the data samples obtained in other conditions. The result is shown in Table 4.4. Accuracy varies among the users, but mostly it is between 50% - 60%, indicating the relative robustness of the recognition system.

	Condition					
	Upwards	Down	Left	Right	Take Off&ON	Average
User A	86.7%	72.9%	84.4%	93.8%	79.4%	83.4%
User B	67.5%	50.4%	48.1%	56.9%	45.6%	53.7%
User C	35.4%	66.7%	56.9%	46.2%	53.1%	51.7%
User D	24.6%	32.1%	65.6%	28.1%	64.4%	43.0%
User E	44.2%	47.9%	63.1%	58.1%	68.1%	56.3%
User F	50.0%	58.8%	59.4%	66.9%	40.0%	55.0%
User G	40.4%	45.8%	61.9%	43.8%	53.8%	49.1%
User H	81.7%	37.5%	46.9%	29.4%	55.0%	50.1%
Average	53.8%	51.5%	60.8%	52.9%	57.4%	

Table 4.4: Accuracy by different conditions compared to looking straight ahead

4.2.2 Evaluation 2: Reliability over Time

The second experiment collected data samples from three of the participants in the first experiment on different days (The data samples were obtained in the looking-straight position only). Users were asked to sit in a chair and put on the device. After the calibration, the experimenter collected data samples of the eight facial expressions three times each (240 data samples: 10 samples per expression per time x 8 expressions x 3 times) on each day. Like Evaluation 1, the experimenter recorded manually while the user kept their maximum pose of each facial expression at regular 50-millisecond intervals. The experimenter conducted the procedure on three different days, and so the system acquired 720 data samples with facial expression labels from each user. The classifier (SVM in the same way as 4.2.1) used the data obtained on two of the three days as a training set (480 data samples) and the data samples from the remaining day as a test set (240 data samples). The results are shown in Table 4.5. The averaged accuracy for the three users was 78.1%. By making bigger the size of the training set, the repeatability can be ensured. It suggests the possibility of long-term usage. In the confusion matrix, the most dominant error was classifying 33.3% of the anger cases and 23.3% of fear cases as disgust.

	Day1	Day2	Day3	Average
User A	83.8%	91.7%	90.4%	88.6%
User E	72.9%	70.8%	69.2%	71.0%
User F	86.3%	69.6%	68.8%	74.9%

Table 4.5: Accuracy on different days

4.2.3 Evaluation 3: Usage during Walking

This experiment evaluated the effect of walking on the recognition of facial expressions because the activity may cause changes in the position of the eyewear, which influence sensor values slightly. The experimenter collected 480 data samples with facial expression labels from each user. Four users participated (two males and two females; two Japanese, one Chinese, and one German. Age range: 25-59). The experimenter collected 10 data samples for the eight facial expressions in the looking-straight position. The experimenter repeated the process three times (Dataset A: 240 data samples). After the data samples had been collected in the stable position, users were asked to walk along a corridor at a natural speed. The experimenter manually collected 10 data samples at regular 50-millisecond intervals for the eight facial expressions three times each while they were walking and holding their maximum pose for each facial expression (Dataset B: 240 data samples). Carrying a laptop, the experimenter walked along with each participant. The experimenter asked him/her to make all facial expressions one by one. Soon after recognizing his/her maximum pose of each facial expression, the experimenter recorded data samples with a laptop. The system used Dataset A as a training set and Dataset B as a test set. SVM in the same way as 4.2.1 was applied. The result was an average accuracy of 73.2%, which is slightly worse than the result found in the evaluation ref4-evaluation2. This is because walking caused the device to shift its position, which leads to the noise for the system.

4.2.4 Evaluation 4: Robustness to Positional Drift

To make the system robust, the noise by the positional drift of the eyewear should be considered as shown in the last evaluation. The fourth experiment evaluated the robustness to the slipping of the device down the nose. The experiment did not consider the slip to the side because it should not be a significant issue if the eyewear is appropriately fitted to the user. On the other hand, the downward slip of the glasses is a common occurrence.

In this evaluation, sensor value distribution is different from other experiments as the system collected data samples on various levels of the positional drift (Levels). Hence the dataset was normalized based on the average and standard deviation of training datasets.

First, to examine the relationship between the distance and the sensor values, the experimenter measured the distance d between the yellow mark and the base of the wearer's ear shown in Figure 4-8 in seven different positions. The distance corresponds to the degree of the positional drift. At the same time, the experimenter measured the sensor values of neutral expression in each position. Principal Component Analysis (PCA) was applied to reduce the 17-dimensional sensor data to the one-dimensional data. The result can be seen in Figure 4-9 (top: raw data, bottom: PCA value). It shows a linear relationship between the distance and the first principal compontent of the sensor values.

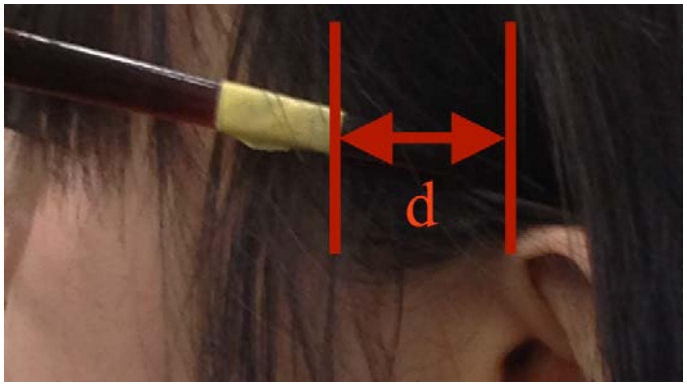


Figure 4-8: The distance between the end of the yellow part and the base of the ear were measured

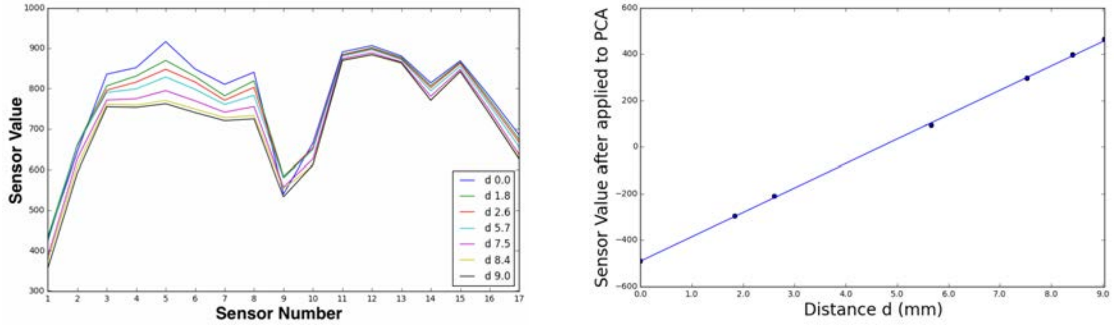


Figure 4-9: The distance and sensor values (neutral) change depending on the level of positional drift. left: raw data, right: PCA

Next, the evaluation was performed to see the effect of the slippage on the robustness of classifying eight facial expressions. The experimenter collected 960 data samples with facial expression labels in total (10 data samples per expression per time x 8 facial expressions x 3 trials x 4 positional drift levels) from five participants (four male and one female). They are all Japanese graduate students aged 22 - 27. The evaluation followed the procedure as Evaluation 1 except that the experimenter collected data samples at different levels instead of different poses. Level 1 is the base state where there is no slippage. The bigger the number of the Level is, the greater the degree of positional drift. The Figure 4-10 shows the snapshots of User B with different Levels.

The Figure 4-11 and Figure 4-12 show the averaged sensor values of all facial expressions at different Levels of Users A and B respectively. The sensor numbers correspond to the ones shown in Figure 3-5. Each color shows the corresponding Levels. As the degree of the slip is user-dependent, the Levels are defined relatively for each user. Sensor value distributions for Users C, D and E are shown in Figure 4-13 - Figure 4-16. These figures focus on particular expressions at different Levels (neutral, happy, angry, surprise) from different users. Two expressions (anger and surprise) from user E are included to show how the sensor value distributions differ by expressions.

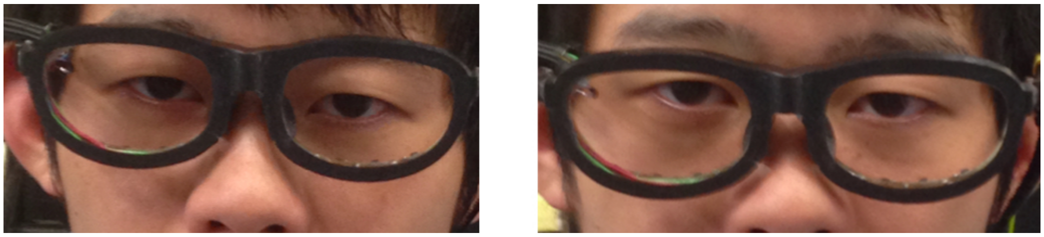


Figure 4-10: The positional drift level of the eyewear (Left: Level1, Right: Level4)

The evaluation was done to see the possibility of predicting facial expressions at one positional drift level using the data samples taken at another Level. The training set includes data samples from two trials at one certain Level (160 data samples: 10 data samples per expression per trial x 8 expressions x 2 trials) while the test set includes data samples from another trial (80 data samples: 10 data samples x 8 expressions x 1 trial) at one Level. The cross-validation method was applied: 3 test sets were evaluated from 3 trials at each Level. The classifier was trained with the training set of Level 1 and tested with the test set of Level 1-4 respectively. The process is repeated for the training sets from each Level and each user. The SVM applied was different from the one previously used (linear kernel, C = 500) because the SVM with rbf kernel did not perform well. The result was then averaged for all users. The matrix can be seen in Table 4.6. At all Levels, the accuracy of facial expression recognition was best when the training set and the test set on the same Level was used (78.0% - 87.8%). The further the distance between the Levels for the training set and the test set become, the worse the accuracy. The experiment concludes that it is hard for the current algorithms to predict facial expressions when the slip of the glasses happens without the dataset that includes the data samples at the level of the slip.

	Level 1	Level 2	level 3	level 4
Level 1	83.7%	39.6%	23.7%	21.8%
Level 2	46.5%	79.6%	45.9%	22.0%
Level 3	25.2%	42.9%	87.8%	38.7%
Level 4	23.0%	26.0%	44.6%	78.0%

Table 4.6: Accuracy of facial expression recognition using training set and test set on different Levels

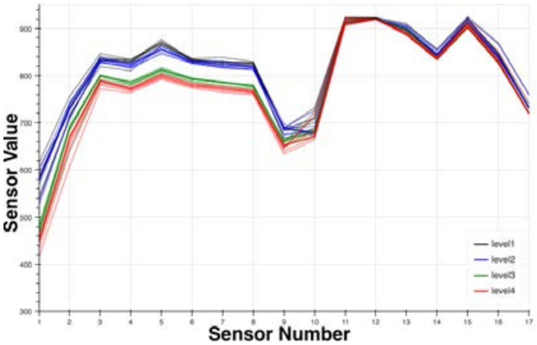


Figure 4-11: Facial expression distribution on each Level(User A)

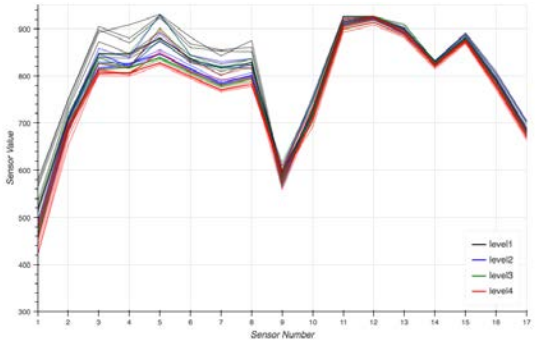


Figure 4-12: Facial expression distribution on each Level(User B)

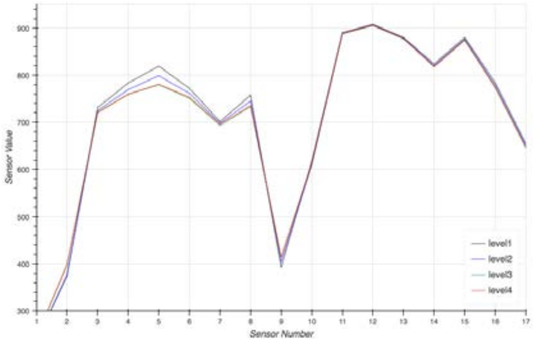


Figure 4-13: Sensor value distribution on each Level (User C, Neutral)

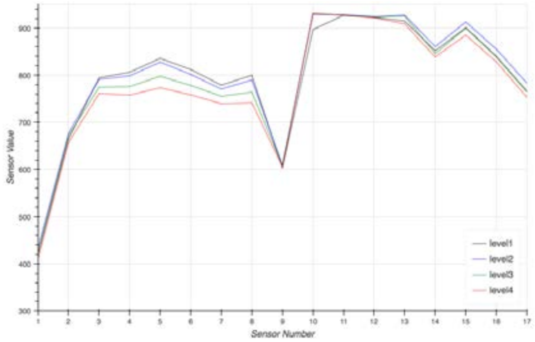


Figure 4-14: Sensor value distribution on each Level (User D, Happy)

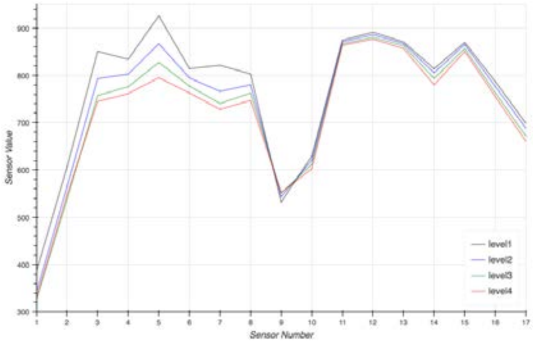


Figure 4-15: Sensor value distribution on each Level (User E, Angry)

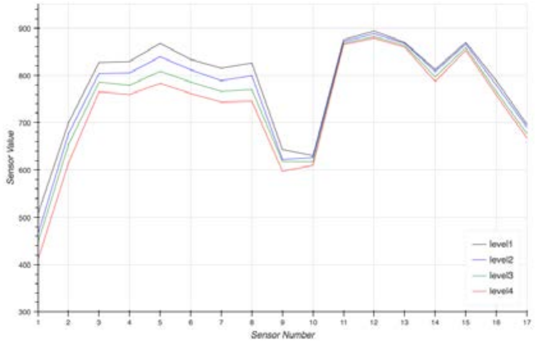


Figure 4-16: Sensor value distribution on each Level (User E, Surprise)

The other evaluation was done to show the accuracy of facial expression recognition with the data samples of all Levels. For each participant's dataset, the data samples were taken in two of the three trials (640 data samples: 10 data samples per expression x 8 facial expressions x 2 trials x 4 Levels) were merged as a training set and the data samples from the remaining trial was used (320 data samples) as a test set. Cross-validation method with SVM (linear kernel, C = 500) was applied to the datasets. The result is shown in Table 4.7. The average recognition rate was 86.8%. PCA was also applied to the training sets. Averaged accuracy of facial expression recognition slightly improved to 87.7%. The best results are shown with (User A:12, User B:13, User C:12, User D:15, User E:16) principal components respectively. Even when the positional drift of the glasses happens, the system can classify facial expressions with robustness by learning the data samples taken at different positional drift levels.

	User a	User b	User c	User d	User e	Average
Result with 17 Sensors	89.3%	85.5%	84.2%	82.1%	92.8%	86.8%
Best result(PCA)	91.0%	86.7%	82.3%	83.5%	94.9%	87.7%

Table 4.7: Accuracy of facial expression recognition using the datasets that includes all Levels

4.2.5 Demonstration at SIGGRAPH Emerging Technologies 2015

The eyewear device has been demonstrated at SIGGRAPH Emerging Technologies 2015. During the demonstration, There were more than 200 users from various international backgrounds, and they tried on the device. As the demonstration proceeded, the observations have emerged that the size and shape of the eyewear had to be adjusted to each user for accurate recognition. Three major issues were present: (1) The device sometimes slipped out of place when some users changed facial expressions. The slippage changed the accuracy of the recognition because the device relies on the proximity sensing between the front frame of the eyewear and the skin surface on the face. (2) For those who had high nose bridges, some of the sensors positioned between users' eyebrows saturated and did not work well. This problem could occur even with the neutral expression. 3) The sensors placed on the top can measure the distance to the eyelid or the eyebrows depending on the shape of the users' face or the position of the eyewear. Therefore, the sensor data can be different for each user, requiring individual training and calibration. Due to these reasons, the eyewear should be customized to each user. There was also the observation that the eyewear seemed to work better for Caucasians compared to Asians as Caucasians tend to be more expressive.

4.3 Discussion

This chapter presented a smart eyewear prototype that can classify the wearer's facial expressions with 92.8% accuracy using user-dependent training in the experimental setting. This result is promising, showing the potential of the proposed approach. Though many of the users remarked that they did not see clear differences between surprise and fear, fear and disgust, and disgust and anger when they looked at the pictures presented for instruction, it was still possible to classify those expressions. The subtlety of differences between some expressions such as anger and disgust may lead to miscategorization. However, this owes more to the ambiguous nature of human facial expressions than to the design of the recognition system. This issue may be one of the reasons that the system require the individual training as it means they showed a slightly different expression for those labels. It is worth noting that there is some empirical evidence that challenges the universality of basic emotions (See [81] for a discussion on cross-cultural recognition of facial expressions).

The analysis using the user dependency matrix showed the need for user-dependent training. It does not pose a significant problem with the device since eyewear is a personalized item as pointed out by Scheirer et al. [84]. The device can be designed as personalized eyewear, intended to be used by a single user. However, having external datasets may help reduce the cost of the learning phase and improve the accuracy of recognition. Ideally, such external datasets should consist of data samples from a large number of users rather than a large volume of samples from a few users as suggested by Girard et al. [24].

The evaluation of the system found that accuracy was reduced in the case of long-term usage and walking. This reduction is due to the physical condition difference of the device such as the effects of ambient light and positional drift. Another factor for the reduction in accuracy was the difficulty of reproducing the same facial expressions across different times and conditions. It was challenging for the users to try to repeat the exact same facial expression. They made slightly different facial expressions with varying intensities. Accommodating the changes is an inherent problem in developing a facial recognition system. However, the findings suggest that it may be made less of a problem by increasing the size of training sets with more trials. In the experiment that measured the effects of positional drift, the accuracy of 87.7% achieved by learning the data samples at different levels of positional drift. Requiring the users to conduct extensive and repetitive training would probably be unrealistic for real-life usage, but creating a new system that utilizes external training datasets may mitigate the issue.

The chapter focused on classifying the wearer's facial expressions by basic emotion categories using the eyewear device. However, our facial expressions may not represent our inner emotions. Our face may reflect mental effort or convey a communication signal. Information other than basic emotions was not considered in this chapter.

4.4 Limitations

In this study, it was assumed that skin deformations around the eyes indicate changes in facial expression. However, other behaviors such as yawning, rubbing one's eyes, and resting one's cheek in one's hand can also cause skin deformation. These normal behaviors can affect the system of facial expression recognition. Moreover, readings from the photo reflective sensors are affected by the condition of the facial skin. Factors such as tanning, sweat, makeup, and facial swelling may require users to do calibration and training again.

The eyewear device only collects data from the sensors around the eyes. Although movements of the mouth are partly detectable, there are several mouth movements the system cannot detect because mouth movements and cheek deformations do not have a one-to-one correspondence. For instance, the action of opening a mouth lowers the cheeks slightly, which makes the distance between the eyewear frame and the skin surface greater. Therefore, when the sensor values on the lower part changed, the system cannot determine either movement of mouth or cheek deformation cause the changes.

During the experiments, the participants made facial expressions intentionally. The posed and natural expressions are similar to a certain extent, but there is no doubt some differences exist between them. Such differences can have an adverse influence on the accuracy of natural facial expression recognition, mainly because our natural facial expressions are more subtle than posed ones.

4.5 Summary and Future Work

This chapter presented the evaluation of the eyewear that classifies a wearer's facial expressions. The conducted series of evaluations focused on classifying eight universal facial expressions. The experimental results showed recognition rates of 92.8% for one-time use regardless of facial direction or removal/remount of the device; 78.1 % for repeatability and multiple-day usage after a training process; and 87.7% if the positional drift of the glasses was taken into account. The robustness in daily scenes can be achieved by learning more data. The system still has room for improvement regarding calibration and accuracy, yet it is a significant step in quantifying the flow of facial expressions in daily life.

The following issues are the future work.

Firstly, it would be better to design a natural learning process to capture natural facial expressions. To this end, the automation of the learning process using the emotion-evoking stimuli can be one option.

Secondly, designing an optical filter to reduce the influence of ambient light, especially sunlight can improve the recognition accuracy. The prototype used photo-reflective sensors that function on the basis of IR reflection. Because sunlight contains enormous amounts of IR light, sensor data saturates when the sensors are exposed to direct sunlight. With the current system, the sensors are not covered by anything and are easily influenced by an intense ambient light even though the system applied light modulation to the LED of the photo reflective sensors. The filter can be designed by considering the directional light.

Thirdly, the calibration process needs to be improved. For robust recognition, the prototype requires each user to calibrate under various situations. Generating an additional dataset based on already trained data can reduce this process. Transferring learning method should be explored so that the system can make use of other users' datasets to calibrate another's effectively.

Chapter 5

Recognition of Spontaneous Expressions in Daily Conversation

In the previous chapter, the focus was on classifying basic emotion categories in the laboratory setting. All of the recorded expressions were posed. However, in real life, people do not make such exaggerated expressions frequently. This chapter introduces the three attempts to capture spontaneous expressions. The first section introduces the prediction of the daily activities recording from pre-training basic emotion categories. The second section describes the summarization and mapping of spontaneous expressions in daily conversations in an unsupervised manner. The last section explains the spontaneous expression detection while reading. It shows the device can consider the eye movement, which leads to the potential of quantifying reading activity that is related to the cognitive process of the human brain. The First section uses the first prototype, and the other sections use the second generation of the prototype. This chapter is also based on the work [58, 59, 60].

5.1 Initial Field Trials

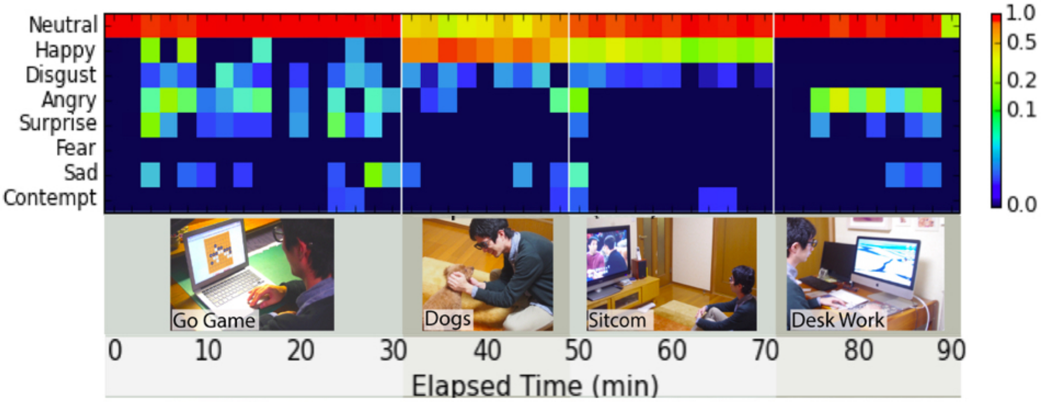


Figure 5-1: Distribution of predicted result of facial expressions based on recorded sensor values for a long time

This section aims at recognizing spontaneous expressions based on pre-trained posed labels. Although there is a discussion that the posed expression and spontaneous one was different, it does not limit to capture spontaneous expressions with pre-trained posed labels.

Initial field trials were conducted for daily life by recording facial expressions of a user in daily living or home scenario. The recording was a 90 minutes time series consisting of four activities: (1) playing Go game with a computer for 32 minutes (a board game involving two players originated in China) (2)playing with a dog for 16 minutes (3) watching an episode of "Friends" for 22 minutes and (4) programming on a computer for 20 minutes. Additionally, playing a shooting game with friends (won and lost: 10 minutes) and watching an episode of a crime drama "Crime Scene Investigation (CSI)" for 45 minutes were recorded on the following day. The aim was to compare (1) the activities that include social interactions (human-human and human-animal) as well as individual activities, and (2) the same activities in a different context (e.g., winning vs. losing games, watching comedy vs. drama series). Figure 5-1 shows the frequency distribution of recognized facial expressions during the recording period. The figure presents a normalized distribution for every 2 minutes with a logarithm contrast enhancement. The enhancement makes less frequent expressions easy to see since sometimes neutral expression was mostly shown during the activities.

The distribution of facial expressions varied depending on the activities. For instance, happy expressions were mostly observed while interacting with the dog. The sitcom also produced happy expressions between intervals of neutral ones. While playing the Go Game, negative expressions sometimes occurred, but they were eventually replaced with a happy expression reflecting the progression of the game (from facing challenges to winning the game). During individual activities, the user tended to show fewer facial expressions. For example, while programming induced some angry expressions, the dominant expression was neutral. On the other hand, the user showed more various facial expressions (other than neutral) while interacting with the dog. The user also showed various facial expressions while playing the shooting game with a friend. The results are shown in Figure 5-2. During the game, the user displayed more various facial expressions such as happy and angry while playing with another person, compared to when the user played alone with the computer (the Go game). It suggests that social interactions induce more different facial expressions. Although this is only a preliminary field trial, it is tempting to speculate that facial expressions can be used as an indication of communication between people. This result is consistent with the work by [33].

The distribution of facial expressions may also be influenced by the nature of the interactions as seen in the case of gaming in the trial (see Figure 5-3). When the user won, he showed more happy expressions, especially in the latter half of playing. When the user lost, negative expressions were displayed, the dominant expression being anger. However, the user frequently displayed happy expressions during the gameplay even when he eventually lost.

While watching the episode of the sitcom "Friends," the user sometimes showed happy expressions. When the user watched an episode of the drama "CSI" that included some graphic scenes such as murder, dissection of the human body and bleeding, disgust and surprise were detected. While this is an unsurprising result, it is once again tempting to speculate whether analyzing the distribution of facial expressions may be able to provide some feedback on the user experience.

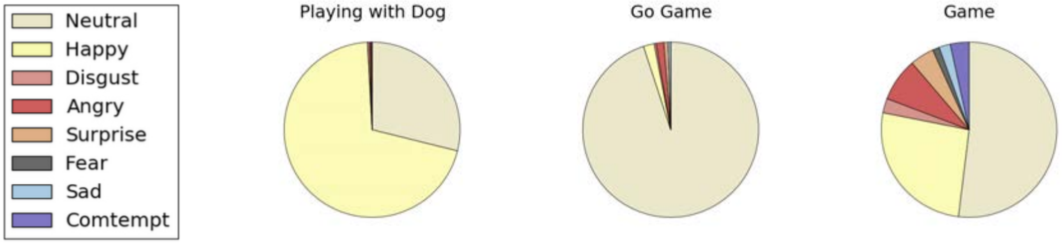


Figure 5-2: Facial epression ratio during the activities

5.1.1 Discussion

Although the ground truth was not recorded, the device could capture the characteristics of each activity. However, some of the results may have been better understood if combined with non-emotional signals or physiological signals. For example, in the field trials, "angry" facial expressions were recognized during computer programming or Go game. It is not difficult to imagine that this was more a reflection of mental effort, confusion or frustration rather than an expression of anger. However, considering the common aspect of anger and concentration which try to avoid the other's inter-

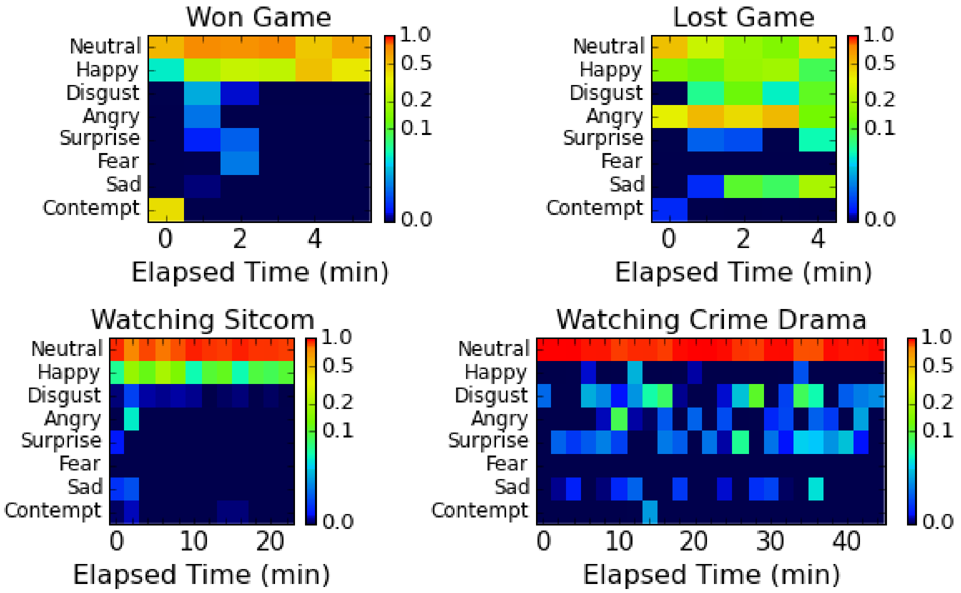


Figure 5-3: Comparison of facial expression distribution between the activities

vention, it is obvious that the basic emotion labels are not the only way to interrupt the facial expressions. Du et al. suggested the existence of 21 expressions [14]. The information provided by complex facial expressions would be useful for understanding the user experience in depth. This evaluation is a first step towards classifying more complicated facial expressions in daily life.

In the field trial, the device classified the expressions that were high in intensity, which suggests that the system trained with the posed expressions were unable to pick up less intense expressions. It means the system could only provide an approximate picture of the pattern of facial expressions in a daily life setting.

5.2 Spontaneous Expressions in Daily Conversations

In the previous section, the classification was based on basic emotion labels. This section shows the way to capture facial expressions differently based on sensor data. The section focuses on the spontaneous expressions in daily conversations where people show diverse expressions as the previous section suggested.

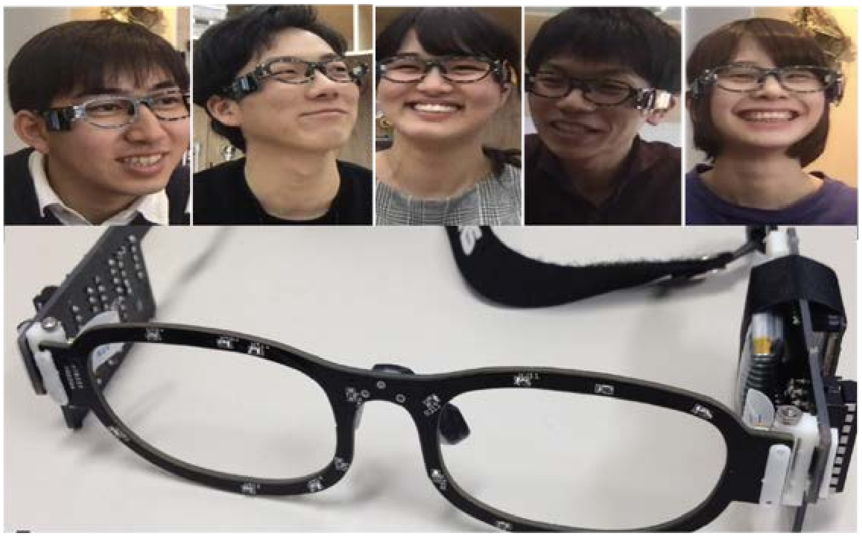


Figure 5-4: People show different kinds of spontaneous expressions in daily conversations. The method maps the similar expressions using the wearable sensors in an unsupervised manner.

5.2.1 Introduction

Humans are inherently social and exchange information with others through multiple channels such as languages, body gestures, audio tones, and facial expressions [22]. According to Frith et al., facial expressions play a dominant role in our daily interaction compared to other nonverbal clues [21,65]. We can often recognize the intentions and emotions unconsciously from facial expressions. Moreover, facial expressions determine our subjective impression of a person. We tend to show facial expressions deliberately in situations in face-to-face communication [33]. Especially we smile more in conversations with friends [36]. The smiles vary in different contexts. For example, we show social smiles for efficient communication while we show smiles after feeling a positive emotion. To analyze the different expressions people convey, this section measures facial expressions during daily conversations.

The most of the facial expression classification methods use the labels of categorical frameworks beforehand. While it allows objective interpretation by labels, it is necessary to record facial expressions with associated labels such as smile, anger and so on. Also, it is difficult to take into account the facial expressions that are not included in the labels. They may reflect the vital information to understand personal differences in facial expressions that are stable over time [10]. For example, it may be different between a smile made by those who often laugh and the smile of another who rarely laugh. One approach to tackle this issue is unsupervised learning that can identify the structure of the data. It makes possible to classify according to individual expression patterns. It is useful for human behavior analysis. For the camera-based approach, there are a few works that applied an unsupervised classification such as [109]. Unsupervised classification of facial expressions using a sensor-based method has not been explored yet.

The purpose of this section is to map and classify natural facial expressions in daily communication, especially that are subtle and typical of each person using a wearable sensor and unsupervised learning method(Figure 5-4). By doing so, it is possible to see the potentially detectable expressions with the device. The five users' facial expressions were recorded for five to ten minutes in daily conversations. In such an unscripted communication, the users showed their spontaneous and subtle expressions. For the recording, a camera captured the facial images, and the device captured the sensor data. For the analysis, the Self-Organizing Map(SOM) [48] was used to visualize and summarize facial expressions in an unsupervised manner.

It is necessary for the smooth and natural interaction between human and computer, and a talking agent to understand the personal difference of their facial expressions. Besides, if the system can correctly detect the communicative signal, the social assistant robot can help with appropriate timing. It is also useful for human behavior analysis. The method can be used to improve communication skills by knowing the personal habit of making facial expressions. For example, the user can get feedback on how to make facial expressions in job interviews or business presentations where the impressions are essential to be successful.

The main contributions of the section are the following:

1. Data collection of facial expressions in daily conversations using the eyewear device and a camera. The data was used to visualize and map the facial expressions of the participants using an unsupervised learning method.

2. Description of summarizing natural facial expressions in each cluster of the map. The frequency information could show characteristic facial expressions for each person.

3. Creation of the map by learning the sensor dataset of multiple users. The method could construct common or similar expression classes among users considering the distribution of facial expressions in daily conversations. It could also capture diverse expressions of the users.

5.2.2 Unsupervised Mapping Using the Eyewear Device

This subsection explains how to process the data from the device and how to make sensor data-driven maps using an unsupervised learning method. For the system, the sensors located at both ends of the upper part were used for measuring the ambient light. The information of these two sensors reflects on the movement of head direction since it does not get influenced by the change of facial expressions.

There are three main advantages to using the eyewear device in combination with a camera to measure spontaneous facial expressions in daily conversations. First, it is wearable. It can measure facial behaviors regardless of the optical occlusion by the hand gesture, directional change of the faces, and the user's body movement. Second, the device is eyewear. It is comfortable to wear. Additionally, it does not disturb the camera-based facial expression recognition as algorithms are already trained to deal with glasses. Third, the device uses optical sensors. The sensors can measure subtle differences in distances. The sensors have the potential to measure even minute changes in facial expressions.

5.2.3 Data Processing Pipeline

The device acquired the sensor data that consists of a 16-dimensional 10 bits values(see Figure 3-7 for the layout of the sensors). The built-in camera also recorded a sequence of pictures. Those two data are synchronized using the attached timestamps. The sampling frequency is about 30 Hz for both. The algorithm applied a five sequences simple moving average filter to each dimension of sensor data in order to reduce the noise. Then outliers were eliminated using the following formula from the acquired dataset. If the sensor data included an outlier in any dimension, the data were excluded. $outliers = (data - mean) > 4xstd$ After the exclusion, the dataset was normalized so that the time-series data points in each sensor dimension have zero mean and unit variance.

Creating facial expression maps

SOM [48] was used for unsupervised classification of the sensor data. SOM is an artificial neural network of unsupervised learning that does not require labels beforehand. It can summarize the non-linear data by preserving the topological properties of the inputs. It is one of the most standard methods that can visualize high-dimensional data into 2D map. The proposed method made use of MiniSom [101]. It mapped the sensor data into a 2-dimensional space with 7 X 7 neurons. For the hyperparameters, sigma was set to determine the range to update the weights of the neurons to 1.0, and a learning rate was set to 0.4. The initial value of the weight for each neuron was set randomly from one of the sensor data in the dataset. SOM was trained with all data in the dataset. The order of the data for the training was randomized, and training had 1500 iterations.

There were many data in one cluster. To determine the facial expression regarded as representative of the cluster (representative facial expression), the following three approaches were taken.

Average Expression

The pictures were processed in the clusters with the following steps. The approach used the face morpher library[1] and dlib library [46]. First, the method found the face areas in the photos of one cluster. This localization used convolutional neural networks based face detector in the dlib library. Then, it extracted the 68 facial

[1] https://github.com/alyssaq/face_morpher

landmarks. It aligned and warped the faces in the cluster using the points. After that, it averaged facial expressions using the extracted feature points.

Facial Landmarks Median

This approach used the facial landmarks extracted in Average Expression method. After normalizing the landmarks in the range of 0-1, it calculated the median values of all the landmark points in each cluster. As a representative facial expression, it picked out the picture that corresponds to the landmarks that have the closest Euclidean distance to the median values of the cluster.

Sensor Data Median

The method chooses the sensor data that is closest to the medians of sensor data in one cluster. Then it regards the corresponding picture as the representative expression of that cluster.

5.2.4 Case Studies of Daily Conversations

The case studies were undertaken to see the potential of the approach to classifying various facial expressions. The experimenter recorded five users' facial behaviors in daily conversations with the eyewear device and a built-in camera on MacBook Pro 2016. They sat down at the one place and talked about what random topics such as the holiday experiences, their favorite sites and so on. They talked with two or three friends including the observer. The experimenter did not limit the topics and any head movement of the user. They sometimes spoke, sometimes listened. The size of a recorded picture is 360 X 640 pixels. All of the recordings have been done indoors at the similar location with a similar ambient light condition. The individual user case studies explore each a specific theme: the User 1 case focuses on a reliable, continuous usage scenario, the User 2 case focuses on variation in facial expressions (especially smile), the User 3 on how interruptible the expressions are.

User 1: Reliability

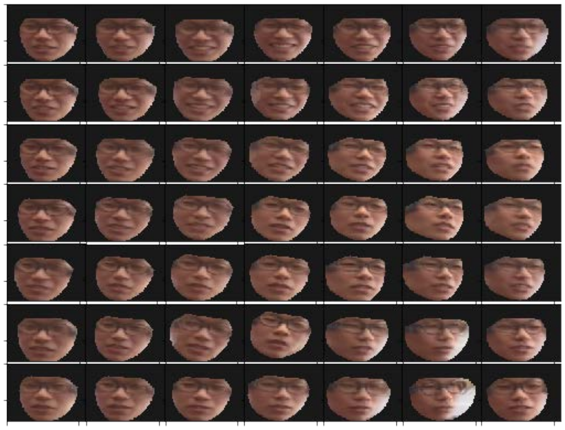


Figure 5-5: Averaged Expression Map

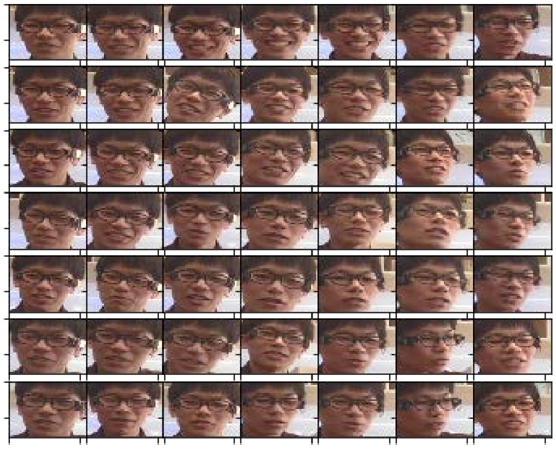


Figure 5-6: Facial Landmarks Median Map

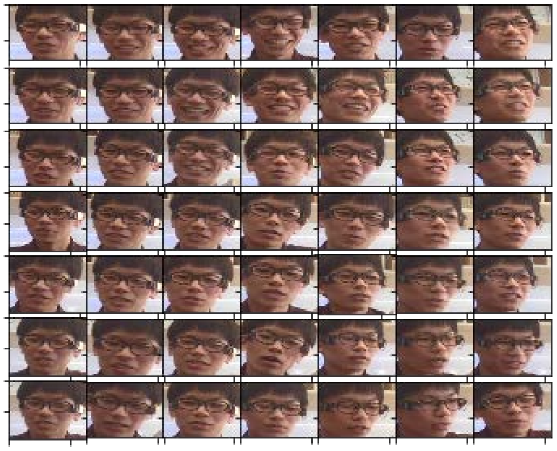


Figure 5-7: Sensor Data Median Map

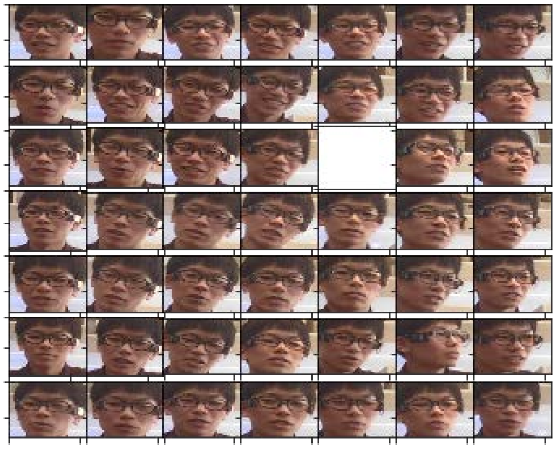


Figure 5-8: Mapped with another dataset

The experimenter recorded the 5 minutes conversation of the first user two times. Between the recordings, the user had a short break for about 30 seconds. For each recording, the experimenter asked the user to start with a neutral face. The experimenter trained the SOM using a dataset from the first conversation.

Figures 5-5,5-6,5-7 show the visualized maps of the first user's dataset of the former conversation using the above methods (Average Face, Landmark Median, Sensor Median). All of three maps show similar expressions in nearby areas. SOM was able to map similar facial expressions to the same or near clusters using the sensor data from the eyewear device. In other words, the sensor data has correspondence with various facial behavior. It summarized the main expressions of the first user. The first method (Average face) could show the impression of facial expressions in each cluster. However, if the landmarks of faces were not accurate, the averaged expression was easily blurred, and difficult to interrupt the meaning of faces. The second method made use of the landmarks. It enabled to pick out the picture that corresponded to the median data of the landmarks. These two methods take time because they require processing all of the pictures to detect faces. The last method picked out the representative facial expression corresponding to the median of the sensor data in the cluster. The result is competitive with the other methods. This method is faster since there is no need to apply face recognition to all the pictures. Also, the method can consider all of the data even when the frontal face is not available in the pictures.

For the first user, there were two main expression clusters. The first cluster is a laugh that appeared on the top three rows. The other cluster is a neutral face. Also, The clusters were made by left-right face direction. It is because the ambient light came from the left direction of the user. Also, the clusters correspond to the up-down face direction. Since people look at the eyes when they have confidence and look down when they have vague feelings in mind, the face direction reflects on the communicational intention. Figure 5-8 shows the result of mapping the dataset of the second conversation to the trained SOM. Each dataset was normalized to 0 mean and unit variance previously as the facial expression appeared on the two conversations were not so different. The blank cluster means no sensor data was assigned. From the map, almost every cluster has a similar representative expression, yet there were different expressions in the same place (for example, the first row and second column). This result suggests the potential to use the method for a semi-supervised approach. For example, measuring the expressions first with the camera and the device, then the user measure using only the eyewear device. If the clusters are made in an unsupervised manner, the labeling to all the sensor data is not necessary as long as some of the data in the cluster have its facial expression labels. It is useful for long-time recording to see how the frequency of the user's facial expression changed based on the user's facial expressions that had already appeared.

User2: SOM Resolution

The second user made a conversation with her friends and the experimenter with everybody sitting. The experimenter recorded the 5 minutes conversation two times. The experimenter has tried the Average Face method. However, the landmarks of the facial features of the user 2 did not achieve good results. It is caused by the fact that the face is not always facing front and her forelock hides the eyebrow. Therefore, there is the result of Sensor Median map only(Figure 5-9).

The characteristics of user 2 are that several kinds of smiles have the different clusters. For example, the right middle areas show the smile of enjoyment where the corner of the eyes wrinkle while the bottom right areas show social smile that the eyes are neutral [19]. Also, the map showed the intensity of the smiles as it can show the transition.

By overlapping the frequency information as the transparency, it is possible to know the typical expression in the recording(Figure 5-10). As for the user 2, the most frequent expression is neutral or the face listening to others.

Moreover, by increasing the resolution of the SOM, it can show more diverse expressions (Figure 5-11). It is also useful to know how well the SOM can visualize the clusters with the sensor data. The similar expressions have its island. Since there is a significant amount of information, the high-resolution map is not appropriate to the summarization of the conversation. On the other hand, it is informative to know the user's trait about what kind of expressions are shown because the map showed the various expressions the user made in order.

User 3: Interruptibility

The experimenter recorded the user 3's facial behavior for ten minutes. Besides, the experimenter recorded the posed facial expressions that relate to the eight basic emotional states (neutral face, smile, disgust, anger, surprise, fear, sadness, contempt) defined by Ekman [17](Figure 5-12). The user 3 made facial expressions for 3 seconds each. The experimenter repeated five times. Figure 5-13 shows the Sensor Median

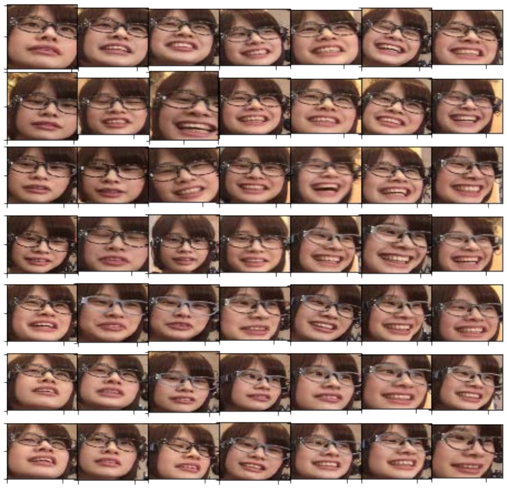


Figure 5-9: The Sensor Median map of the user 2

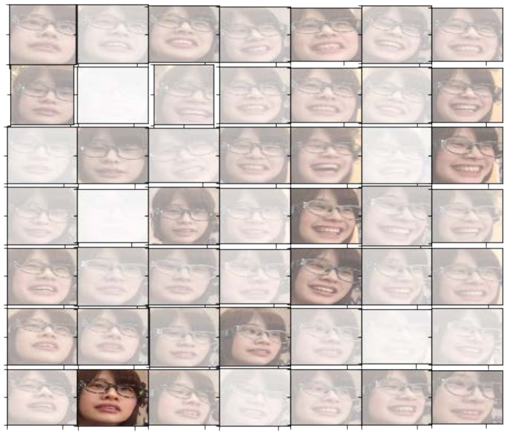


Figure 5-10: Overlapping the frequency information to the map of the user 2

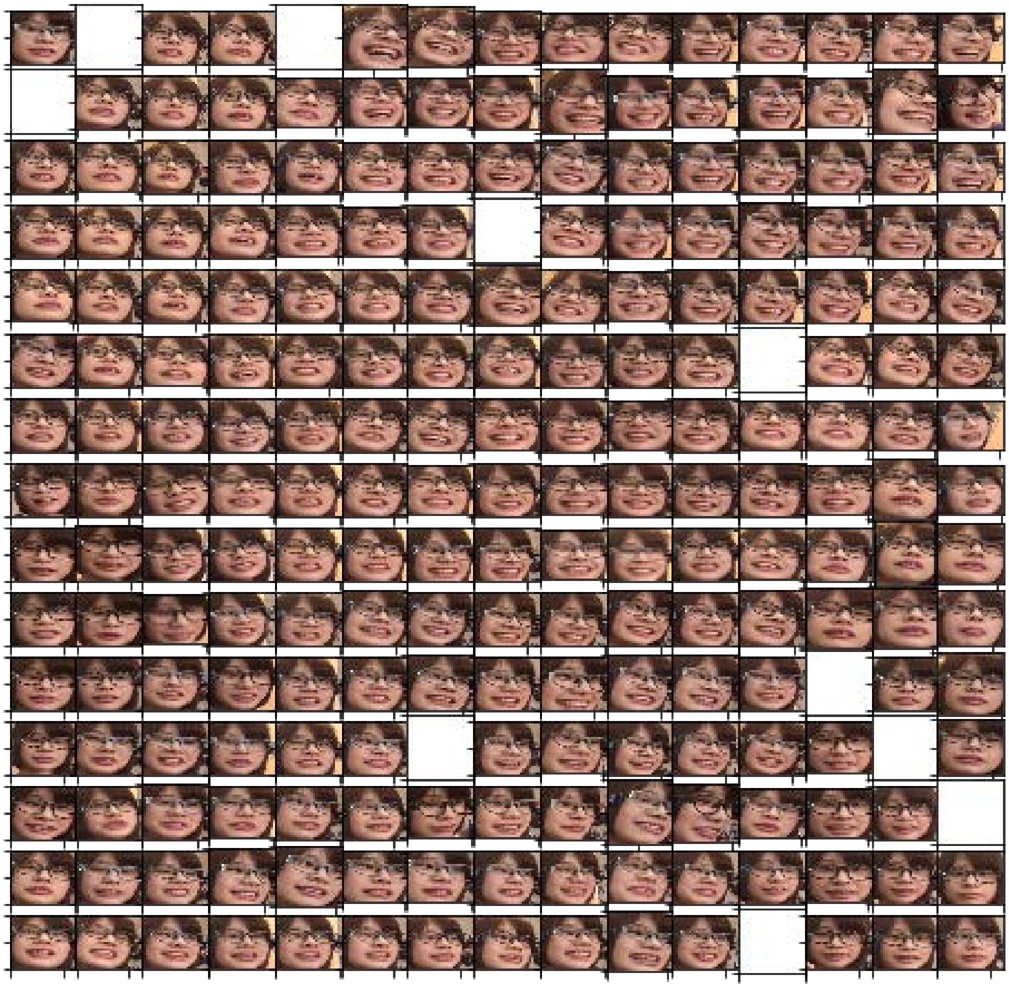


Figure 5-11: The high-resolution map of the user 2

map of the user 3. As similar to the user 2, the clusters of smiles were mapped according to the intensity of the smiles. The enjoyment of the smile appeared on the lower left areas. The smaller smile appeared on the middle right and bottom middle areas.

The data with basic emotion labels were normalized using the mean and standard deviation of the conversation dataset. Then, The data related to basic emotions were mapped to the trained SOM (Figure 5-14). It helps to understand the map. For example, the posed smiles were mapped to the clusters of the smaller smiles. It means that the smaller smiles are similar to the posed one that did not appear as an enjoyment. Another example is that the surprise arose on the first row, the fifth column. It suggests that the expressions on that cluster have the faces with the raised eyebrow. On the other hand, the other emotional states were mapped to the same area as the neutral one. It is because the expressions were very different from the ones shown in the conversation.

The Map Trained with Multiple Users

The map was trained using all five participants' datasets trimmed to five minutes conversation per each. Each dataset was normalized to mean 0 and unit variance respectively and was merged into one dataset. Figure 5-15 shows the Sensor Median expression of the dominant users in each cluster. The map shows the clusters only when the number of one of the user's data occupies the half in one cluster. From the left bottom areas of the map, two users' smiles are smaller than the others' smiles of adjacent areas. Although they show different smiles, their smiles are similar concerning the distribution of facial expressions in the conversations. In other words, the method can compare how different smiles the users make in the same situation.

Figure 5-16 shows the result of applying each user's dataset. The alpha value for each cluster is the number of each person's data divided by the sum of all users' data in the cluster. Each user's data was distributed into various clusters. Regardless of the users, the data of the neutral faces and the smiles are mapped into the same or close clusters respectively. It means the map could make the clusters across the users.

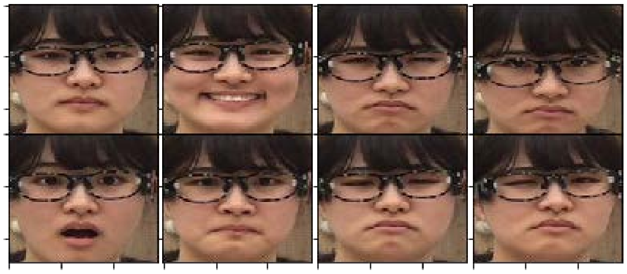


Figure 5-12: The posed basic emotions made by the user 3

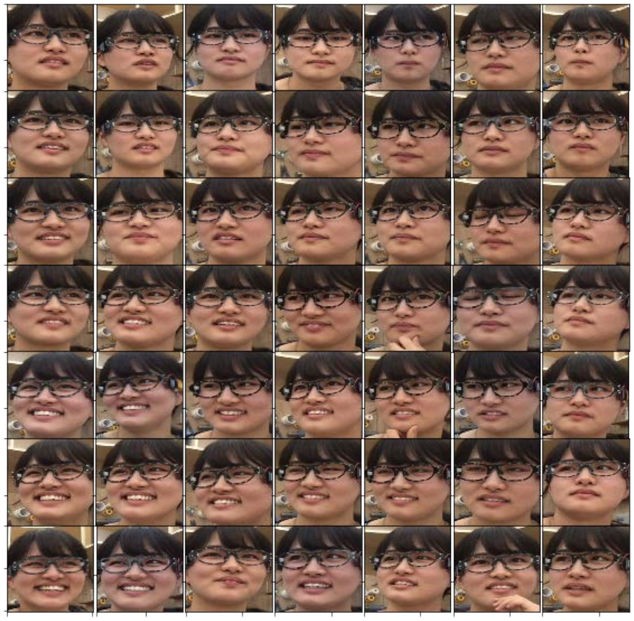


Figure 5-13: The Sensor Median map of the user 3

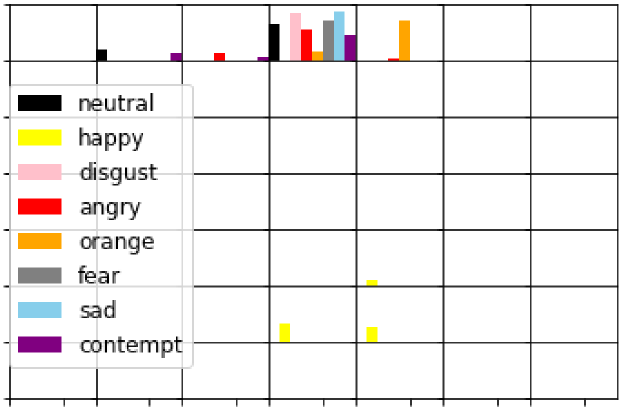


Figure 5-14: The mapped amount of basic emotions data to the clusters in Figure 5-13

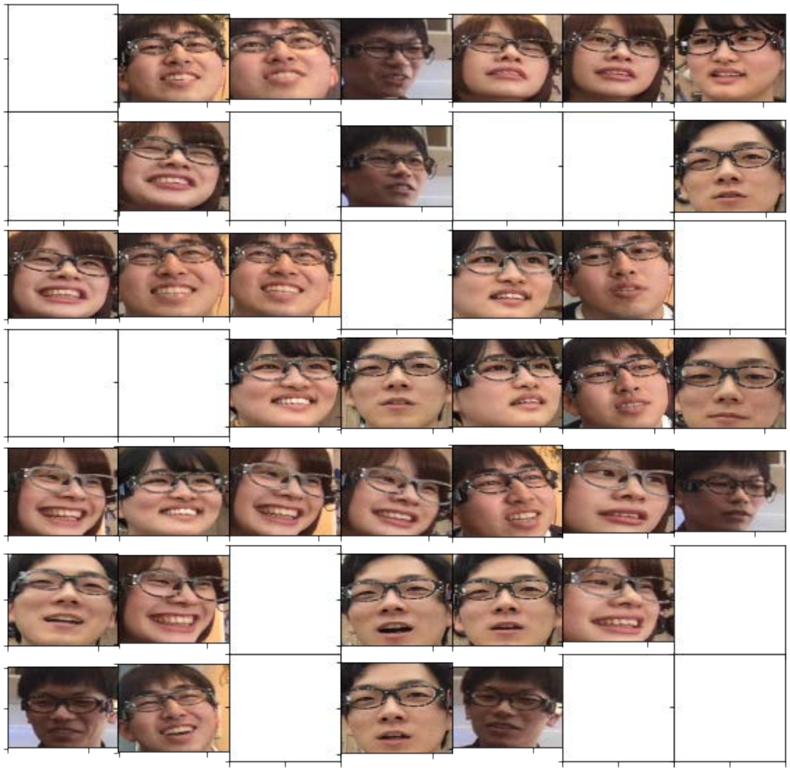


Figure 5-15: The map trained with multiple users' data

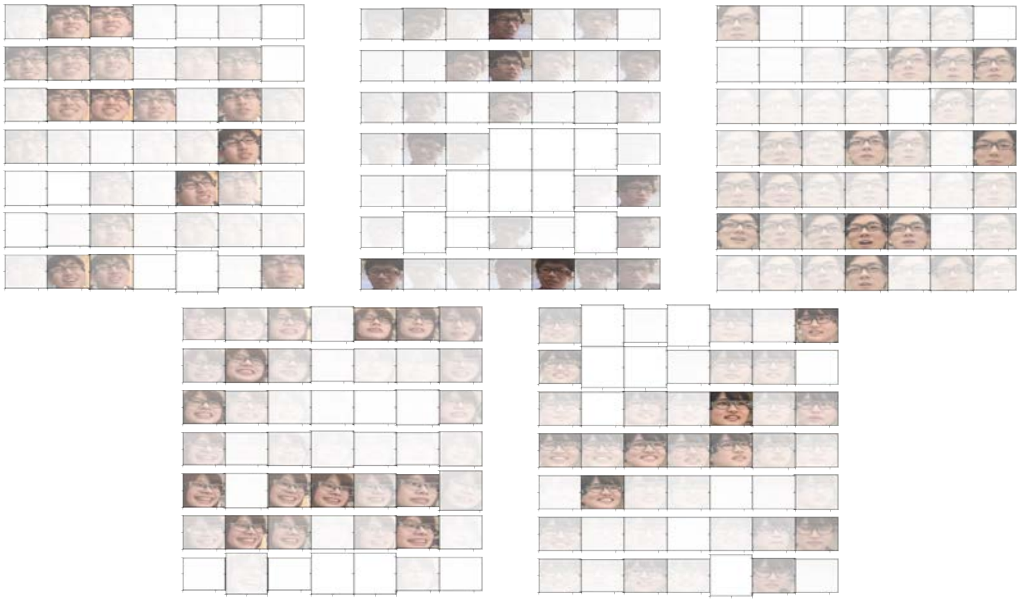


Figure 5-16: The mapping result of each user's data to the SOM trained with multiple users data with frequency information overlapping

5.2.5 Diverse Expressions of the Users

Up to this point, the maps showed the main expressions frequently appeared on the dataset because all of the examples in the dataset have the same weights and the outliers were eliminated. These maps can see what expressions dominate on the map. On the other hand, there are the expressions that have characteristic of the user expressions even if they appear less frequently. To see the diverse facial expressions of the users, a map was trained with "diverse" dataset where the sensor data are distributed far from the mean. It used the data that satisfies the following formula in any sensor dimension.

$diverse = (data - mean) > 3xstd$

Figure 5-17 shows the result of the user 2. The map could show more diverse expressions than the previous map while the similar expressions are still in the close cluster. Figure 5-18 showed the map with all user's "diverse" dataset. The map picked out the expressions in the same way as All Data section. As the expression shown have the characteristic sensor data, what it visualized is the characteristic expression of each user in the datasets at one map.

5.2.6 Discussion and Future Work

In the case studies, the map showed the facial images based on the sensor data structure using SOM. The device can map based on the face direction, the intensity of smiles, various types of smiles. The method can visualize the main expressions the users showed. Besides, the method can map the expressions across the users.

The method can help understand how the users make facial expressions. It is a qualitative analysis by summarizing and visualizing the data. It needs to combine a quantitative analysis for the better understanding of facial expressions in the conversations. To this end, the annotation by human coders to the video is required in the future work.

Only five students participated in the study. Further work includes examining how the method can generalize with diverse nationalities, gender, and ages in different

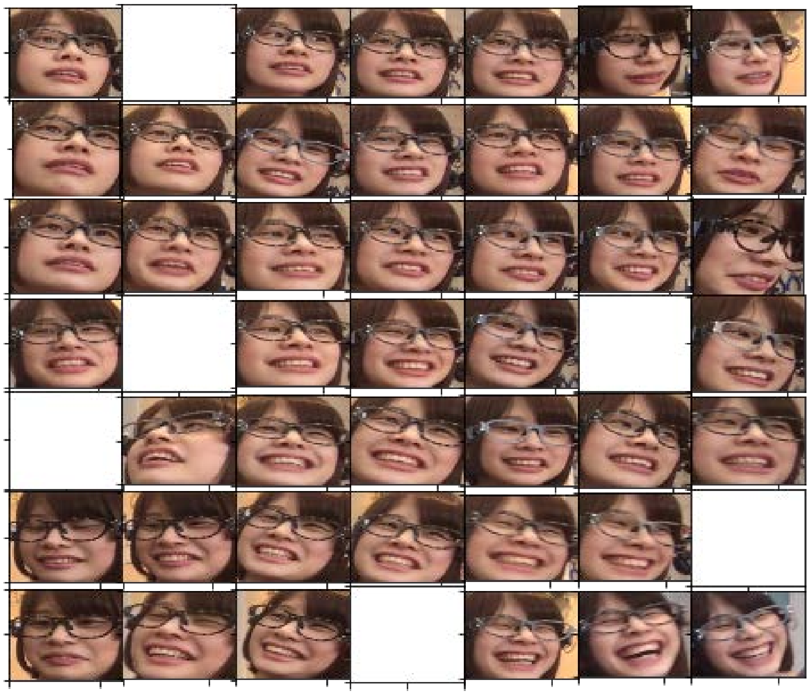


Figure 5-17: The Sensor Median map showing the diverse expression of the user 2

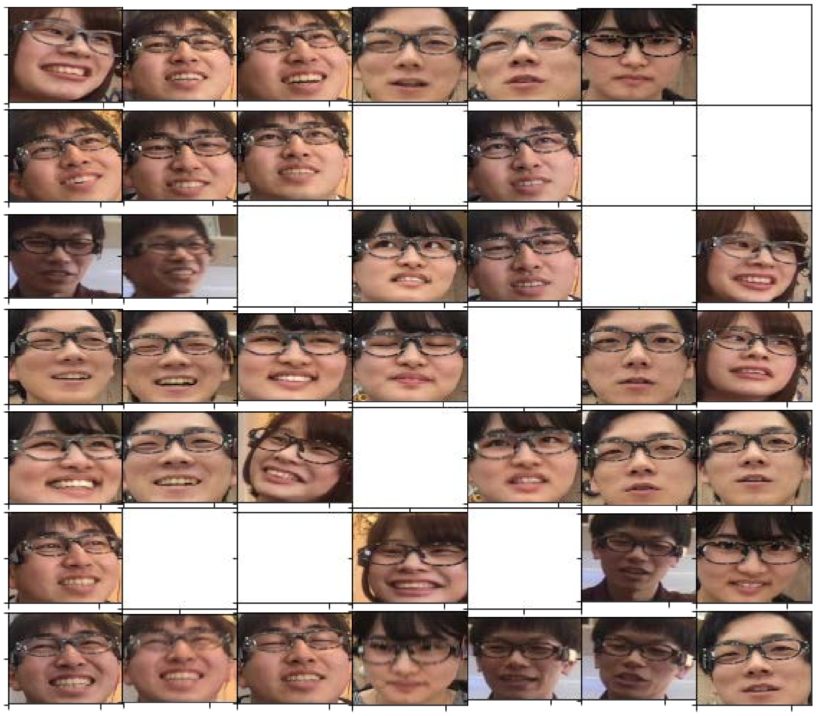


Figure 5-18: The Sensor Median map showing the diverse expression of multiple users

ambient light conditions.

If the facial expressions in the test dataset are not similar to the ones in the training dataset, they cannot be classified correctly. Also, the method needs to make correspondence between training data and test data by calibration if those come from the different distributions.

The SOM used the static sensor data. The data corresponds to the geometric change. Time-series information was not considered although the dynamics of facial expressions have abundant information about how the facial expressions of the user change. The future work considers temporal features of facial expressions by using such as recurrent neural networks.

The conversation was analyzed from one-side. However, the interaction is always at least two-sided, and people affect each other's expressions and (re-)actions. It would be better to analyze conversations both-ways to understand social interactions in greater detail.

5.3 Eye Movements and Reading Detection

Up to this point, the primary focus was facial gestures. This section focuses on eye movements that are also important aspects of facial expressions. Some researchers show that eye movements and blinks reveal about the mind of people [27,88]. Therefore, facial expression and eye movements are both critical to understanding people's inner states and behavior.

Since the system can detect the skin deformation around the eyes, the system can be used to measure the eye movements. This section investigates how accurate the eye movement can be measured and potential of reading quantification. Mainly, the experiments attempted to measure line breaks by the eyewear device. With the information, the system can estimate how many words the user read [50]. Also, facial responses caused by reading texts were analyzed.

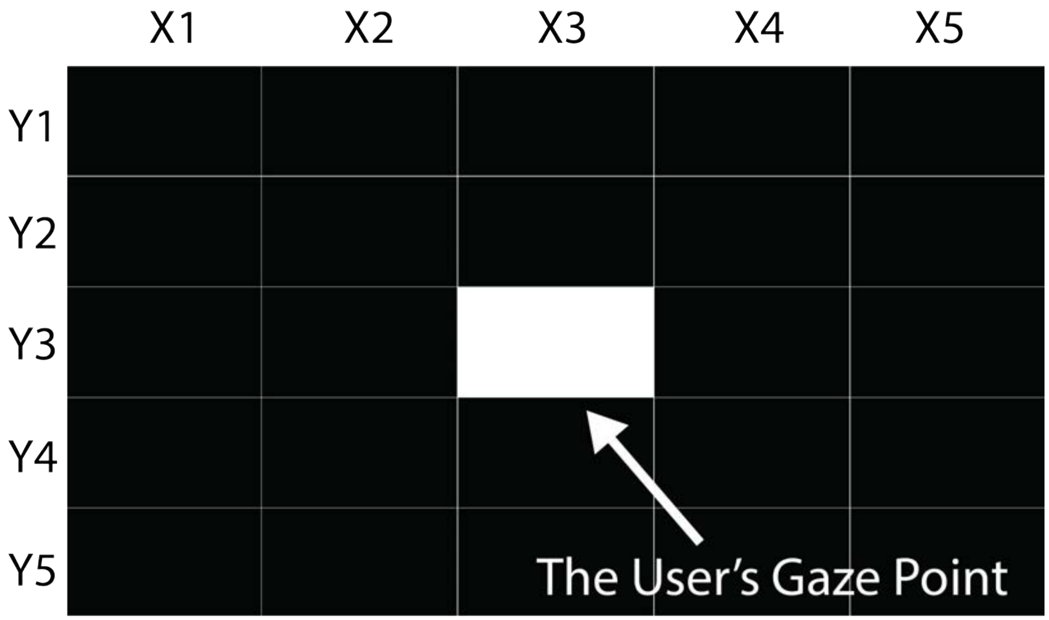


Figure 5-19: The screen shown to the participants

5.3.1 Evaluation: Eye Gaze Position

The subsection evaluated the accuracy of estimating eye gaze position with the device. The estimation of the position was done from the skin deformation caused by the directional change of the eyeballs.

The experimenter had the five participants (all in the 20s, one female). Each of the participants wore the device. They sat down at the distance of about 60 cm from a 23 inches screen. In the screen, 5 x 5 matrix was shown (Figure 5-19). It means each class has approximately 5 degrees (vertical) and 10 degrees (horizontal) of the field of view. The experimenter asked the participants three things during the experiment;(1) look at the colored rectangle on the screen. (2) hold neutral face and blink only during the transition time to reduce the artifacts of the user's behavior (including facial expression change) to the sensor values. (3) follow the colored place with eyes only and not move their head.

After the participants started the software, the colored place changed in order from (X1, Y1), (X2, Y1),...,(X5, Y1),(X1, Y2),..., to (X5, Y5). Whenever the position changed, the color of the rectangle turned to gray for the first 500 ms as the transition.

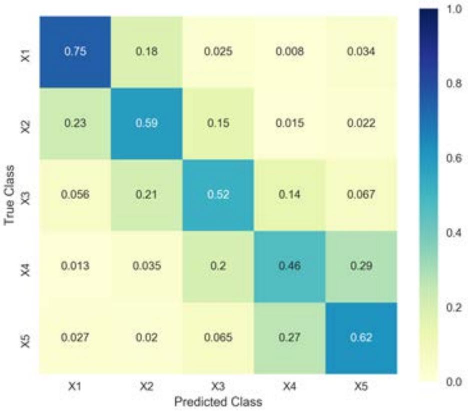


Figure 5-20: Confusion matrix of estimating horizontal gaze direction.

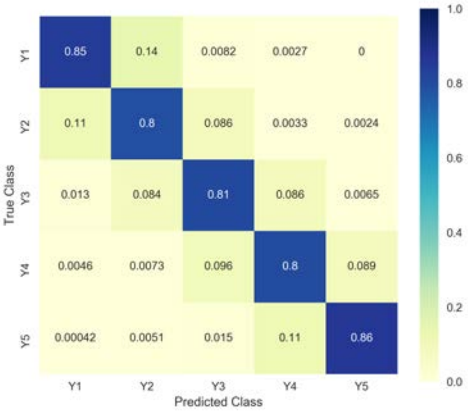


Figure 5-21: Confusion matrix of estimating vertical gaze direction.

In the meantime, the participants changed the gaze position to the gray rectangle. After the color changed to white from gray, the software recorded the sensor data samples(1000ms). This process is to record when the participant gazed at the correct position. Each person repeated the process of looking at 25 positions eight times. It means the dataset of each participant includes the data samples of 8 seconds for all of 25 positions. The data samples from 25 position are labeled based on the position.

Since some of the participants blinked during the recording time, the outliers of each position class were rejected from the acquired dataset. The outlier rejection used the following formula. D is the data samples in each class. d is a data sample that belongs to D. mu and std are 16-dimensional values(average and standard deviation) calculated for each dimension of sensors within the class.

$$outliers = abs(d - mean(D)) < 2xstd(D)\ \forall d \subset D \tag{5.1}$$

Later, each dimension of the datasets was normalized to zero mean and unit variance. Then, 25 classes of the data samples were categorized into five classes in two ways: horizontally and vertically (the five classes in each column or row made a new class). Then 5-fold cross-validation was applied using an SVM classifier (kernel

= rbf, C = 1000) to each dataset. The experimenter repeated the process for every participant.

5.3.2 Results

Figure 5-20 and Figure 5-21 show the average accuracy of each participant's result with individual training. The figures indicate that the sensor data and eye gaze position are correlated. Especially, the vertical movements show a higher correlation(average accuracy 82.4%) than the horizontal movements(average accuracy 58.8%). It means that the vertical movements of eyes cause more skin deformation around the area measured by the device than the horizontal ones. Most of the false predictions are classified as the area next to the True classes. The accuracy is higher on the corner comparing to the central area. It is hard to identify the exact position because the approach only estimated based on the skin deformation measured by the device. The direction change of eyes caused the deformation. However, the device can approximately measure where the user looks at in an experimental condition.

5.3.3 Feasibility Study: Reading Detection

To demonstrate the potential of the device for the analysis of facial expression and eye movements in daily contexts, the feasibility study of reading detection was done. Since reading is essential for learning, reading detection is useful for quantifying and managing the activity to let users read more [50]. Implicit tagging of the facial expressions to the contents can help users to search their potentially favorite contents. It is also beneficial to analyze the contents and to make the contents recommendation.

Advantages of using the device for the purpose instead of wearable eye trackers is that it is possible to consider the information of facial behavior to estimate the emotional response and cognitive states. On the other hand, if the aim is just to quantify reading, the information is noise, which means the accuracy of detecting line break would be worse than the eye trackers.

One participant (a male in the 20s) wearing the device read 10 English jokes.

Jokes were chosen to induce non-neutral facial expression (positive). The jokes are retrieved based on [15]. The lengths of the jokes are from 2 lines to 11 lines. He read the texts shown on the screen in the text box of 900-pixel width on a 23-inch screen (1920 x 1080 p). Soon after he finished reading each joke, he pressed the keyboard in order to record the sensor data samples of the only reading activity. Then, the user also evaluated each of the jokes by 1) how well the user understood the joke (1: not wholly understand - 9: completely understand) 2) how funny it was (1: not funny - 9: very funny) with the 1 - 9 Likert scale. The experimenter also recorded the videos of the wearer's face in the study. It was used to check and count his eye movements and the facial expression change manually.

To the recorded data samples, the system applied a simple moving average of 5 sequences. Later, FastICA from SciKit-learn library was used to process the data samples into 4-dimensional time-series data. To detect independent components (skin deformation caused by facial muscle activity and the skin deformation caused by eye movements are assumed independent), ICA is chosen as it is the standard statistical method for general-purpose. Among the four dimensions, Each dimension of the data corresponds with 1) facial expression change 2) horizontal eye movements 3) blinks and the user's behavior and 4)the other factors such as ambient light noise. The experimenter manually categorized the data. The system applied a moving average of 2-40 sequences depending on the category and the amount of the noise (for blinks: 2-5 sequences, for line breaks: 10 - 20 sequences, and for facial expression: 20 sequences). The peak detection algorithm was used (the Python version of "findpeaks function" from MATLAB Signal Processing Toolbox). The parameter of peak detection algorithm was manually adjusted for each result.

From the recordings, the section introduces three specific examples. The one shows the data of reading activity only when the user kept a neutral face; another one indicates the data with the facial expression change (neutral to positive); the other one illustrates the data with facial expression change, head motion, and body movements. Note that the time scale of each figure is different depending on the length of the jokes.

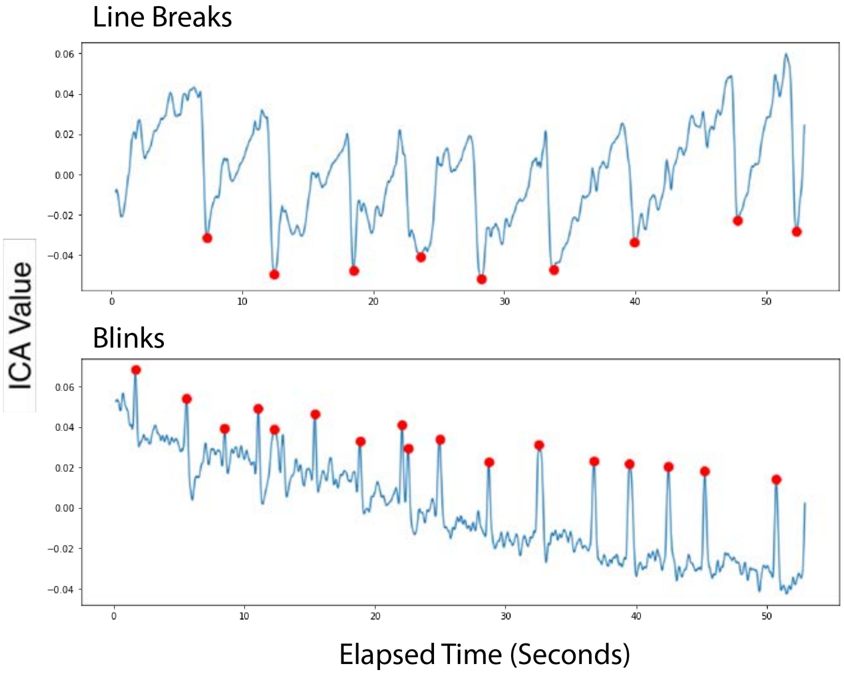


Figure 5-22: The time-series data after applied FastICA to sensor data samples while the user read with a neutral face. Top: line breaks, Bottom: blinks

The first example is the data of 9.5 line joke that the user understood (8 points), and he evaluated as little funny (6 points)(Figure 5-22). From the video, there was no facial expression change. The above figure shows the data of horizontal eye movements. Each peak corresponded with the eye movement that went from the end of a line to the new line. The red dots in the bottom figure show the blinks of the user. All blinks except the last one were successfully detected. The last blink was not detected because the recording ended in the middle of the blink. This example demonstrates the potential of quantifying how many lines or the words the user read [50] using the device.

The second example is the data of 2.5 line joke that the user understood (9 points) and he evaluated as funny (8 points)(Figure 5-23). From the video, the false positives (blue dots, 5 seconds and 18 seconds) on the line break figure were actually backward saccades: the behavior of looking back a couple of words behind. The system successfully detected all of the blinks rather than the last blink since the

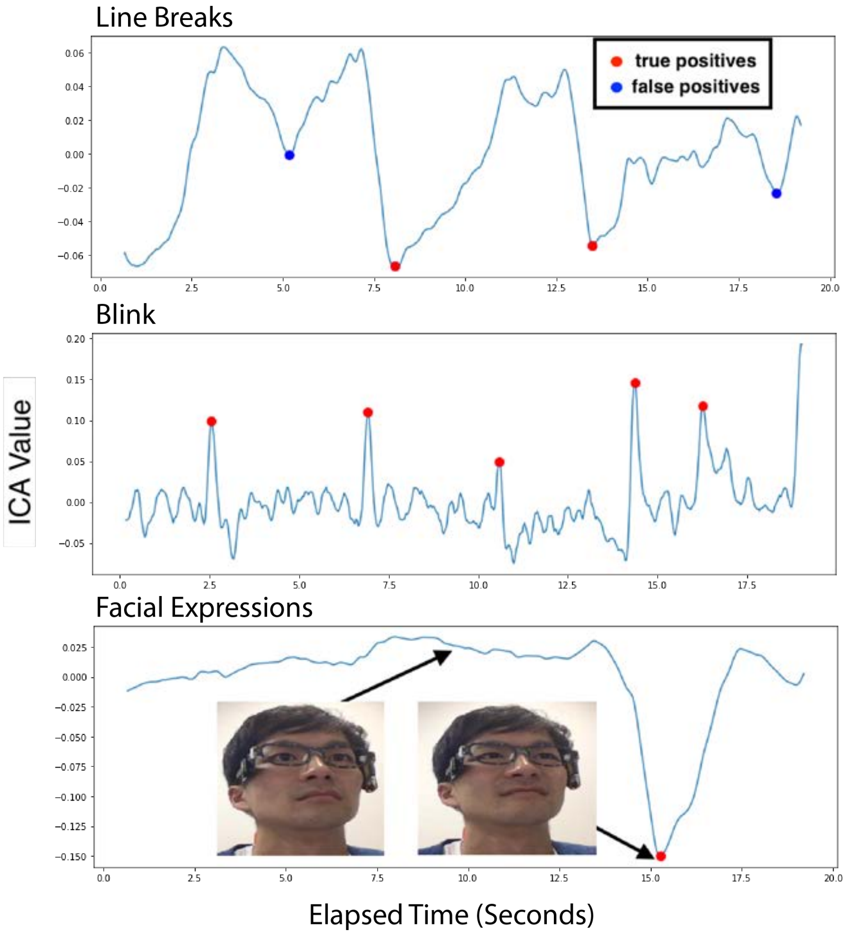


Figure 5-23: The time-series data after applied FastICA to sensor data samples while the user read and smiled in the end. Top: line breaks, Middle: blinks, Bottom: Facial expressions

recording ended in the middle of the last one. The bottom figure shows that one of the ICA time-series data correlated with the actual facial expression change. This example could demonstrate it is still possible to detect line breaks and blinks while detecting the change in facial expressions.

The last example is the data of 3.5 line joke that the user understood(9 points) and he evaluated as funny(8 points)(Figure 5-24). In the above figure, there was the influence of the user's behavior of moving forward to the screen and of looking down on the keyboard. The system detected all of the line breaks, but the influence of the blink caused the wrong line break detection(false positive). If blinks happened

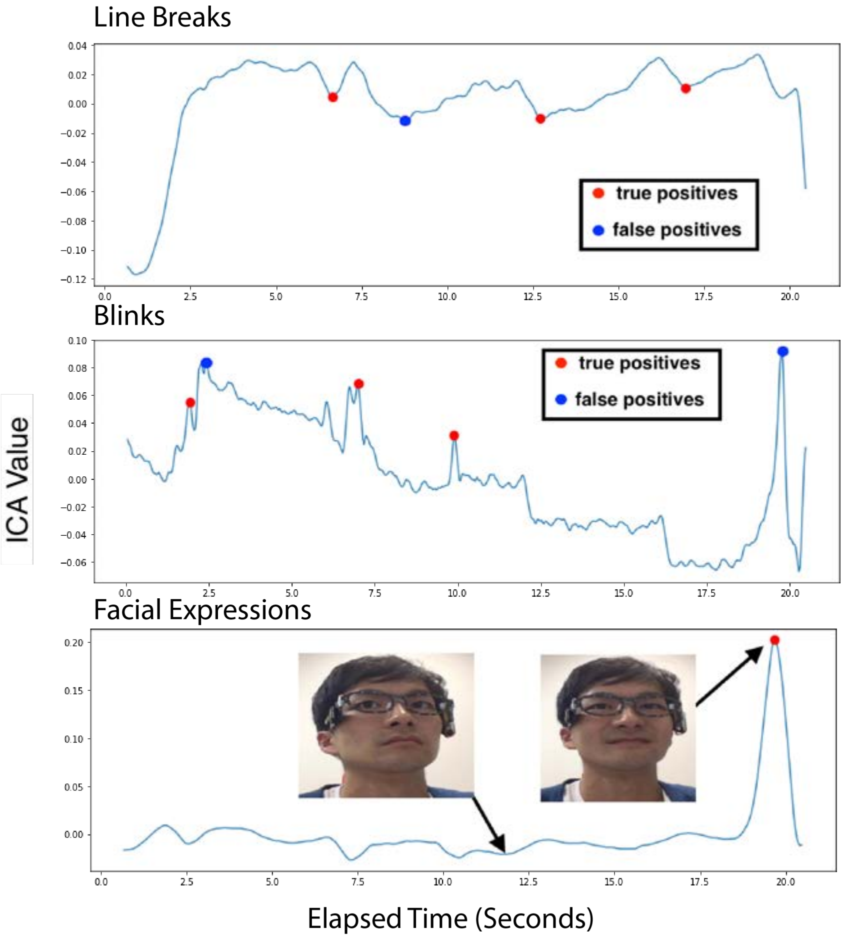


Figure 5-24: The time-series data after applied FastICA to sensor data samples while the user moved, read and smiled . Top: line breaks, Middle: blinks, Bottom: Facial expressions

in the middle of the user's behavior, the blink was not detected. The user's action also caused the wrong detection of the blink. It means the ICA algorithm could not separate all the factors correctly for this example. The figure of the blinks shows that the time-series data included the blinks, the movement of the line break and facial expression change. However, the bottom figure demonstrates the successful detection of the facial response.

5.3.4 Discussion

Although the proposed device can estimate approximate gaze position, it is hard to recognize the exact position of the user's gaze. It means the device cannot be used for eye pointing. However, the information from the device can be combined with an existing eye tracker information. It may improve the accuracy of eye pointing with the eye tracker.

Also, the feasibility study of reading detection showed a potential for implicit tagging and content analysis. The current reading detection can work with stable head positions. To measure correctly, the preprocessing is required to separate facial and eye behaviors from other noises. The future work investigates to tackle the issue and have a large number of the participants with more extended and various contents.

The gaze information is one of the critical aspects of facial expression. People communicate their intention by looking at a talking partner or by avoiding eye contacts. In this sense, combining facial behavior with gaze information could help analyze social interaction deeply.

5.4 Summary

This chapter first proposed the field trials using the posed expression labels. It showed that the device could capture the pattern of the daily activities. Also, the potential of capturing the cognitive load was suggested. Secondly, the chapter showed a method to visualize and summarize facial expression information in a conversation by combining the eyewear device with the array of photo-reflective sensors and image information from the camera. The SOM could map facial expression clusters of individuals in order. With the five to ten minutes of case studies of five users, the section could show that the method could visualize the main expressions and diverse expressions the users showed. The map trained with multiple people showed the clusters shared among the users. The map could show the characteristic expressions of the users and also allow to compare how one user's expression corresponded to the others'. Finally, the chapter showed the potential of detecting eye movements with the device. The experiment demonstrated the possibility of estimating approximately gaze positions in an experiment condition. Another experiment investigated the potential application case of reading detection. Although a user's behavior caused the false positives, the system could detect the blinks and line breaks in addition to facial expression change by applying FastICA.

Chapter 6

Input from Face: Eye Gesture Detection and Hand Over Face Gesture Detection

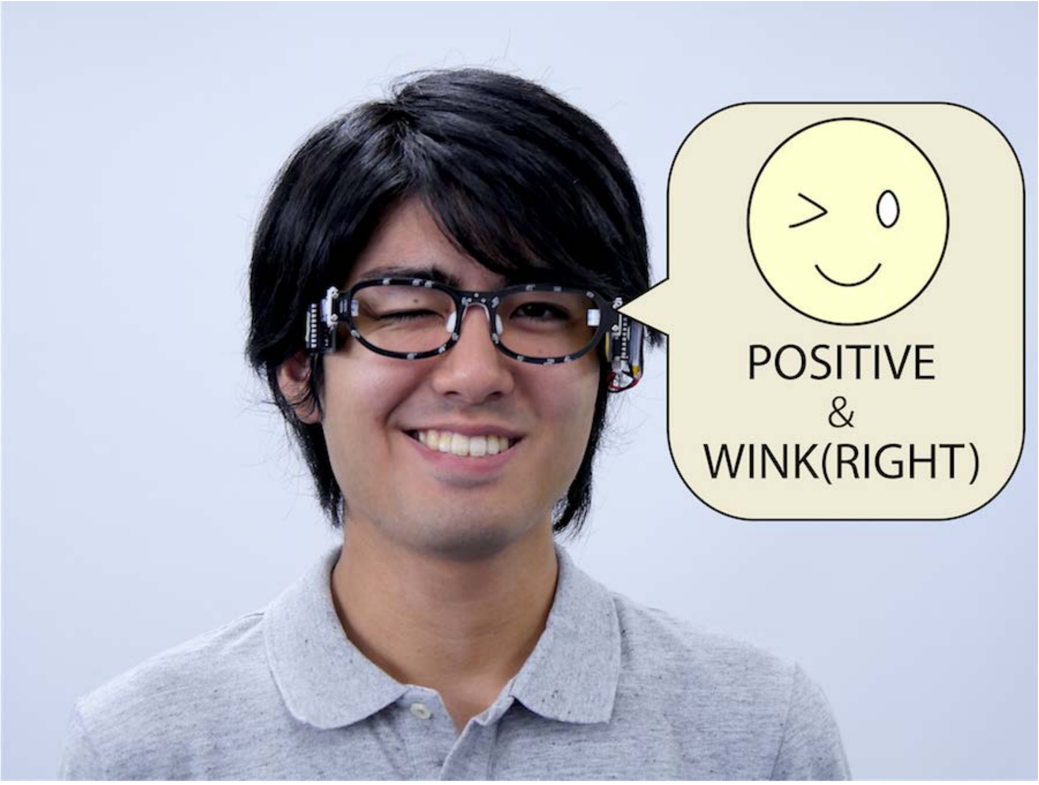


Figure 6-1: The user wears the prototype. The system can detect facial expression and eye gestures

This chapter introduces two input techniques using smart eyewear. The first section describes eye gestures with facial expression recognition. Next section describes the input technique by rubbing face that can be recognized independent from facial expression.

6.1 Eye Gesture Detection

In the field of Affective Computing [75], nonverbal information such as facial expression obtains a great interest to improve the system that detects the emotion of the user. For example, Pham and Wang proposed AttentiveVideo to understand emotional responses to mobile video using physiological signal and facial expressions [74]. They used two cameras on a cell phone. Also, nonverbal information can facilitate efficient and natural interaction between human and computer. Chao et al. proposed a stroke grouping method in sketch recognition using eye gaze information [7]. They showed eye gaze information improved the efficiency of stroke grouping. On the other hand, researchers used information from faces as an input modality. For example, Face typing is a vision-based interface for hands-free text entry that utilizes face detection and facial gesture recognition [26]. These examples suggest that facial expressions and eye gestures have potential as an input interface by providing a command or its context. Most of the previous works used the camera-based method. However, it requires high processing cost and not suitable to use in everyday's life contexts because of mobility limitation. Furthermore, no single wearable device can simultaneously detect both facial expressions and eye movements without cameras. This section presents eye gesture detection pipeline as well as detecting the facial expression states (neutral/positive/negative) of the wearer and his/her eye gestures (up, down, left, and right, wink (left and right), blink) with embedded optical sensors (Figure 6-1) on the device. Both the eye movements and facial expression changes cause the skin deformation around the eyes that the embedded optical sensors measure. The detection algorithms were Support Vector Machine (SVM) and Dynamic Time Warping (DTW) respectively.

The system allows the user to input information to a computer naturally and intuitively with eye gestures because the system can reflect on the user's condition through facial expressions. Since the system is wearable, the users do not need to install any setting in their environments. They just wear the devices to input a command to the computer. The input using face information from the wearable device allows hands-free interaction to the users. It is effective for people who are hand caught or who can not move their hands like people who are driving a car or using the wheelchair. Also, the processing cost is much smaller than the camera since the data from the sensors have lower dimensions (16-dimensional 10-bit values per readings).

The contribution of this section is

1. Development of the algorithms that can detect eye movements with facial expression state.

2. Technical evaluation of detecting eye gestures and facial expressions. The experimenter recorded 210 gestures (The seven kinds of the gestures on three different facial expression condition, ten times) from each of nine people participated in the experiment. The accuracy of detecting seven kinds of the gestures are 92.9% with the user-dependent templates.

6.1.1 Related Work

The proposed work is based on the works from the field of a wearable eye tracker system and the interaction based on information from faces.

A wearable eye tracker such as Pupil [41] can measure eye movements robustly. However, as visual information from a built-in camera requires processing cost, the system needs to have appropriate processors. To overcome this limitation, InvisibleEye uses four low pixel cameras for gaze estimation [97]. This research shows the potential for mobile eye tracking in daily life. JINS MEME is commercial eyewear that measures electrooculogram (EOG) signals and detects eye movements and blinks. The appearance is almost the same as regular glasses. Ishiguro et al. proposed Aided Eyes for human memory enhancement in daily life [35]. Their prototype sensed eye activities using small phototransistors and infrared LEDs. The entire system can be attached to the glasses. However, their system has to be in front of the eyes. It occluded the wearer's vision. While those researchers showed the possibility to detect eye movements in daily life, they do not measure facial expressions. The photosensors on the proposed device are capable of estimating the eye movements. It measures the skin deformation around eyes for estimating the eye movements. The proposed method is also capable of measuring facial expressions.

There are the works using information from the eyes as an interaction technique. Jota and Wigdor explored the design space of eyelid gestures using a commodity camera. They proposed various application cases [39]. A wearable EOG glasses proposed by Bulling et al. allowed the wearer to play a desktop computer game using eye movements [3]. Manabe presented the earphone based interface to detect eye gestures by EOG measurement [56]. They considered the usage in daily life and developed a music player application. Špakov and Majaranta proposed the hands-free interaction system combining gaze pointing and head gestures [89]. Face typing utilized face detection and visual gesture detection to manipulate a scrollable virtual keyboard [26]. Surakka combined the use of two modalities, voluntary gaze direction and facial muscle activation for object pointing and selection [90]. They attached EMG electrodes to the user's face. Their works consider facial and head gestures using cameras and electrodes. Their researchers suggest that the interaction has the potential for hands-free interaction. Cameras require high processing cost while electrodes on face do not fit everyday usage as the appearance is not okay and it is not comfortable to wear for a long time. Besides, Considering the mobility of the device and daily-usage, low signal sensors were preferred. The method considers facial expression and eye gestures in an unobtrusive way.

From related work, the system is the first wearable prototype that has the function of detecting eye gestures and facial expression states while keeping the form factor of ordinary eyewear.

6.1.2 Application Scenario

The advantages of using the system as an interaction technique are 1) hands-free control, and 2)context-based input. The system is suited for a simple interface such as turning on/off the lights, turning the page of e-books, and playing or stopping the music. For example, while streaming a random music playlist, the user can skip to happy music with winks if the user holds a happy face. Also, the combination of eye movement inputs can also be used for a command input like the application of Manabe et al. [56]. Another simple application is inserting emoticon while texting on a mobile phone. The user can input by making a smile and a wink. Since the verbal information sometimes communicates emotions improperly, the emoticon insertion can contribute to smooth communication.

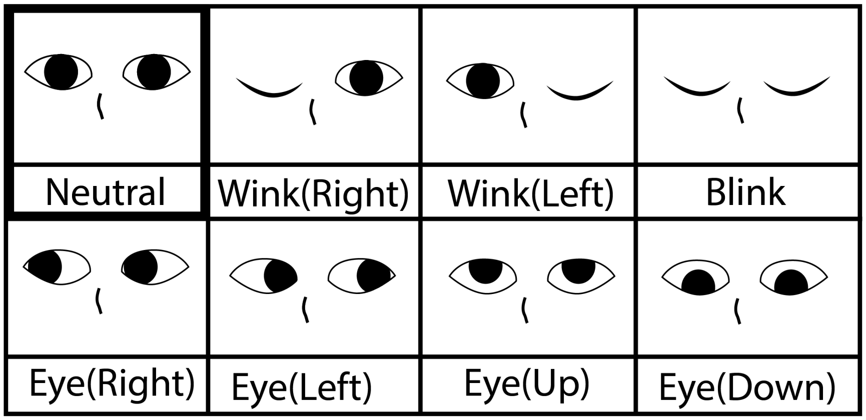


Figure 6-2: The set of gestures. All gestures start and end at neutral eye position on the top left.

6.1.3 Gesture Set

For facial expression states, this section considered three states for simplicity's sake: positive (smile), neutral and negative (frown). When the positive expression is recognized, zygomatic major muscle is activated while negative emotion activates the corrugator supercilii muscles [51]. As such, the experimenter asked the users to activate those muscles in the experiment.

Figure 6-2 shows the gestures to detect with the device. All of the seven kinds of the gestures are temporal gestures, starting from and ending at neutral eye position shown on the top left of the figure. Among the gestures, people make winks only explicitly. It is the advantage of using as an input because it is possible to avoid unconscious inputs. Basic four directions of eye gestures are used for simplicity. For these four gestures, the system considers when the user moves his eyes to specific direction as much as the user can. This assumption can clarify the difference between the four directional gestures.

Blink Detection

Blinks happen involuntary or voluntary. It is better to differentiate between them if the users input by blink in order to avoid an unexpected input. The system makes

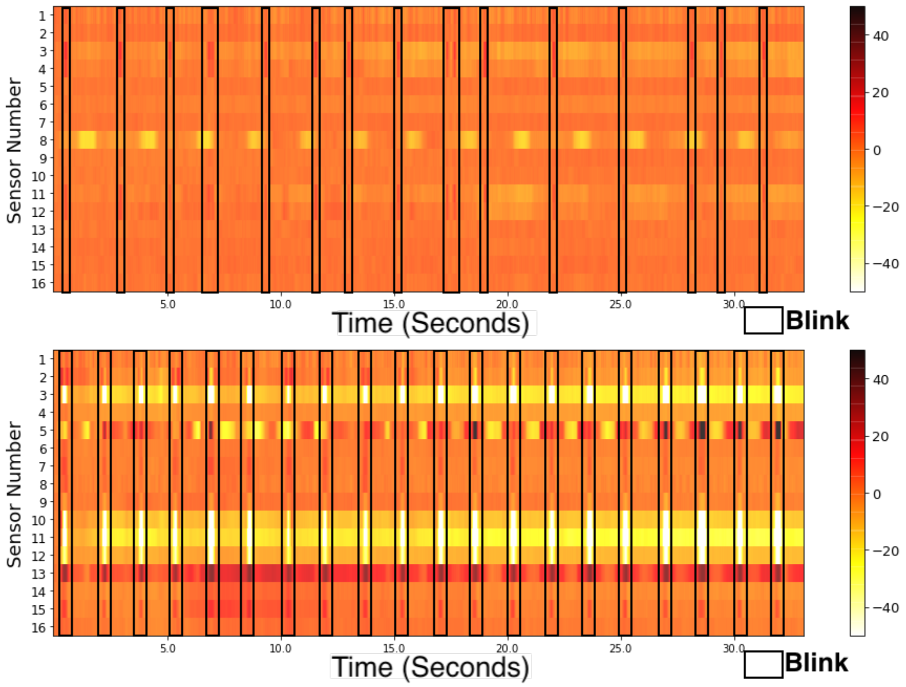


Figure 6-3: The sensor recordings of involuntary blinks (top) and strong voluntary blinks (bottom).

use of voluntary blinks because people can make stronger than involuntary ones. To see the difference between the sensor data of both, the experimenter made two recordings with one participant as a preliminary experiment. He held a neutral face during this study. For involuntary blinks, the experimenter recorded 35 seconds of the data samples while the user watched a neutral video. For voluntary stronger blinks, the experimenter recorded the 20 blinks for 35 seconds. For both recordings, the experimenter recorded videos of the user wearing the glasses. Figure 6-3 shows the heatmap of the results. In the figure, the experimenter annotated the blinks manually by checking the video. The values on the heat maps are the subtraction of the raw data samples in time series and initial raw data sample. The figure shows that the sensor data and the blinks have corresponded. Strong voluntary blinks cause the more significant change to the values from the sensors located in the different places. Moreover, the values of some sensors only change for voluntary blinks. It is because a stronger blink causes the deformation of the upper cheeks. Therefore, strong voluntary blinks can be differentiated from involuntary blinks.

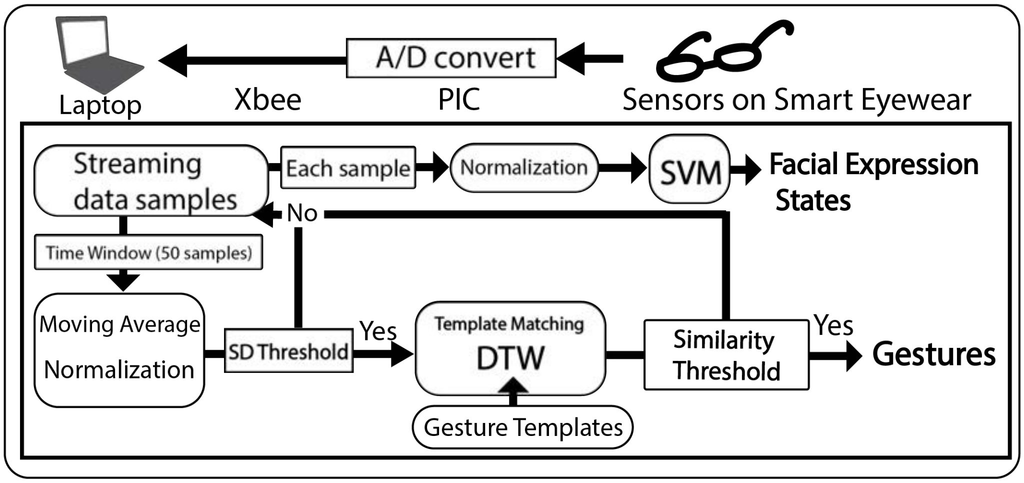


Figure 6-4: The overview of the system. The data samples from the sensors are applied to detect eye gestures with DTW and facial expressions with SVM.

6.1.4 Algorithm

Figure 6-4 describes the overview of the system. The process of eye gesture classification algorithm consists of two stages: data acquisition and preprocessing, and template matching.

Data Acquisition and Preprocessing

The device acquires a 16-dimensional data sample per reading. The sampling frequency is around 30 Hz. From the data streaming, the system read in the latest sample into a buffer. The size of the buffer is for 70 data samples. The system calculates the standard deviation of the data samples for each sensor in the buffer. If the summation of the standard deviation is lower than a threshold, it classifies as no gesture. Otherwise, it regards a gesture in the buffer. Then, it makes a new array that is a simple moving average of 10 sequences to the buffer. It smoothes out the noise. Then, the algorithm normalizes each sensor dimension of the time-series samples in the array separately to a zero mean and unit variance.

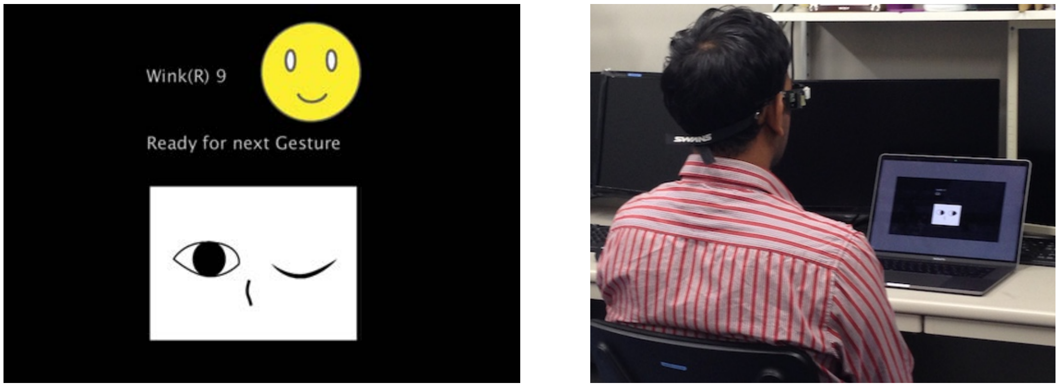
Template Matching

If the gesture is detected, the system compares the time-series with matching templates of all of the seven gestures. If there is the template that is similar enough to the array, the system regards the array as one of the seven gestures. The core of this algorithm is template matching.

The system used one of the most standard time-series similarity measures: Dynamic Time Warping (DTW). This algorithm calculates the distance between two different time-series. The shorter distance means the two are similar. Considering the possibility of real-time detection, the system applied a FastDTW [82]. This algorithm is an approximation of DTW that has a linear time and space complexity. As the signals are multi-dimensional, the system used a distance measure as the summation of absolute difference in all sensor dimensions [93]. The formula for calculating the distance(D) between two K-dimensional time-series, i-th sample of A and j-th sample of B is as follows.

$$D = \sum_{k=1}^{K} |A_i(k) - B_j(k)| \tag{6.1}$$

By performing DTW on the first-order derivatives of the feature values, it is possible to consider the high-level feature of the shape of the time-series [43]. In this case, the system used the derivatives because the data sample at the starting point of the wave differs depending on the position of the device and facial expression states, but how the data samples change over time is consistent to some extent. Therefore, the algorithm compared the similarities between the derivatives of the buffer with the derivatives of all matching templates. The system used 7 matching templates by averaging the resized buffer to 70 samples for each kind of the gestures in the experiment. Through the comparison, the system calculated cost matrix(CS) of the seven calculated distances. The templates are used to classify the buffer signals as the closest gesture template($argMin(CS)$) if (1) $Min(CS)$ is lower than a threshold (2) the relative similarity(RS) is bigger than similarity threshold th. RS is calculated


Wink(R) 9

Ready for next Gesture

Figure 6-5: (left) The user interface used for the recording in the experiment. (right) The experiment setup.

with the following formula.

$$RS = (Min(CS) - SecondMin(CS))/Average(CS) \tag{6.2}$$

The thresholds reject possible confusing gestures or different gestures. The bigger *th* helps to improve the accuracy of classification. However if the threshold it too big, the correctly classified gesture can also be rejected.

6.1.5 Evaluation

The goal of this study is to evaluate the accuracy of detecting eye gestures while users keep three facial expression states. The nine users(eight users are male. They are all in the 20s.) participated in the study. The experimenter recorded the sensor data samples of the gestures with Processing language. To avoid the influence of intense ambient light, the study was run in a quiet room far from windows(Figure 6-5). The following analysis was done in Python environment.

Procedure

Figure 6-6 shows the summary of the procedure in the study. The experimenter recorded 210 gestures (7 kinds of the gestures x 3 facial expression condition x 10 times) for each participant.

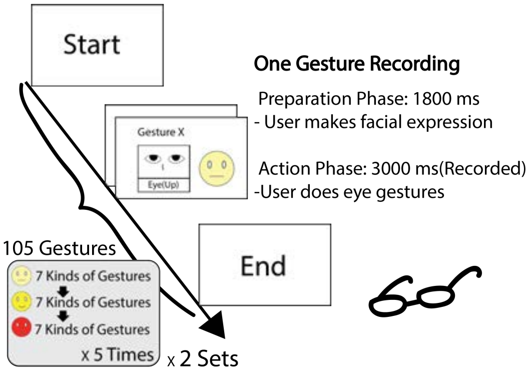


Figure 6-6: The summary of the experiment procedure.

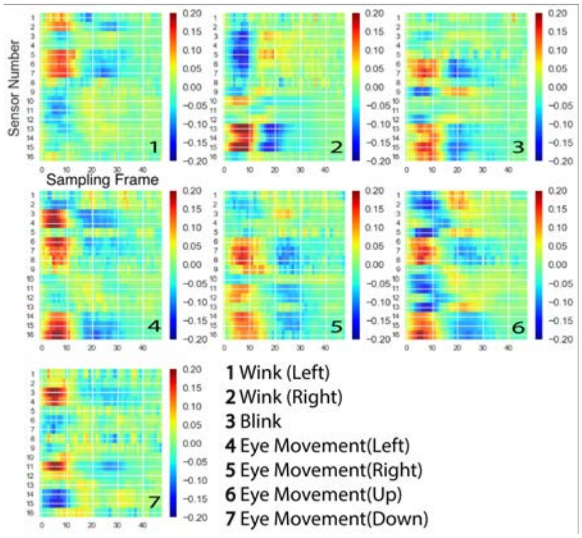


Figure 6-7: The averaged eye gesture templates of all 9 users

The recording of each gesture was divided into two phases. In the preparation phase, the software instructed the user's next gesture and facial expression. In the action phase, the software told to start the gesture. The use of two phrases allowed the participants to start almost the same timing during the recording. It helps to make matching templates. The software recorded the sensor data samples with around 30 Hz only during the action phase.

Firstly, each participant was asked to sit on a chair in front of the laptop on the desk. They wore the prototype with eyewear band strap for stability. The observer introduced the software for the experiment to the participant. The observer explained that the participants make seven different eye gestures ten times each on three different facial expression conditions(neutral, positive and negative) and it took around 20 to 30 minutes in total. The observer told them that each gesture should start and end on neutral eye position(starting at the center of computer screen) and the order of the gesture was periodic so as the participants not to make the wrong gesture. After the general instruction, the observer repeated the following process.

1. The observer start the software for the experiment and remind the participant to keep specific expression in the action phase.

2. The software instruct which kind of gesture and facial expression the participants do during the preparation phase (1800 ms) with the text and the images. Figure 6-5(left) shows the screenshot of the software. In this phase, the user holds the instructed facial expression until the next preparation phase.

3. The software asks the participants to make the instructed gesture in the action phase (3000 ms). The software records the data samples from the sensors in the phase.

4. The process of (3) and (4) are repeated for seven gestures five times each.

5. After the software stops, the observer gave a short break, and go back to the process (1). he observer repeats the process (1) to (5) twice in total.

The observer recorded the video with the built-in camera of the laptop during the experiment. This recording was used to check manually if the participant held a right facial expression and made a right gesture. From the recordings, some participants held the wrong facial expressions when they needed to change. Therefore, the data of the first gesture after the user changed facial expressions was removed from the dataset for facial expression classification. For the one user, the device temporally didn't work in the middle of the recording, and the dataset included 203 gesture data. Therefore, The data for the eye gesture classification come from 1883 (210 x 9 - 7) gestures with 9 participants. For the experiment, těthe standard deviation threshold and similarity threshold was not used since all of the recordings include specific gestures.

Result:Facial Expression

For the classification of facial expressions, SVM (linear kernel, C = 100) was used as a classifier. SVM was trained with the dataset acquired from each participant separately. Each dimension (it includes the time-series sensor values of one sensor from one participant's recorded gestures) from the experiment is normalized to zero mean and unit variance. Then, 10-fold cross-validation was applied to the dataset with

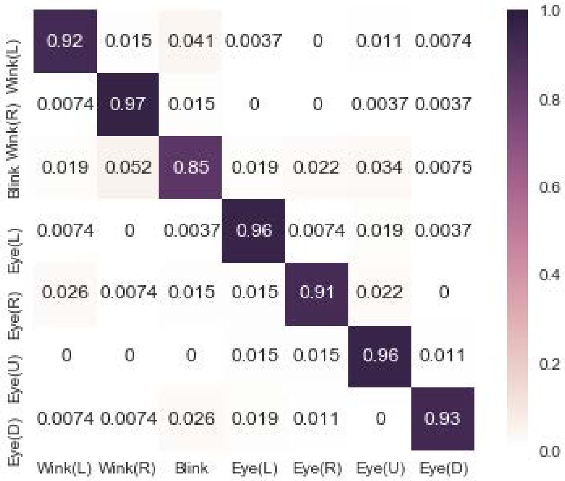


Figure 6-8: Confusion matrix of the accuracy when the system classified seven gestures using the user-dependent templates.

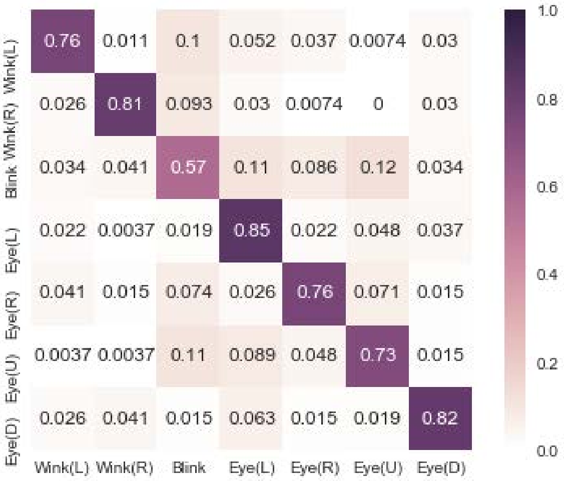


Figure 6-9: Confusion matrix of the accuracy when the system classified seven gestures using the averaged templates of all users.

the SVM classifier. The average accuracy of classifying three facial expression states are 90.9% with individual training. The robustness of facial expression classification regardless of the eye gestures with the proposed eyewear.

Result:Eye Gesture

Templates matching used first 70 data samples of each gesture. Matching templates consisted of all of the recorded gestures and averaged for each kind of the gestures. The average accuracy of classifying seven gestures from 9 participants are 92.9% with user-dependent templates. Figure 6-8 shows the confusion matrix of the accuracy of classifying the seven kinds of the gestures. Among the seven kinds, the system recognized blinks least robustly. The French male showed the lowest accuracy that is 75.2%. It is because when he winks, he tended to close both eyes. 43% of the left eye winks are classified as either the right eye wink or blink. 33% of the blinks are classified as a right eye wink. Also, the experimenter could not control the starting timing of the gesture for him. Every time he made the gestures, he started on his arbitrary timing. It made the features of the templates weak as the templates were the average of the gesture data samples in time series for each kind.

The average templates were made for all kinds of the gestures using the data samples of all the participants (Figure 6-7). The sensor number corresponds to Figure 3-7. The different sensors on both upper and lower side of the frame react to the different gestures. With the templates, the average accuracy is 75.7%. Figure 6-9 shows the confusion matrix. Compared to the individual template, the accuracy gets lower, especially blinks. A part of the reason for this is that the strength of the blinks was not stable within the trial and among the users. However, right eye wink, left eye movement, and down movement are recognized with more than 80%. The average processing time is 63.6 milliseconds per gesture with MacBook Pro (2.9 GHz Intel Core i7).

6.1.6 Discussion

As eyes move implicitly to look at people or the surroundings, the proposed system could not recognize whether the gesture input is intentional or not. It is a common problem in using eye gesture as an interaction technique. This issue can be solved by using an explicit gesture (e.g., wink) as a trigger command. From another perspective, detecting implicit eye gestures opens up the possibility of an ambient interface that understands people. It can facilitate natural interaction with the environments or robots.

The set of the gestures was chosen by focusing on the eye movement-related ones. However, the device has the possibility to recognize facial gesture as well since the device was able to measure the skin deformation caused by facial expressions like [104]. As a trigger command mentioned above, the system can consider the other gestures rather than the set of the gestures since winks are the only explicit gestures in the sets. The winks may not be suitable for some users since, in the experiment, they are not good at winking their eyes.

The demographics of the experiment is biased. Most of the participants are male. Although the shape and features of a face are different depending on the nationality, gender, and ages, the method can work as long as it can measure skin deformation around eyes with close distance. It is possible to adjust the register values for the phototransistors of the sensors. Also, another way is to control the distance by changing the design of the nose pad since the nose pad of the device can be replaced. Considering those factors, the system can work for various people.

6.1.7 Limitation and Future Work

The eye gestures only change a subtle amount of sensor values. The system works with only a stable condition. It means if there are other influences during the gesture, the system may not work well. Possible influences are head motions, facial expression change, the device displacement, and ambient lights.

With the proposed algorithm, there is a risk of classifying non-defined gestures as a target gesture. For future work, it would be better to record other gestures that mostly could happen implicitly such as squinting, and identify what feature should be used to avoid the risk.

The system could classify four directions of eye gestures. In future work, the possibility of classifying diagonal eye gestures such as top-left or bottom-right direction would be explored.

Head motion is also related to non-verbal communication and can be detected with IMU. Besides, IMU can be useful to compensate the sensor value because the head motion and the user's behavior changed the sensor values. With the sensor fusion, the device enables to detect facial expression, eye movement, and head motion at the same time with the form of the everyday glasses.

6.1.8 Summary

This section presented the system to detect eye gestures and facial expressions. The system used DTW for classifying eye gestures and SVM for classifying facial expressions. The average accuracy of detecting seven different eye gestures and classifying three facial expression states are 92.9% and 90.9% respectively with user-dependent training.

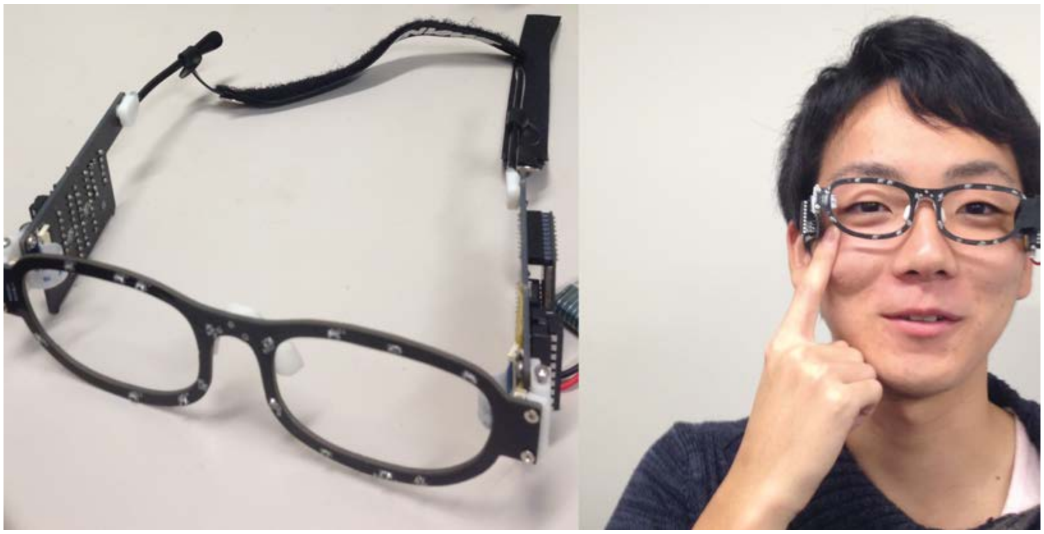


Figure 6-10: Left: the eyewear device used for the recognition. Right: the user makes a rubbing gesture on face.

6.2 FaceRubbing

The focus of this section is hand over face gesture using the smart eyewear. This section is based on the work [62]

6.2.1 Introduction

In recent years, various wearable devices became available such as smart watches, smart headsets, smart clothes and so on. Among wearable devices, smart glasses are attracting attention [2]. Many companies develop optical see-through displays (OST) for augmented reality application such as Hololens developed by Microsoft, MOVERIO BT-300 by Epson and so on. JINS made smart eyewear (JINS MEME) that can measure the physical condition of the user. Due to the development of the MEMS technology, these devices become almost same as ordinary glasses. It allows people to wear such devices in public places (meetings, parties, academic conferences and so on) more often. However, the wearable devices have only the limited input space because of the small size. Operating the device from additional devices is cumbersome. Also, in the situations where OST is used, it is preferable for the user to control the information without being noticed by the other party. For example, people would like to cut off unexpected notifications during meetings or access manuscripts during a presentation while keeping eye contacts on the audiences. However, current input technologies like tapping or flicks use only a limited input space and mid-air gestures are noticeable and not subtle.

This section introduces a new input method of hand-to-face gesture using optical sensors on the smart eyewear (Figure 6-10). Although the device is developed for facial expression recognition, the sensors data information can be used as an input method to a computer. This method allows users to input various commands by rubbing the different areas of the facial surface. The technology can be integrated not only into the proposed device but more general eyewear computing devices such as OSTs because of the small factor of the optical sensors. The advantages of adopting the rubbing gesture are as follows. First, the system can recognize the rubbing gestures independent from facial expression because there is no periodic motion change in facial expression change. It means the gesture recognition method can be used while the system measures facial expressions with the device. Although a directional touch on the face could be recognized using optical sensors on the device, there has the risk of misclassification as a facial expression change. Second, it allows a subtle interaction. The gesture is less obvious than a mid-air gesture or touching gesture to a device since rubbing a face is one of the physiological behaviors that can occur in daily life.

The contributions of this section are 1) Algorithm development for recognizing rubbing gestures on the face. The gesture recognition pipeline consists of three phases for the recognition of the rubbing gesture: pre-processing, gesture detection and gesture classification. 2) Technical Evaluation of the gesture recognition. The study was run with five participants. Then, the section describes the results and the implications. The accuracy of detecting rubbing gesture is 97.5%. The classification accuracy of 10 gesture input spaces is 88.7% with user-independent datasets, 91.7% with user-dependent datasets.

6.2.2 Related Work

Researchers proposed the interaction methods to overcome the limited input space of the devices with a small factor. Harrison et al. presented Skinput that uses the skin on the arm and hand for an input surface [31]. SkinWatch enabled an interaction modality for a smart watch [69]. GestureSleeve is capable of detecting different gestures on touch-enabled sleeves [85]. It allowed the user to control a smartwatch without touching it.

Not only the input method on arms, but many researchers also explored the hand-to-face input method. Serrano et al. investigated facial surface as an input for Head-worn displays [86]. They showed the cheek is a natural surface for the input. They explored the design space for input gestures to faces. They used an infrared optical tracking system with six cameras and a proximity sensor to detect the gestures. Kikuchi et al. proposed EarTouch [44]. The device is a sensor-equipped earphone to enable input to a computer by touching ear. They used photo reflective sensors to recognize the directional touch to the ear. A similar input method of directional touch on cheek surface is developed by Yamashita et al [104]. It made use of photo reflectors to detect the cheek deformation of the user. Hairware is a capacitive touch sensor integrated into a hair extension [100]. It detected a variety of touches to the hair extension for the triggering of different devices. Itchy Nose [53] detected various finger movements to a nose using EOG sensors in smart eyewear. These researchers showed that the facial surface has a potential for input to a computer. In the gesture set of Itchy Nose, rubbing is included. This section focuses on rubbing gestures and consider various areas rather than the nose.

6.2.3 System

The sensors on the device collect the information about periodic skin deformation caused by rubbing gestures. Since the sensors are scattered on the front frame, the system can distinguish the rubbing gestures among different areas on the face. The input space considered in this work is shown in Figure 6-11.

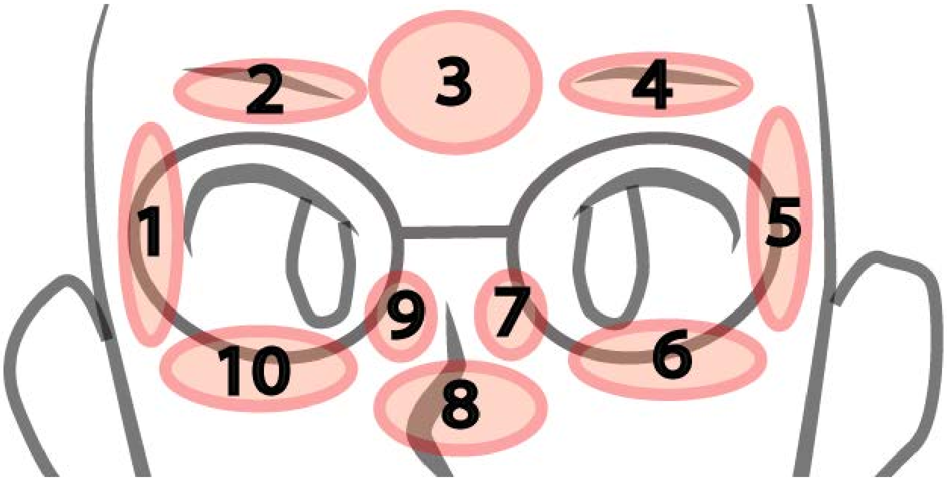


Figure 6-11: The input space for rubbing gestures

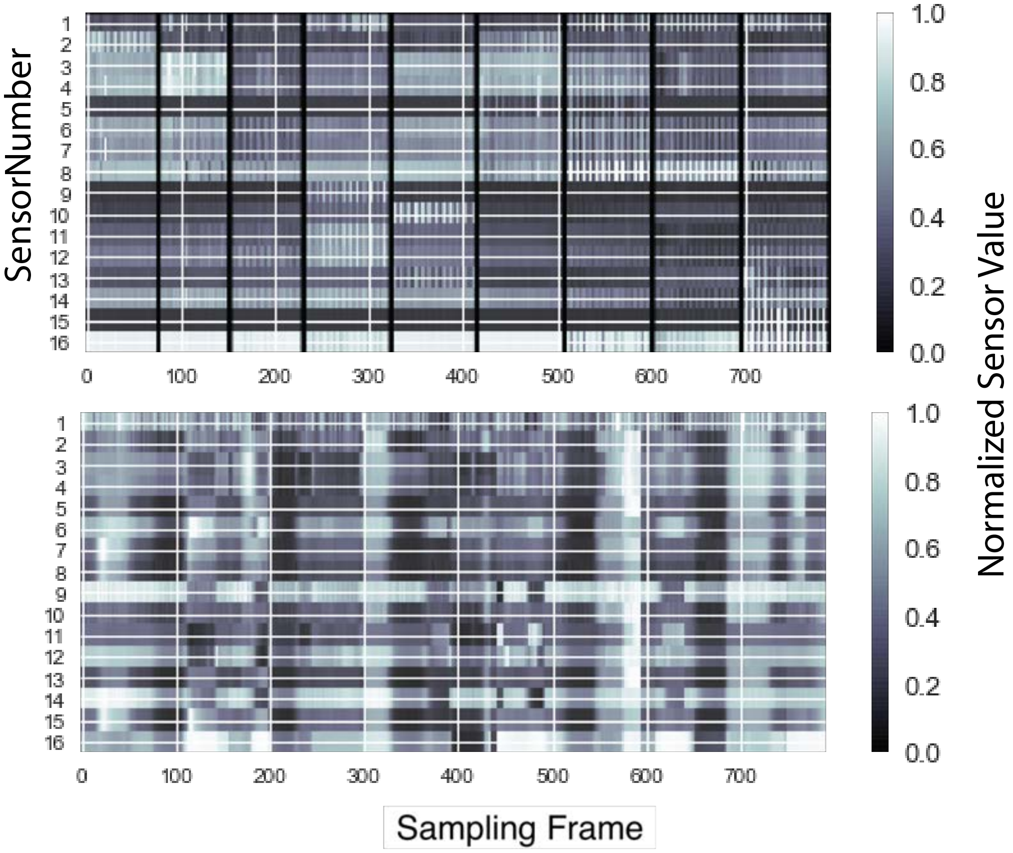


Figure 6-12: The sensor recordings of 1)Top: when rubbing various areas on facial surface 2)Bottom: when moving the facial muscles randomly

6.2.4 Algorithm

Firstly the user recorded the sensor data from the device to see the difference between the changes caused by rubbing gestures and facial gestures (Figure 6-12). The top figure shows sensors values changed while the user rubbed various areas on the facial surface. The bottom figure displays when the user made facial gestures randomly. The figures show normalized each sensor value in the range from 0 to 1 for better visualization. The rubbing gestures caused the periodic sensor value changes in a short period while the facial gesture changed sensor values with low frequency. Based on the characteristics, this section proposes a gesture recognition algorithm as follows. The algorithm consists of three phases; pre-processing, gesture detection, and gesture classification. The algorithm did not misdetect any rubbing gesture from the recording of facial gestures in Figure 6-12 despite the various change of sensor values. The implementation of the algorithm was done in Python.

Pre-Processing

The device acquires a 16-dimensional data sample per reading. The data sample is a subtraction of the sensor values when the infrared LEDs of the sensors are on and off. All the sensors are actuated at the same time. The sampling frequency is around 30 Hz. The system applies a simple moving average of three sequences to the data samples. The system uses a sliding window of 40 samples (i.e., around 1.3 seconds). The algorithm is run every ten frames for the gesture detection.

Gesture Detection

The system applies PCA to the data samples. PCA reduces the dimensions to one-dimensional time-series data. It includes the most dominant trend from the data samples. After normalizing the time-series data to fit in the range from 0 to 1, the system applies peak detection algorithm (the python version of MATLAB peak detection algorithm). It used the threshold determination to detect the rubbing gestures. The system counts the number of upper and lower peaks. If there are more than five peaks in the data, the system regards the data as the rubbing gesture to a particular area. It means the algorithm detects the gesture if there are the three rubbings in the time window. Since there is a slight chance that the noise from ambient light can also cause many small peaks when the user does not move and make any gesture, the method excluded the data from the detection algorithm if the summation of absolute values of the derivative of the data samples in the window is less than a threshold.

Gesture Classification

If the rubbing gesture is detected, the system extracts the features from the data samples in the window. The features are 16-dimensional: the summation of the absolute values of the derivatives of each dimension of the data samples. It normalizes the features so that the summation of the feature values is 1. Then, the system applies a random forest classifier(max-depth is 10) to the features in order to classify which area on the face is rubbed. For the machine learning algorithm, the most important issue is robust recognition. Therefore, by comparing the basic algorithms such as SVM, adaBoost, a random forest classifier is choosen as it showed best accuracy in the preliminary experiment.

6.2.5 Technical Evaluation

The goal of this evaluation is to investigate the accuracy of the proposed method. The five users (all of them are male in the 20s) participated in the study. For this evaluation, the experimenter recorded the sensor data samples of the rubbing gestures. The recording was done using Processing language. To avoid the influence of intense ambient light, the experimenter conducted the study in a quiet room far from windows. The following analysis used Python environment. The system assumed the user makes the gestures only with a neutral face. Therefore, the participants held neutral face during the experiment.

Firstly, each participant sat on a chair in front of the laptop on the desk. They wore the prototype with an eyewear band strap for stability, and the observer started the software after explaining the experiment detail.

In the first stage, the observer recorded 100 gestures (Rubbing gestures to 10 areas X 10 times) each by each from every participant. The numbers of the input areas are corresponded to Figure 6-11. The order of the gestures the participants make is periodic (1 to 10). The software showed the figure of the input space and the number of the gesture area so as the participant can quickly understand which area to rub. In order not to limit how to rub, the experimenter did not instruct which hand the user should use for the rubbing, the speed, and direction of the rubbing during the recordings. Each recording of the gesture is divided into two phases. In the preparation phase (2000 ms), the users settled the rubbing position and started rubbing on a particular area. In the recording phase (3000 ms), the software recorded the sensor data samples while the users were rubbing.

In the second stage, the observer recorded 20 gestures (10 areas X 2 times) from each participant. The recording was divided into four trials. The observer told each participant to make the gestures in specific ways (the number of the gesture area: 1-2-3-4-5, 6-7-8-9-10, 9-10-1-2-3, 4-5-6-7-8). In one time of the trial, the users made five gestures in 25 seconds. Every time the users made a gesture to a specific area, the users put their hands on their kees. This procedure created "no detection time" between the gestures. There were 5 seconds break between the trials.

Results

First, the analysis was done for the data from the first experiment. 80 data samples from the recording of each gesture were extracted first. For the analysis, these samples were split into the two (first 40 samples and last 40 samples), i.e., two gesture data from one rubbing gesture. After excluding the outliers due to a lack of enough recording samples or a strong noise, the acquired dataset includes 937 gesture data from the recordings of five participants in total. The dataset was shuffled randomly and applied five cross-validation method. Overall, the accuracy of classifying ten gestures is 88.7% with user-independent training. Figure 6-13 shows the confusion

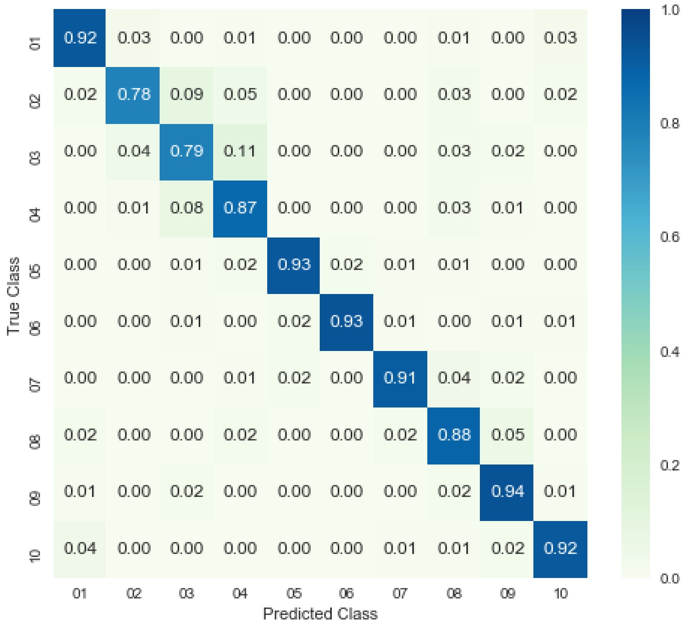


Figure 6-13: Confusion matrix of accuracy detecting rubbing gestures to ten different spaces

matrix. Most of the false positives come from the adjacent areas of the true positives. With user-dependent datasets, the average accuracy is 91.7%. Table 6.1 shows the result of each user. The accuracy of user A and B are around 80% that is about 10 % lower than the accuracy of the others. It is because some specific areas of the gestures of user A and B showed relatively less accuracy. For example, 38% of the gestures of the User A to the area 2 is predicted as the gesture to area 3. Also, 41% of the gestures of the User B to the area 4 is predicted as the gesture to area 3. In order to hold higher accuracy, the system can make use of only the areas not close to each other. For example, the accuracy of classifying five areas (1,3,5,7,9) is 95.6% with user-independent datasets.

User	A	B	C	D	E	Average
Accuracy(%)	82.2	82.7	96.7	98.5	98.4	91.1

Table 6.1: The accuracy of classifying 10 gestures with user-dependent datasets

Regarding the dataset acquired in the second experiment, the recording of each participant was merged into one time-series data. Then, the algorithm was applied to

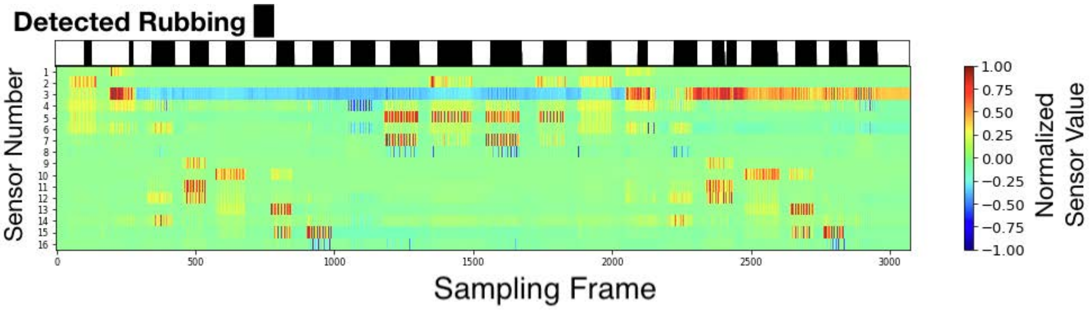


Figure 6-14: The time-series recordings

the dataset. The classifier was trained individually using each participant's dataset from the first experiment. Each detected gesture have the sequences of the predicted results because the length of the gestures the users made is different every time. The dominant prediction result in the sequences is used for estimating accuracy. The dataset of User A was eliminated since the strong noise was measured. As the device stopped for a short time while the user E used, the one gesture was not recorded. Table 6.2 illuminates the summary of recordings from the second experiment. The average accuracy of detecting the gesture is 97.5%. The F1-Score is 0.987. The average accuracy of classifying the gestures of the true positives is 91.0%.

User	B	C	D	E	ALL
Detection(TP)	20	20	20	18	78
Detection(FP)	1	0	0	0	1
Detection(FN)	0	0	0	1	1
Detection(F1-Score)	0.976	1.00	1.00	0.973	0.987
Classification Accuracy(%)	80.0	90.0	100	94.4	91.0

Table 6.2: The summary of recording from the second stage. TP:True Positive, FP: False Positive, FN: False Negative.

Figure 6-14 shows the heatmap of the sensor data and the result of the gesture recognition of User D who showed the best result among the users. To make better visualization, the figure shows (all the data samples - the initial values of the time-series data samples)/maximum absolute value for each sensor dimension. Although the system detected all the gestures, there was an unrecognized gap in the series of the gesture. This case was regarded as one gesture because the gap was short.

6.2.6 Discussion and Limitation

This section focused on the area around eyes for the input since rubbing on the area caused a significant change of sensor values. However, rubbing on the cheek can also cause the skin deformation on the areas covered by the sensors. It means that the system can use the broader area of the face. It can expand the input areas, but it also could reduce the accuracy of the gesture recognition.

There is a trade-off between the length of rubbing and robustness of the gesture recognition. The window size of 40 was used for the gesture recognition. If the size is too big, the algorithm may not detect the gestures correctly. Also, the user has to make rubbing for a long time to be detected. On the other hand, if the size is too small, it causes false detections of the gesture. For example, when the user blinks continuously in short time, the gesture detection may recognize as a gesture.

Ambient light may inhibit from detecting gestures because intense light makes sensor values saturated. Therefore direct sunlight should be avoided when users use the proposed method.

An intentional rubbing gesture may cause some problems. If the user rubs too much on one particular area, this may lead to rough skin. Besides, the user with makeup may hesitate to make a rubbing gesture because the makeup comes off by the gesture.

6.2.7 Summary and Future Work

This section presented the input method by rubbing a facial surface using photo reflective sensors on smart eyewear. The input is subtle since rubbing a facial surface can occur as a physiological behavior in daily life. Although the system is developed for facial expression recognition, the system can recognize rubbing gestures to various areas independent of facial expression change. The accuracy of classifying rubbing gestures in 10 different areas of the face is 88.7% with user-independent datasets and 91.7% with user-dependent datasets. F1-Score of the gesture detection is 0.987.

One of the future works is to make the applications based on the recognized gesture patterns. Simple aplication is to cut off an unexpected phone call during the meeting with a subtle gesture. Another application is to measure human mind through the hand-to-face implicit movements. Implicit movements are an effective means to know the state of the person. According to [68], the hand-to-face gesture can reveal the mind of the person such as. The knowledge is useful to analyze daily user behaviors. Also, the hand-to-face gestures sometimes are not appropriate behavior in public. The system can be applied to monitor those gestutes to improve a social manner.

Chapter 7

Conclusion

This dissertation presents novel smart eyewear that classifies the wearer's facial expressions in daily scenarios. The device can keep track of facial expressions in daily life, and it considers socially acceptance because of its typical form of the glasses. The device uses embedded photo reflective sensors and machine learning for the classification. The approach focuses on skin deformations around the eyes that occur when the wearer changes their facial expressions. Small photo-reflective sensors on the device measure the distances between the skin surface on the face and the array of sensors embedded in the eyewear frame. The sensors can cover various facial muscle movements with low processing cost. Also, they are small and light enough to be integrated into daily-use glasses. The device can classify facial expressions regardless of face directions and the occlusions such as hand gestures and hairs.

For basic emotion classification, a Support Vector Machine (SVM) algorithm is applied to the information collected by the sensors. The evaluation of the device shows the robustness to the noises from the wearer's facial direction changes and the slight changes in the glasses' position, as well as the reliability of the device's recognition capacity. The recognition accuracy of classifying eight basic facial expressions in daily scenes was 92.8% accuracy regardless of facial direction and removal/remount. The device can classify eight basic facial expressions with 78.1% accuracy for repeatability and 87.7% accuracy in case of its positional drift.

For spontaneous facial expression recognition, initial field trials in a daily life setting were undertaken to test the usability of the device. Besides, the system was used to detect facial expressions in daily conversations. It shows a novel unsupervised way to summarize and visualize the individual spontaneous facial expressions using a wearable device. The method synchronized the data with camera images to create the visualization. Through the case studies of five minutes, unscripted communications of the five to ten minutes each revealed the approach could map the main facial expressions and the diverse expressions of the users in order. The study also demonstrated that the map trained with the datasets of five users could categorize the similar expressions of each user into the shared clusters among the users. The map enabled to compare the difference in each user's distribution of facial expressions. Similar expressions of each user were mapped into the shared clusters among the users. Not only facial expressions but also eye movements could be detected with the device. The evaluation showed the accuracy of eye gaze position estimation with five users holding a neutral face. The system showed higher accuracy to detect y position than x position. The system was also capable of analyzing both facial expressions and eye movements in daily contexts as the feasibility study of one user reading jokes while wearing the device shown.

For the interaction purpose, the smart eyewear can input information by eye gesture within the context of facial expressions. The system allows hands-free interaction in many situations. The evaluation was done to see the accuracy of detecting the gestures and facial expressions with 9 participants. The average accuracy of detecting seven different eye gestures and classifying three facial expression states are 92.9% and 90.9% respectively with user-dependent training. The method used Support Vector Machine (SVM) for facial expression classification and Dynamic Time Warping (DTW) for gesture recognition. An input technique to a computer by rubbing face was also proposed to make use of the limited input space. Since rubbing gesture occurs in daily life, the system enables a subtle interaction between the user and a computer. The embedded optical sensors measure the skin deformation caused by rubbing on the face. The system detects the gestures using principal component analysis (PCA) and peak detection. It classifies the area of the gesture with a random forest classifier. The accuracy of detecting rubbing gesture is 97.5%. The classification accuracy of 10 gesture area is 88.7% with user-independent training.

The dissertation demonstrated the system that can measure various facial gestures including eye gestures. It could open up the potential of quantifying diverse facial expression activities in daily life and new interaction methods for smart glasses.

For the future work, the evaluation of the spontaneous emotional expression is important. One possible evaluation is to associate the subjective rating of the emotional stimuli video and the sensor data of facial behaviors while the user is watching the video. In addition, the cognitive part of facial expression can be explored. For example, fatigue detection for the patients, confusion detection in learning, or lie detection are an interesting direction. Besides, the dynamics of facial expression should be considered by using the temporal features based learning algorithms.

Based on the dissertation, I would like to explore the wearable computing to support to keep user fit and to acquire the new skills. As an ultimate goal, the use of the computing will make people do what they intend to do eventually without the power of the computing. It is because I believe that people feel joy in the sense that they can do they want with their own ability. For this purpose, I believe the embodiment of the device is important. It is necessary to consider both two aspects of our body at the same time. First one is the senses. We measure the environmental stimuli or inner state through the senses. Mainly I focus on the measurement of the mind, which is the essence of our behavior. It does not mean that the body movement is not the target, but I consider the movement as the embodied mind. The critical step of the measurement is to make people aware of what they unconsciously do. Second is the actuation. We actuate to the environment or other people mentally or physically through the hands or facial expressions and so on. For example, we smile to change the other's social interaction. The key step of the actuation is to make people do what they intend to do without consciousness. By considering both aspects, the wearable device can naturally be a part of the body. The feedback from the device is immediate based on the measurement, which can accelerate the learning. It can help the user to change their behavior to achieve their goal.

Bibliography

[1] V. Bettadapura. Face Expression Recognition and Analysis: The State of the Art. *CoRR*, abs/1203.6, 2012.

[2] A. Bulling and K. Kunze. Eyewear Computers for Human-computer Interaction. *interactions*, 23(3):70–73, apr 2016.

[3] A. Bulling, D. Roggen, and G. Tröster. It's in Your Eyes: Towards Context-awareness and Mobile HCI Using Wearable EOG Goggles. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, UbiComp '08, pages 84–93, New York, NY, USA, 2008. ACM.

[4] A. Calvo R. and S. D'Mello. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *Affective Computing, IEEE Transactions on*, 1(1):18–37, jan 2010.

[5] W. B. Cannon. The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory. *The American Journal of Psychology*, 39(1/4):106–124, 1927.

[6] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time High-fidelity Facial Performance Capture. *ACM Transactions on Graphics*, 34(4):46:1–46:9, jul 2015.

[7] B. Chao, X. Zhao, D. Shi, G. Feng, and B. Luo. Eyes Understand the Sketch!: Gaze-Aided Stroke Grouping of Hand-Drawn Flowcharts. In *Proceedings of the*

22Nd International Conference on Intelligent User Interfaces, IUI '17, pages 79–83, New York, NY, USA, 2017. ACM.

[8] A. Clark. *Being There: Putting Brain, Body, and World Together Again.* MIT Press, Cambridge, MA, USA, 1st edition, 1996.

[9] J. F. Cohn and F. la Torre. Automated Face Analysis for Affective Computing. *The Oxford handbook of affective computing*, pages 131–150, 2014.

[10] J. F. Cohn, K. Schmidt, R. Gross, and P. Ekman. Individual Differences in Facial Expression: Stability over Time, Relation to Self-Reported Emotion, and Ability to Inform Person Identification. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ICMI '02, pages 491—-, Washington, DC, USA, 2002. IEEE Computer Society.

[11] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, Jun 2001.

[12] C. Darwin, P. Ekman, and P. Prodger. *The expression of the emotions in man and animals.* Oxford University Press, USA, 1998.

[13] U. Dimberg, M. Thunberg, and K. Elmehed. Unconscious Facial Reactions to Emotional Facial Expressions. *Psychological Science*, 11(1):86–89, 2000.

[14] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[15] R. I. M. Dunbar, J. Launay, and O. Curry. The Complexity of Jokes Is Limited by Cognitive Constraints on Mentalizing. *Human Nature*, 27(2):130–140, jun 2016.

[16] P. Ekman. The argument and evidence about universals in facial expressions. *Handbook of social psychophysiology*, pages 143–164, 1989.

[17] P. Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[18] P. Ekman. Facial Action Coding System (FACS). *A Human Face*, 2002.

[19] P. Ekman and W. V. Friesen. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, 6(4):238–252, jun 1982.

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

[21] C. Frith. Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1535):3453–3458, 2009.

[22] C. D. Frith and U. Frith. Mechanisms of Social Cognition. *Annual Review of Psychology*, 63(1):287–313, 2012.

[23] K. Fukumoto, T. Terada, and M. Tsukamoto. A Smile/Laughter Recognition Mechanism for Smile-based Life Logging. In *AH*, New York, NY, USA, 2013. ACM.

[24] J. M. Girard, J. F. Cohn, L. A. Jeni, S. Lucey, and F. D. Torre. How much training data for facial action unit detection? In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8, may 2015.

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

[26] Y. Gizatdinova, O. Špakov, and V. Surakka. Face typing: Vision-based perceptual interface for hands-free text entry with a scrollable virtual keyboard. In *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*, pages 81–87, jan 2012.

[27] R. P. G. Van Gompel, M. H. Fischer, W. S. Murray, and R. L. Hill. Chapter 1 - Eye-movement research: An overview of current and past developments. In Roger P G Van Gompel, Martin H Fischer, Wayne S Murray, and Robin L Hill, editors, *Eye Movements*, pages 1–28. Elsevier, Oxford, 2007.

[28] J. J. Gross. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299, 1988.

[29] J. J. Gross. *Handbook of Emotion Regulation*. Guilford publications, 2013.

[30] A. Gruebler and K. Suzuki. Design of a Wearable Device for Reading Positive Expressions from Facial EMG Signals. *Affective Computing, IEEE Transactions on*, 5(3):227–237, jul 2014.

[31] C. Harrison, D. Tan, and D. Morris. Skinput: Appropriating the Body As an Input Surface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 453–462, New York, NY, USA, 2010. ACM.

[32] T. F. Heatherton, J. Polivy, C. P. Herman, and R. F. Baumeister. Self-Awareness, Task Failure, and Disinhibition: How Attentional Focus Affects Eating. *Journal of personality*, 61(1):49–61, 1993.

[33] R. Hess, U., Banse and A. Kappas. The intensity of facial expression is determined by underlying affective state and social situation. In *Journal of Personality and Social Psychology*, volume 69, pages 280–288. American Psychological Association, 1995.

[34] L. Inzelberg, D. Rand, S. Steinberg, M. David-Pur, and Y. Hanein. A Wearable High-Resolution Facial Electromyography for Long Term Recordings in Freely Behaving Humans. *Scientific Reports*, 8(1):2058, 2018.

[35] Y. Ishiguro, A. Mujibiya, T. Miyaki, and J. Rekimoto. Aided Eyes: Eye Activity Sensing for Daily Life. In *Proceedings of the 1st Augmented Human International Conference*, AH '10, pages 25:1–25:7, New York, NY, USA, 2010. ACM.

[36] E. Jakobs, A. S. R. Manstead, and A. H. Fischer. Social Motives, Emotional Feelings, and Smiling. *Cognition and Emotion*, 13(4):321–345, 1999.

[37] W. James. *The principles of psychology, Vol II.* Henry Holt and Company, NY, US, 1890.

[38] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. De La Torre. Continuous au intensity estimation using localized, sparse facial feature space. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–7, April 2013.

[39] R. Jota and D. Wigdor. Palpebrae Superioris: Exploring the Design Space of Eyelid Gestures. In *Proceedings of the 41st Graphics Interface Conference*, GI '15, pages 273–280, Toronto, Ont., Canada, Canada, 2015. Canadian Information Processing Society.

[40] T. Kanade, J. .F Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53, 2000.

[41] M. Kassner, W. Patera, and A. Bulling. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14 Adjunct, pages 1151–1160, New York, NY, USA, 2014. ACM.

[42] D. Keltner, P. Ekman, G. C. Gonzaga, and J. Beer. Expression of emotion. *Handbook of Affective Sciences*, pages 411–414, 2003.

[43] E. J. Keogh and M. J. Pazzani. Derivative Dynamic Time Warping. In *Proceedings of the 2001 SIAM International Conference on Data Mining*, pages 1–11. SIAM, 2001.

[44] T. Kikuchi, Y. Sugiura, K. Masai, M. Sugimoto, and B. H. Thomas. EarTouch: Turning the Ear into an Input Surface. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '17, pages 27:1–27:6, New York, NY, USA, 2017. ACM.

[45] S. Kimura, M. Fukuomoto, and T. Horikoshi. Eyeglass-based Hands-free Videophone. In *Proceedings of the 2013 International Symposium on Wearable Computers*, ISWC '13, pages 117–124, New York, NY, USA, 2013. ACM.

[46] D. E. King. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[47] L. Knapp, M., J. A. Hall, and T. G. Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[48] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, Jan 1982.

[49] K. Kulkarni, C. A. Corneanu, I. Ofodile, S. Escalera, X. Baro, S. Hyniewska, J. Allik, and G. Anbarjafari. Automatic Recognition of Facial Displays of Unfelt Emotions. *arXiv preprint arXiv:1707.04061*, 2017.

[50] 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*, UbiComp '15, pages 87–96, New York, NY, USA, 2015. ACM.

[51] J. T. Larsen, C. J. Norris, and J. T. Cacioppo. Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, 40(5):776–785, 2003.

[52] D. H. Lee and A. K. Anderson. Reading What the Mind Thinks From How the Eye Sees. *Psychological Science*, 28(4):494–503, 2017.

[53] J. Lee, H. Yeo, M. Dhuliawala, J. Akano, J. Shimizu, T. Starner, A. Quigley, W. Woo, and K. Kunze. Itchy Nose: Discreet Gesture Interaction Using EOG Sensors in Smart Eyewear. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ISWC '17, pages 94–97, New York, NY, USA, 2017. ACM.

[54] H. Li, L. Trutoiu, K. Olszewski, L. Wei, T. Trutna, P. Hsieh, A. Nicholls, and C. Ma. Facial Performance Sensing Head-mounted Display. *ACM Transactions on Graphics*, 34(4):47:1–47:9, jul 2015.

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

[56] H. Manabe, M. Fukumoto, and T. Yagi. Conductive Rubber Electrodes for Earphone-based Eye Gesture Input Interface. *Personal Ubiquitous Comput.*, 19(1):143–154, jan 2015.

[57] B. Martinez and M. F. Valstar. *Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition*, pages 63–100. Springer International Publishing, Cham, 2016.

[58] K. Masai, K. Kunze, Y. Sugiura, M. Ogata, M. Inami, and M. Sugimoto. Evaluation of Facial Expression Recognition by a Smart Eyewear for Facial Direction Changes, Repeatability, and Positional Drift. *ACM Transactions on Interactive Intelligent Systems*, 7(4):15:1—-15:23, dec 2017.

[59] K. Masai, Y. Sugiura, M. Ogata, K. Kunze, M. Inami, and M. Sugimoto. Affectivewear: Recognizing wearer's facial expression by embedded optical sensors on smart eyewear (in japanese). *Transactions of the Virtual Reality Society of Japan*, 21(2):385–394, 2016.

[60] K. Masai, Y. Sugiura, M. Ogata, K. Kunze, M. Inami, and M. Sugimoto. Facial Expression Recognition in Daily Life by Embedded Photo Reflective Sensors on Smart Eyewear. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, IUI '16, pages 317–326, New York, NY, USA, 2016. ACM.

[61] K. Masai, Y. Sugiura, M. Ogata, K. Suzuki, F. Nakamura, S. Shimamura, K. Kunze, M. Inami, and M. Sugimoto. AffectiveWear: Toward Recognizing Facial Expression. In *ACM SIGGRAPH 2015 Emerging Technologies*, SIGGRAPH '15, pages 4:1–4:1, New York, NY, USA, 2015. ACM.

[62] K. Masai, Y. Sugiura, and M. Sugimoto. FaceRubbing: Input Technique by Rubbing Face Using Optical Sensors on Smart Eyewear for Facial Expression Recognition. In *Proceedings of the 9th Augmented Human International Conference*, AH '18, pages 23:1–23:5, New York, NY, USA, 2018. ACM.

[63] D. Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 16(4):363–368, 1992.

[64] D. McDuff, S. Gontarek, and R. W. Picard. Remote measurement of cognitive stress via heart rate variability. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2957–2960, aug 2014.

[65] A. Mehrabian. *Nonverbal communication.* Transaction Publishers, 1972.

[66] H. Nakamura and H. Miyashita. Control of Augmented Reality Information Volume by Glabellar Fader. In *Proceedings of the 1st Augmented Human International Conference*, AH '10, pages 20:1—-20:3, New York, NY, USA, 2010. ACM.

[67] N. Nakazato, S. Yoshida, S. Sakurai, T. Narumi, T. Tanikawa, and M. Hirose. Smart Face: Enhancing Creativity During Video Conferences Using Real-time Facial Deformation. In *Proceedings of the 17th ACM Conference on Computer*

Supported Cooperative Work &#38; Social Computing, CSCW '14, pages 75–83, New York, NY, USA, 2014. ACM.

[68] M. Nicas and D. Best. A Study Quantifying the Hand-to-Face Contact Rate and Its Potential Application to Predicting Respiratory Tract Infection. *Journal of Occupational and Environmental Hygiene*, 5(6):347–352, 2008.

[69] M. Ogata and M. Imai. SkinWatch: Skin Gesture Interaction for Smart Watch. In *Proceedings of the 6th Augmented Human International Conference*, AH '15, pages 21–24, New York, NY, USA, 2015. ACM.

[70] M. Ogata, Y. Sugiura, Y. Makino, M. Inami, and M. Imai. SenSkin: Adapting Skin As a Soft Interface. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, pages 539–544, New York, NY, USA, 2013. ACM.

[71] M. Ogata, Y. Sugiura, H. Osawa, and M. Imai. iRing: Intelligent Ring Using Infrared Reflection. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, UIST '12, pages 131–136, New York, NY, USA, 2012. ACM.

[72] B. Parkinson. Do Facial Movements Express Emotions or Communicate Motives? *Personality and Social Psychology Review*, 9(4):278–311, 2005.

[73] M. Perusquía-Hernández, M. Hirokawa, and K. Suzuki. A Wearable Device for Fast and Subtle Spontaneous Smile Recognition. *IEEE Transactions on Affective Computing*, 8(4):522–533, oct 2017.

[74] P. Pham and J. Wang. Understanding Emotional Responses to Mobile Video Advertisements via Physiological Signal Sensing and Facial Expression Analysis. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*, IUI '17, pages 67–78, New York, NY, USA, 2017. ACM.

[75] R. W. Picard. *Affective Computing.* MIT Press, Cambridge, MA, USA, 1997.

[76] R. W. Picard and J. Healey. Affective wearables. *Personal Technologies*, 1(4):231–240, 1997.

[77] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[78] G. Rhodes, A. Calder, M. Johnson, J. V. Haxby, J. W. Tanaka, and I. Gordon. Features, configuration, and holistic face processing.

[79] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca. Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification. *IEEE Transactions on Cybernetics*, pages 1–11, 2018.

[80] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[81] J. A. Russell. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, 115:102–141, 1994.

[82] S. Salvador and P. Chan. Toward Accurate Dynamic Time Warping in Linear Time and Space. *Intelligent Data Analysis*, 11(5):561–580, oct 2007.

[83] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, June 2015.

[84] J. Scheirer, R. Fernandez, and R. W. Picard. Expression Glasses: A Wearable Device for Facial Expression Recognition. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '99, pages 262–263, New York, NY, USA, 1999. ACM.

[85] S. Schneegass and A. Voit. GestureSleeve: Using Touch Sensitive Fabrics for Gestural Input on the Forearm for Controlling Smartwatches. In *Proceedings*

of the 2016 ACM International Symposium on Wearable Computers, ISWC '16, pages 108–115, New York, NY, USA, 2016. ACM.

[86] M. Serrano, B. M. Ens, and P. P. Irani. Exploring the Use of Hand-to-face Input for Interacting with Head-worn Displays. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 3181–3190, New York, NY, USA, 2014. ACM.

[87] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vision Comput.*, 27:803–816, 2009.

[88] D. Smilek, J. S. A. Carriere, and J. A. Cheyne. Out of Mind, Out of Sight. *Psychological Science*, 21(6):786–789, 2010.

[89] O. Špakov and P Majaranta. Enhanced Gaze Interaction Using Simple Head Gestures. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 705–710, New York, NY, USA, 2012. ACM.

[90] V. Surakka, M. Illi, and P. Isokoski. Gazing and Frowning As a New Human–computer Interaction Technique. *ACM Trans. Appl. Percept.*, 1(1):40–56, jul 2004.

[91] K. Suzuki, M. Yokoyama, S. Yoshida, T. Mochizuki, T. Yamada, T. Narumi, T. Tanikawa, and M. Hirose. FaceShare: Mirroring with Pseudo-Smile Enriches Video Chat Communications. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 5313–5317, New York, NY, USA, 2017. ACM.

[92] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, jun 2014.

[93] G. A. ten Holt, M. J. T. Reinders, and E. A. Hendriks. Multi-dimensional dynamic time warping for gesture recognition. In *Thirteenth annual conference of the Advanced School for Computing and Imaging*, volume 300, 2007.

[94] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2016.

[95] Y. Tian, T. Kanade, and J. F. Cohn. Recognizing Action Units for Facial Expression Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, feb 2001.

[96] S. Tomkins. *Affect imagery consciousness: Volume I: The positive affects.* Springer publishing company, 1962.

[97] M. Tonsen, J. Steil, Y. Sugano, and A. Bulling. InvisibleEye: Mobile Eye Tracking Using Multiple Low-Resolution Cameras and Learning-Based Gaze Estimation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):106:1–106:21, sep 2017.

[98] H. Tsujita and J. Rekimoto. Smiling Makes Us Happier: Enhancing Positive Mood and Communication with Smile-encouraging Digital Appliances. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 1–10, New York, NY, USA, 2011. ACM.

[99] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: An addition to the mmi facial expression database. *Proc. Int'l Conf. Language Resources and Evaluation, Workshop EMOTION*, pages 65–70, 2010.

[100] K. Vega, M. Cunha, and H. Fuks. Hairware: The Conscious Use of Unconscious Auto-contact Behaviors. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 78–86, New York, NY, USA, 2015. ACM.

[101] G. Vettigli. MiniSom: minimalistic and NumPy-based implementation of the Self Organizing Map.

[102] S. Wan and J. K. Aggarwal. Spontaneous facial expression recognition: A robust metric learning approach. *Pattern Recognition*, 47(5):1859–1868, 2014.

[103] J. Whitehill, M. Bartlett, and J. Movellan. Automatic facial expression recognition for intelligent tutoring systems. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, jun 2008.

[104] K. Yamashita, T. Kikuchi, K. Masai, M. Sugimoto, B. H. Thomas, and Y. Sugiura. CheekInput: Turning Your Cheek into an Input Surface by Embedded Optical Sensors on a Head-mounted Display. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, VRST '17, pages 19:1–19:8, New York, NY, USA, 2017. ACM.

[105] K. P. Yao, W. H. Lin, C. Y. Fang, J. M. Wang, S. L. Chang, and S. W. Chen. Real-Time Vision-Based Driver Drowsiness/Fatigue Detection System. In *2010 IEEE 71st Vehicular Technology Conference*, pages 1–5, may 2010.

[106] S. Yoshida, T. Tanikawa, S. Sakurai, M. Hirose, and T. Narumi. Manipulation of an Emotional Experience by Real-time Deformed Facial Feedback. In *Proceedings of the 4th Augmented Human International Conference*, AH '13, pages 35–42, New York, NY, USA, 2013. ACM.

[107] S. Yoshimura, W. Sato, S. Uono, and M. Toichi. Impaired Overt Facial Mimicry in Response to Dynamic Facial Expressions in High-Functioning Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, 45(5):1318–1328, 2015.

[108] R. Zhao, Q. Gan, S. Wang, and Q. Ji. Facial Expression Intensity Estimation Using Ordinal Information. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3466–3474, jun 2016.

[109] F. Zhou, F. De la Torre, and J. F. Cohn. Unsupervised discovery of facial events. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2574–2581, jun 2010.