Embedded Facial Surface Sensing and Stimulation: Toward Facial Surface Interaction in Virtual Environment

February 2022

Fumihiko Nakamura

A Thesis for the Degree of Ph.D. in Engineering

Embedded Facial Surface Sensing and Stimulation: Toward Facial Surface Interaction in Virtual Environment

February 2022

Graduate School of Science and Technology
Keio University

Fumihiko Nakamura

Abstract

Since the emergence of interactive computers, a large number of human-computer interaction methods have been developed. Generally, humans interact with computers by inputting to and receiving feedback from computers. A typical computer has a keyboard, mouse, display, and speakers as input/output devices. With the development of computers, body gesture-based input methods have been introduced in consumer devices, like smartphones. The display technology has also developed as well as the input technology, thereby leading to the emergence of consumer head-mounted displays (HMDs). HMDs occlude the user's eyes with the display, and, thus, the conventional interfaces are difficult to use for VR users. In addition, in terms of the feedback, appropriate integration of multi-modalities improves the virtual experience. Most HMDs have high-quality visual and audio systems, but their haptic modality is not rich. To enable suitable input and rich interaction, the body is employed as an interface. In particular, the face can be controlled in minutes detail and is one of the most sensitive parts of the entire body. Therefore, the face is useful for interaction in a virtual environment. This dissertation presents two systems that are required for face-based interaction; one is the mouth shape recognition using mouth shape recognition embedded into an HMD; the other is the spatial directional guidance technique using robotic arms attached to an HMD.

First, an embedded optical sensor-based mouth shape recognition technique is proposed. This technique classifies mouth shapes into six classes using optical sensors embedded in an HMD. Moreover, this technique automatically gives labels to the training dataset by vowel recognition. In the experiments, with five participants, the classification accuracy of our method were compared for six mouth shapes in manual and automated labeling conditions. The results reveal that our method achieves an average classification accuracy of 99.9% and 96.3% under the manual and automated labeling conditions, respectively. These findings indicate that automated labeling is competitive relative to manual labeling, although the classification accuracy of the former is slightly higher than that of the latter. Furthermore, using the mouth shape recognition technique, a mouth expression transfer application was

also developed. This application blends six mouth shapes and then applies the blended mouth shapes to avatars.

Thereafter, Virtual Whiskers, a spatial directional guidance technique by cheek haptic stimulation using tiny robot arms attached to an HMD is presented. The tip of the robotic arm has photo reflective sensors to detect the distance between the tip and the cheek surface. The robot arms stimulate a point on the cheek obtained by calculating an intersection between the cheek surface and the target direction. In the directional guidance experiment, it is investigated how accurately participants identify the target direction provided by our guidance method. The difference between the actual target direction and the direction pointed by the participant was evaluated. The experimental result reveals that our method achieves the average absolute directional error of 2.54 degrees in the azimuthal plane and 6.54 degrees in the elevation plane. Moreover, a spatial guidance experiment to evaluate task performance in a target search task were conducted. In the evaluation, task completion time, system usability scale (SUS) score, and NASA-TLX score were compared in three conditions-visual, visual+audio, and visual+haptic conditions. The averages of task completion time were M=6.39 s, SD=3.34 s in the visual condition; M=5.62 s, SD=3.12 s in the visual+audio condition; and M=4.35 s, SD=2.26 s, in the visual+haptic condition. The SUS score was M=55.83, SD=20.40 in the visual condition; M=47.78, SD=20.09 in the visual+audio condition; and M=80.42, SD=10.99 in the visual+haptic condition. The NASA-TLX score was M=75.81, SD=16.89 in the visual condition; M=67.57, SD=14.96 in the visual+audio condition; and M=38.83, SD=18.52 in the visual+haptic condition. Statistical tests revealed significant differences in task completion time, SUS score, and NASA-TLX score between the visual and the visual+haptic conditions and the visual+audio and the visual+haptic conditions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Body as a User Interface

With the incredible evolution of electronic computers, they have become essential for our activities in human society. In the course of their development, there has been significant change in their operation methods. In operating computers, we input instructions into the system with our body and receive feedback triggered by our actions. During the operation, we access user interfaces, such as a keyboard, mouse, headphone, and display. For example, we use the touchpad on a laptop computer to click an application icon and see the display to receive visual feedback. In other words, it is essential for computer interaction to sense input from users and present stimuli according to the input, which is essential for human-computer interaction. Since the invention of the mouse and Sketchpad [2], various types of user interfaces have been developed. For a long time, computer systems have adopted characters and graphics in the form of character user interface and graphic user interface. These interfaces enable users to manipulate computer systems with metaphors that abstract instructions, while users have to memorize how to use various devices, like keyboards. However, the development of sensing and display technology enables intuitive interaction such as touches and gestures. Such techniques have been introduced in our daily life. For example, a smartphone, a portable device with a telephone and computing functions, has a touch interface. In addition, tangible user interfaces have emerged [3]. The tangible user interface enables interaction with digital world with physical objects. Moreover, there is an attempt to use our own body as an input/output interface, which is called on-body interaction. Such interfaces bring us closer to computers by merging operating and perceptional space, which makes the manipulation intuitive. In particular, since on-body interaction requires our body as an interface, it required less effort to learn how to use the device on users than physical instruments.

Advances in computing have driven virtual reality (VR) as well as the user interface. VR offers artificially reproduced stimulation, such as vision, audio, haptic, smell, and taste, to users. In VR, head-mounted displays (HMDs) are popular and have been developed since The Sword of Damocles [4]. The HMD displays computer graphics according to the head posture through the displays located in front of each eye to immerse users in a virtual environment. Moreover, most HMDs have audio interfaces and a pair of handheld controllers to integrate audio-visual-haptic modalities since multi-modal stimulation makes for better immersion. The advent of a low-cost and VR-ready HMD released by Oculus Rift [5] has made it easier for individuals to have HMDs. After the emergence of Oculus Rift, more VR applications have appeared than ever before and the market has expanded rapidly. Moreover, a metaverse has fascinated people since the late 2010s. The metaverse includes technologies and services that enable social interaction in the virtual world. In particular, owing to the widespread use of HMDs and the advances of network technologies, there has been remarkable improvement in the quality of virtual experience, which boosts the development of the metaverse. The metaverse is expected to expand interactions and business in our society and numerous companies have begun to invest substantial amounts of money in this field. Facebook, a famous social networking service company, has invested heavily in the metaverse, and has changed its name to Meta [*1]. Further, one of the most popular VR applications is VRChat [6]. In VRChat, users can interact with objects and even communicate with others via an avatar in a virtual environment. In addition, due to the COVID-19 pandemic, there has been a strong impetus to the use of VR, as it is deemed to be rather helpful in supporting social activities. However, VR systems have difficulty with regard to user interaction. It is difficult for HMD users to see the physical environment due to occlusion by the display. Most HMD-based systems adopt handheld controllers to manipulate the virtual environment, while the controllers limit the interaction because most activities need our hands. Therefore, VR needs hands-free and intuitive interfaces. On-body interaction is a solution for this, which allows the users to input/output through the body.

Although on-body interaction enables users to understand operation intuitively, the appropriate integration of multi-modality is important to convey accurate information. In on-body interaction, the body is used as an input/output interface. The hands are one of the most sensitive parts to mechanical stimuli and have a high degree of freedom. Therefore, since the early days of VR, many hand-worn devices have been developed. Such devices detect hand movements [7–11] and provide haptic feedback [12–15]. Apart from the hand, the torso has also been used for interaction. Suit-type devices are often leveraged to track body movements and stimulate a specific point on the torso. However, the information from the environment is encoded into the head-centered frames [16], which causes a shift in the perceived stimulation position. To avoid such issues, the face is employed as the other highly sensitive and highly

[*1] https://about.fb.com/news/2021/10/facebook-company-is-now-meta/

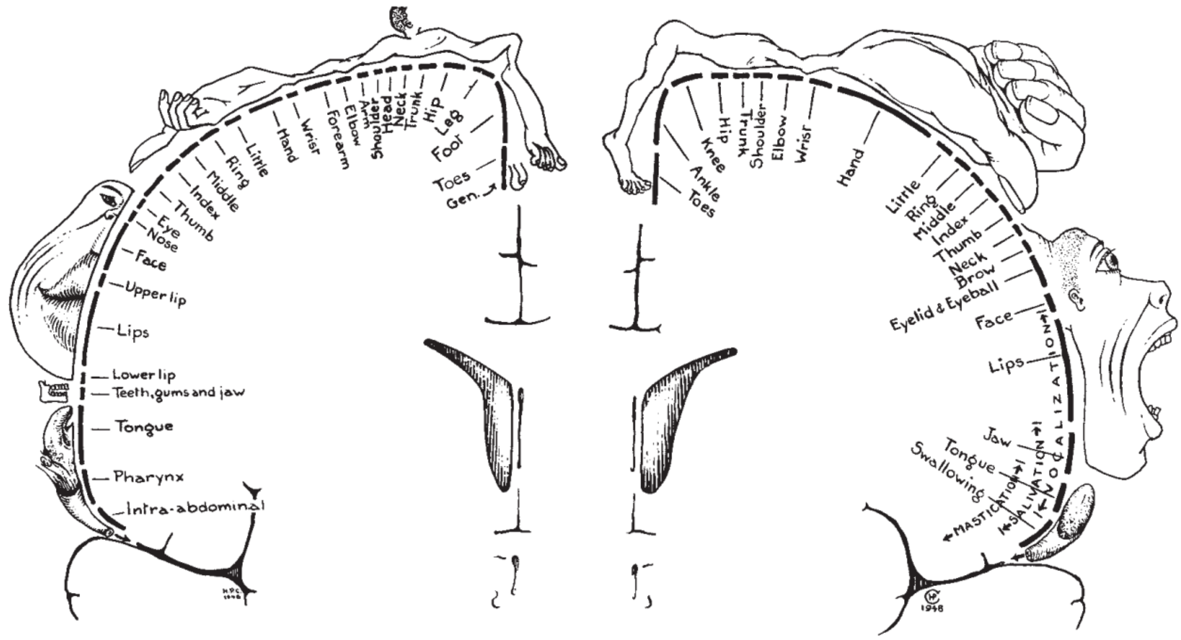


Figure 1.1: Cortical Homunculus [1]. (Left) Cortical Sensory Homunculus. (Right) Cortical Motor Homunculus.

controllable part, as indicated in Penfield's Cortical Homunculus (Fig. 1.1).

The face is a rather important channel for input/output in human-computer interaction. The face includes much information—such as sex, age, and nationality—to identify individuals. In addition, Penfield's cortical homunculus indicates that the face occupies a large part of the sensory and motor brain areas. We can identify the stimulation position precisely within the 2.0mm-4.0mm error. In the motor function, the face is activated by 20 kinds of facial muscles and four kinds of mastication muscles to convey complex information through facial expressions. Face recognition technology has been explored for many years. The most popular approach is the computer vision-based approach. In particular, the great advance of machine learning caused by deep learning has powerfully driven the image-based face recognition technique. Subsequently, face recognition systems were introduced at airports and the 2020 Summer Olympics. However, the face is susceptible to damage because most sensory organs-such as the eyes, lips, and ears-are located in specific positions on the face, and the facial skin is sensitive to stimuli. In particular, the skin of the lips is rather thin, thereby making them vulnerable to even small force. In addition, owing to the large number of muscles distributed on the head, the face can be formed with complex and fine geometry. Therefore, to exploit the potential of the face as an interface, it is important to detect the facial surface and stimulate it in a safe manner.

To interact with a facial surface in an immersive environment, trials were conducted to recognize facial expressions and provide feedback on the face [17, 18]. In terms of sensing facial movements, embedded cameras [19–21] and contact sensors [22, 23] were leveraged. When we move the face, the facial muscles are activated, the facial geometry deforms, and the facial appearance changes. In embedded camera-based approaches, the appearance of specific facial parts changes, such as the eyes and the mouth, and these are captured with color or infrared information to recognize facial expressions and reconstruct facial geometry. In particular, recently, image sensors were integrated with high-performance HMDs to obtain biometric information, such as gaze and mouth movements [24–26]. The camera-based methods enable robust sensing, while they require high computing power due to the processing of much information contained in the images. As one of the approaches with a low processing cost, contact sensors-such as electromyography (EMG) sensors and strain gauges-were adopted. The contact sensors were arranged around specific parts, like the eyes, to detect facial muscle movements or geometrical changes. In addition, the EMG-based sensing attachment for the HMDs was released [27]. The EMG-based sensing methods indicated that even low-dimensional data from the contact sensors was sufficient to reconstruct facial geometry. However, the sensing performance is affected by the contact states, and the contact sensors encounter difficulty in long-term use. In terms of haptic feedback on the face, haptic actuators were attached to certain portions of the HMD, like the bottom side, the front side, and a facial interface [28–31]. Moreover, to enable tangible interaction on the face, Tseng et al. installed several widgets on the front side [32]. Such attempts translated cues from virtual environments, like spatial information, into vibrotactile or thermal stimulation. However, vibrotactile stimulation conveys little information, and the tactors were arranged in a sparse layout. In addition, the tangible systems provided the stimulation indirectly through the HMDs and were not able to provide localizable stimulation. In short, the sensing methods encounter difficulty in computational cost costs due to the rich information of images and the contact states; on the other hand, the stimulation techniques have limited presentable haptic information due to the device configuration and the sparse layout of the actuators.

1.2 Goal

This dissertation presents facial surface sensing and stimulation techniques with embedded approaches to enable facial surface interaction in a virtual environment. Facial surface interaction requires recognition of targets and provision of stimulation based on the recognized states. This dissertation targets the bottom portion of the face, the mouth and the cheek, as Penfield's cortical sensory and motor homunculus indicates that the mouth and the cheek occupy a wide area among all the facial parts. To detect the facial skin surface, multiple proximity sensors are attached to an HMD. As the proximity sensor, optical sensors are employed. The optical sensors detect the distance between the sensors and the skin surface by measuring the light intensity reflected from the target. The sensors measure the skin surface around the mouth. The sensor data is low-dimensional, thereby enabling it to be processed quickly. In addition, as the optical sensors are unaffected by the contact states, it is considered that the optical sensors are suitable for long-term use. To stimulate the facial surface, robotic arms are integrated into an HMD. The robotic arms can move to a specific position so that they stimulate localizable stimulation on the cheek. This dissertation proposes two systems related to the facial surface-one is a system to recognize mouth shape with optical sensors; the other is a spatial directional guidance using cheek haptics.

In mouth shape recognition, the system detects the mouth shape of the user, which is the deformation of the skin surface around the mouth, with embedded optical sensors. Two types of optical sensors, namely, photoreflectors and position sensitive detector (PSD) distance sensors, are attached on the bottom of the HMD to measure the deformation around the mouth. By attaching the optical sensors to an HMD, a prototype is created to detect the mouth shape. A machine learning approach is used to recognize mouth shapes. A mouth shape classifier is constructed by learning sensor data with a support vector machine (SVM). In addition, the procedure of collecting training data is automated by recognizing vowels from the user's voice and giving them as labels to sensor data. With our system, an application is built to transfer the user's various mouth shapes on avatars by blending several templates of the mouth shape.

In the spatial directional guidance system, spatial directional cues are presented on the cheek using robot arms integrated into an HMD. An HMD-based facial haptics system that offers stimulation to the cheek is developed. The device consists of two robotic arms attached to the bottom of an HMD with proximity sensors. In advance of stimulation, the cheek surface is estimated using proximity sensors attached to the end effector of the robotic arm. Based on the estimated cheek surface, spatial directional information is encoded into a point on the cheek and touch it with the robotic arms to provide directional cues.

1.3 Organization

This dissertation investigate the HMD-based facial surface sensing and stimulation technique for facial surface interaction. Chapter 2 reviews related work to clarify the position of this dissertation. The chapter discusses interaction techniques using various body parts, including a face and user interface for VR systems. Chapter 3 introduces the facial surface interaction. Chapter 3 show the effectiveness of the facial surface as an interface in virtual environments. Chapters 4 and 5 describe our research projects. Chapter 4 presents the mouth shape recognition techniques [33]. Chapter 5 presents the spatial directional guidance technique using cheek haptics [34]. Chapter 6 concludes this dissertation.

Chapter 2

Related Work

This chapter reviews interaction techniques using various body parts and VR interfaces. The first section describes methods to sense and stimulate the body and applications using body interaction (Section 2.1). The second section (Section 2.2) focuses on face-related interaction technologies. The above two sections focus mainly on the interfaces in the physical environment. The third section (Section 2.3) presents the interactive systems for VR. The last section (Section 2.4) discusses the standpoint of this dissertation.

2.1 Body-based Interaction

Numerous researchers have attempted to leverage the body for interaction. Measuring and stimulating body parts, such as hands, arms, feet, legs, and a torso, enables interaction. This section reviews the measurement and stimulation methods, respectively.

2.1.1 Sensing Body

Basically, the actions of various body parts are often leveraged as an input method. For body sensing, camera-based approaches are popular for detecting body pose estimation [35], activity recognition [36], and hand gesture recognition [37]. In particular, the emergence of low-cost depth cameras [38–41] has driven image-based human sensing techniques [42, 43]. With the advance of camera technology, cameras have become smaller and higher resolution, thereby making them wearable. In several studies, wearable cameras were placed on a user's chest to recognize the user's bodily gestures [44] and capture the user's pose and motion [45]. Lin et al. mounted cameras on the back of a hand to capture hand gestures [46]. The image-based method enables accurate and robust body sensing, while it has difficulties in terms of limited field of view and high processing cost.

To avoid such issues, embedded sensor-based approaches have been proposed. In the embedded sensor-based approach, wearable devices with integrated sensors are installed on the body and are able to detect various activities of the user. The embedded sensing approach has frequently been adopted for the recognition of hand gestures and various sensors have been used, for example, myoelectric (EMG) sensors [7], strain sensors [47], bend sensors [10], optical sensors [8, 48], capacitive sensors [49], ultrasonic sensors [9, 50], and inertial measurement units (IMU) [51, 52]. In other body parts, there were attempts of pressure sensor-based foot gesture recognition [53], IMU-based gait classification [54], etc. Such embedded sensors have the advantages of less spatial limitation and low dimensional data, which leads to the lower computational intensity and battery power. Moreover, owing to the tremendous development in machine learning, even a sparse layout of stretch sensors was sufficient to estimate a hand pose [11]. The embedded approach is effective in recognizing the activities of specific body parts quickly with low-dimensional data.

Not only body gestures but skin gestures have also been adopted for input. The skin gesture is a deformation caused by the user's action, such as pinching, pushing, and stretching the skin. Since skin gestures are smaller than body gestures, it is difficult for cameras to capture the deformation. Therefore, typically, sensitive sensors have been deployed. Harrison et al. proposed Skinput, an armband-type device to detect touch with acoustic signals on the skin [55]. In their study, they detected touch on the forearm using acoustic signals and integrated the sensing method into an on-body projection system. Furthermore, Weigel et al. found that, in terms of skin input for mobile computers, on-skin gestures extended the conventional touch interfaces, and users are likely to use the forearm and hands [56]. Ogata et al. embedded photo reflective sensors into wearable devices to recognize finger gestures [57] and forearm skin deformation [58]. As in other methods, capacitive sensing [59] and acoustic sensing [60] were also attempted. The on-skin input detects subtle skin deformation with various modalities. In particular, the optical sensors measure the distance between the sensor and skin in order to capture spatial geometrical changes. Therefore, optical sensors are suitable for recognizing facial geometry.

2.1.2 Stimulating Body

In stimulating the body, a variety of haptic stimulation was provided to several body parts. Hands, which are sensitive to mechanical stimuli, are often employed for stimulation. To stimulate the hands, hand-worn or handheld devices were developed. Günther et al. developed a glove-type device that provided spatial information encoded with vibration patterns [13]. Chen et al. proposed a pin array-based handheld device to present spatial directional cues on the user's palm [12]. These studies attempted a stimulation directly on the body, while ambient stimulation, such as air jets [14] and ultrasonic cues [15], was also employed. Ion et al. presented a tactile display that dragged the skin [61]. In stimulating the torso, a jacket-type force feedback device was proposed. To present stimuli to the torso, suit-type devices are often used, including jacket-type devices that present force sensation [62] and commercial haptic suits [63]. In addition, stimulation of the feet has also been explored by rendering the tactile sensation [64]. Matsuda et al. leveraged the neck as a haptic display to present spatial awareness by providing tactile patterns on the neck [65].

2.2 Interaction using Head

Face can be formed in complex geometry and is one of the most sensitive to mechanical stimuli. Therefore, numerous studies have attempted to recognize facial expressions and provide haptic feedback for human-computer interaction [66–69]. Here, previous research on sensing and stimulating faces is discussed.

2.2.1 Facial Expression Recognition

Numerous facial recognition systems, as well as body sensing systems have been developed. In human-computer interaction, the face is used as an input modality. Computer vision-based techniques have been studied for a long time. In facial expression recognition, most studies detected descriptors of facial expressions [70]. Since the emergence of convolutional neural network (CNNs), deep learning has been used for facial expression recognition [71, 72]. Several studies have recognized facial movements from 2D images to leverage them as input [66, 67]. Deepateep et al. recognized 3D facial movements as input for mobile devices [73]. Yan et al. detected a frown for interrupting unexpected action from smart speakers [74].

Furthermore, wearable devices have been deployed for sensing facial expressions. In a remarkable method, Masai et al. embedded multiple facial expressions into glasses to recognize facial expressions [75]. They revealed that by arranging photoreflectors in a sparse layout considering Facial Action Coding System (FACS), high facial expression classification could be achieved. In other studies using photoreflectors, Masai et al. detected facial skin movements caused by rubbing the face for input [76]; Kikuchi et al. recognized ear gestures with earphones integrated photoreflectors [77]; Hashimoto et al. detected tongue gestures with a mouth piece embedded with several photoreflectors [78]. As with other sensing methods, capacitive sensing is also leveraged by integrating electrodes into glasses [79] and a mask [80]. Goel et al. proposed a tongue gesture recognition method with head-mounted microwave motion sensors [81]. Xu et al. presented an on-face interaction method by detecting gestures on the face using acoustic signals [82].

This section mentions several sensing methods used for facial expression recognition.

However, to capture the surface statically, optical sensing is most appropriate, as the other methods focus on temporal changes.

2.2.2 Facial Stimulation

The face has several parts, and each parts have different levels of sensitivity. In facial stimulation, haptic stimulation is designed to safely convey appropriate information. Several studies translated surrounding information into haptic stimulation around the head [83–85]. A forehead has a wide surface and is an area that is easy to provide stimulation to. Kajimoto et al. designed a forehead stimulation system with electro-tactile cues to convey spatial information [69]. The cheek also has a large surface and is more sensitive than other facial parts. Sato et al. designed an interactive cheek haptic display using an air vortex to modify the user's stress [86] and found that air vortex-based cheek haptics affected task performance and physiological responses [87]. Yoshida et al. developed a glasses-type wearable device that released water on the cheek to increase sadness [88] Gil et al. investigated the perception of ultrasonic cues on the face and found that the cheek could detect the location of the stimulus as precisely as the glabella and above the eyebrows [89]. Hashimoto et al. provided vibrotactile feedback on the cheek by modulating hand gestures to waveform [68]. In addition, ears have haptic receptors to detect haptic stimulation. Lee et al. assessed the possibilities of ear haptics in conveying spatial-temporal information [90]. Nasser et al. investigated thermal feedback to the skin around the ears [91]. Not only the facial surface but also the oral cavity is leveraged as a haptic display [92]. Most facial stimulation methods used weak stimulation, such as ultrasonic, vibration, and thermal as the face is easily injured.

2.3 User Interface for VR

In VR systems, we interact with virtual environments through devices. For example, an HMD shows virtual environments according to a head position and posture. In such a case, it is important to sense users' actions and provide feedback to users. This section discusses the user performance sensing techniques and the methods for providing haptic stimulation.

2.3.1 User Performance Sensing

This section focuses on facial sensing. Users wear VR devices to immerse themselves in virtual environments. The significant difference from sensing in the physical environment is that the display occludes the user's face. Several studies have attempted to capture the entire body with cameras [93]. However, the sensing techniques for VR are not different from those for the physical environment.

Facial Sensing of HMD Users

Immersive HMDs occlude most of the face by the displays, thereby making it difficult to recognize facial expressions with conventional camera-based methods. To overcome these issues, many studies have proposed facial sensing systems.

Although most of the HMD user's face is hidden by the display, the mouth is exposed. Several studies have attempted to recognize the mouth to detect mouth gestures [94] and whole facial expressions [95]. However, it is difficult to recognize entire facial expressions due to a lack of information around the eyes. To obtain such lacking information, Li et al. attached multiple strain gauges on the facial interface and an RGB-D camera [22] They attempted to reconstruct facial geometry by combining depth images of the mouth and skin deformation caused by forming facial expressions. In addition, there were approaches to embed tiny cameras into HMDs for capturing eye images and recognizing facial expressions. Hickson et al. infered facial expressions by using eye images of HMD users [19]. Several studies used several facial cameras to transfer facial expression on avatars [96] and generate a photo realistic face [97]. However, camera-based approaches entail high computational costs due to the rich information available from the images. This is critical to wearable HMDs because the HMDs spare many computational resources to process other functions.

For reducing the processing cost, embedded sensors have been employed. As the face is occluded by the HMD, by integrating sensors into the HMD, many studies have attempted to understand gestures. Since HMDs usually have inertial measurement units (IMUs) and microphones, IMU-based head gesture sensing methods [98] and acoustic-based hand gesture sensing on the HMD surface [99] were proposed. In the case of recognizing facial expressions, additional sensors were integrated into the HMD. Several studies placed acoustic sensors [50] and EMG sensors [27, 100–102] on the facial contact area. As other methods, optical sensors, like photoreflectors were leveraged to detect facial surfaces. Photoreflectors were placed to measure the deformation of a specific facial location. Nakamura et al. detected changes in the glabella with a photoreflector embedded into an HMD to control the volume of the displayed information [103]. Li et al. detected the skin movement on both sides of the face to predict continuous jaw motions [104]. Sakashita et al. placed a photoreflector array in front of the user's mouth to recognize mouth movements [105]. Yamashita et al. attached photoreflectors on the bottom of optical see-through HMD to detect cheek surface deformation [106]. Kim et al. embedded several couples of infrared emitter and receiver into a facial cushion of an HMD to detect facial gestures, including head movements [107]. By measuring multiple locations on the face, several studies attempted to capture the entire complex facial surface deformation. Suzuki et al. installed multiple photoreflectors into the interior of the HMD to recognize facial expressions [108]. As described above, the embedded sensor-based approaches achieved face-related movement recognition, including specific facial parts and the entire face. In terms of sensing the mouth, previous studies targeted simple jaw movements, while few studies measured complex mouth shapes, even though the mouth is a more expressive part.

2.3.2 Feedback in Virtual Environment

We receive haptic feedback through interaction. Haptic feedback is provided on various body parts. In particular, hands are popularly used for stimulation. Fang et al. rendered haptic feedback of virtual objects using a shoulder-mounted device that retracted finger-worn wires [109]. AI-Sada et al. leveraged wearable robotic arms to provide multiple haptic modalities with various haptic actuators attached to the end effector [110, 111]. They stroked the face with soft brushes to provide haptic feedback [110]. Hoppe et al. leveraged quadcopters to render haptic feedback [112]. In VR, it is essential to integrate haptic stimuli with other modalities in order to improve the virtual experience.

Facial Stimulation of HMD Users

The head is sensitive to mechanical stimuli, and the perception of haptic stimulation is likely to encode a head-centered frame [16]. Therefore, the head is effective in appropriately integrating haptic stimulation into other modalities. Since the HMDs have a contact area to the face, most studies involve the installation of haptic actuators there. Wang et al. presented a facial skin stretching system that installed shear actuators on an HMD's facial interface [113]. Oliveira et al. translated spatial direction into vibrotactile cues on an HMD's facial interface [28]. Kameoka et al. mapped the finger information to suction stimulation on the face [114]. Peiris et al. installed thermal actuators on the facial contact area to explore the ability of thermal haptic feedback to the forehead for spatial awareness [29] and enrich the virtual experience by providing thermal feedback on the face [30]. Chang et al. proposed a force feedback method by pushing an HMD to the face [115]. Another approach was to mount haptic actuators on the interior or the exterior of the HMDs. Although the interior of the HMD is small, there was an attempt to place an air jet-based tactile feedback device on the lens [116]. On the exterior of the HMD, relatively large actuators were attached. Tsai et al. presented a force to the head through devices, that generated impact force, attached to the front of the HMD [31]. Peng et al. provided unobtrusive tactile feedback on the back of the head during walking to reduce VR sickness [117]. Several studies have attempted to place haptic actuators on the bottom and both sides of the HMD, close to the mouth and cheek. Ranasinghe et al. attached wind and thermal actuators to the bottom of an HMD to enhance the virtual experience [118, 119]. Liu et al. synchronized visual oscillation caused by walking to cheek haptic stimulation to reduce VR sickness [120]. Wilberz et al. provided spatial directional cues with multiple modalities using a single robot arm mounted on the HMD [121]. Thus, it is evident that previous studies aimed at several localized positions for haptic feedback, while few studies attempted to stimulate arbitrary locations on the face.

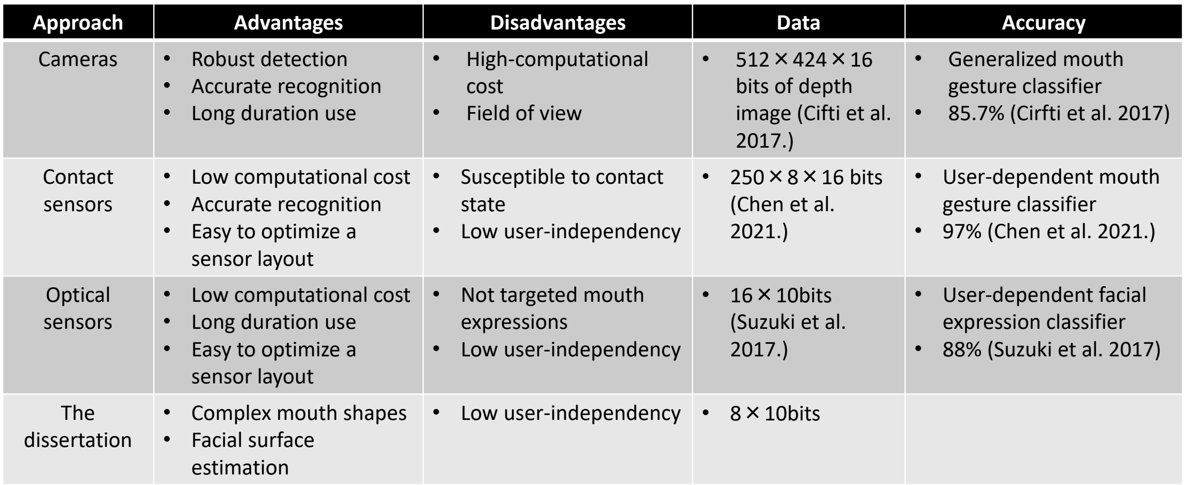
2.4 Standpoint

This section describes the position of this dissertation. The previous sections discuss interactive technologies in physical/virtual environments. Those technologies leveraged various modalities for both input and output.

In terms of the sensing, computer vision techniques were frequently introduced to interactive systems in stationary and wearable approaches, while the processing cost is high due to high-dimensional data (i.e., if the image is VGA, the data dimension is 640 pixels × 480 pixels × 3 channels = 921600 bytes). The previous study on camera-based mouth gesture classification [72], which classified six mouth shapes (silence and five Japanese vowels) using facial images of 1024 pixels (32 pixels × 32 pixels) on a mobile device, achieved an average classification accuracy of 92.4% in an average elapsed time of 2.33 seconds. The camera-based HMD user's facial expression recognition [94] used a high-performance computer to classify seven mouth gestures (neutral, mouth stretch, smile, dislike, lip puckered, left lip corner puller, right lip corner puller) from input facial depth images (512 pixels × 424 pixels × 16 bits) with an accuracy of 85.7% in 18 ms. To reduce the data for processing, there were contact sensor-based attempts to enable gesture recognition. In [101], with EMG sensors attached to the facial interface of HMD, seven mouth gestures were classified with 97% accuracy using data of 32000 bits (1 s × 250 Hz × 8 measures × 16 bits). Contact sensors are easily arranged according to the sensing target. However, in the contact sensor-based approaches, the sensing performance depends on the contact states. Therefore, contactless sensors, such as photo-reflective sensors, were integrated into wearable devices. Photo reflective sensors are also to optimize the sensor layout according to sensing locations. In particular, combining photo reflective sensor-based sensing and a simple machine learning approach, like SVM, achieved high facial expression classification accuracy in physical and even virtual environments [75, 108]. In the virtual environment, most embedded facial sensing technologies target the occluded portions, such as the eyes and the eyebrows. For mouth gesture sensing, there were several methods to capture jaw movements using embedded sensors [104] within an error of 8.19mm, while the complicated mouth shapes could not be detected.

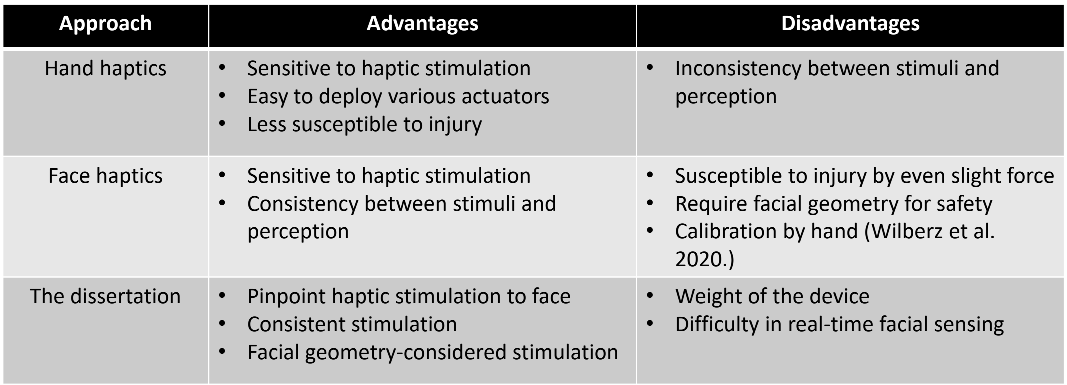
In the stimulation methods, vibrotactile methods are primarily introduced to embedded systems. Haptic stimulation translates various types of information. Penfield's cortical

Table 2.1: Standpoint in terms of sensing



Approach	Advantages	Disadvantages	Data	Accuracy
Cameras	• Robust detection • Accurate recognition • Long duration use	• High-computational cost • Field of view	• 512 × 424 × 16 bits of depth image (Cifti et al. 2017.)	• Generalized mouth gesture classifier • 85.7% (Cirfti et al. 2017)
Contact sensors	• Low computational cost • Accurate recognition • Easy to optimize a sensor layout	• Susceptible to contact state • Low user-independency	• 250 × 8 × 16 bits (Chen et al. 2021.)	• User-dependent mouth gesture classifier • 97% (Chen et al. 2021.)
Optical sensors	• Low computational cost • Long duration use • Easy to optimize a sensor layout	• Not targeted mouth expressions • Low user-independency	• 16 × 10bits (Suzuki et al. 2017.)	• User-dependent facial expression classifier • 88% (Suzuki et al. 2017)
The dissertation	• Complex mouth shapes • Facial surface estimation	• Low user-independency	• 8 × 10bits	

Table 2.2: Standpoint in terms of stimulation



Approach	Advantages	Disadvantages
Hand haptics	• Sensitive to haptic stimulation • Easy to deploy various actuators • Less susceptible to injury	• Inconsistency between stimuli and perception
Face haptics	• Sensitive to haptic stimulation • Consistency between stimuli and perception	• Susceptible to injury by even slight force • Require facial geometry for safety • Calibration by hand (Wilberz et al. 2020.)
The dissertation	• Pinpoint haptic stimulation to face • Consistent stimulation • Facial geometry-considered stimulation	• Weight of the device • Difficulty in real-time facial sensing

sensory homunculus indicates the wider area of the face so that the face can perceive haptic stimulation precisely. Moreover, the previous study on cheek stimulation [89] revealed the participants identified the location of ultrasonic cues on the bridge between the eyes, the eyebrows, and the cheek. In the case of VR, most studies embedded tactors into the facial interface of the HMD. Several studies provided feedback to the region around the mouth with winds and thermal cues [118, 119]. Most studies explored only ambient cues despite the excellent sensitivity of the face for haptic stimuli. Wilberz et al. provided directional cues using a robotic arm mounted on the HMD to improve the virtual experience [121]. However, in their study, the cheek surface was roughly detected by hand, was not accurately estimated, and an arbitrary position could not be calculated.

Since the rise of VR, many companies have invested heavily in VR. These companies targeted high-fidelity photo realistic rendering to transfer user on avatars [21] and high-quality haptic rendering to provide realistic haptic stimulation [122]. Such technologies significantly improve virtual experiences. However, for immersivity, it is not required to have photo realistic avatars and consistent stimuli. For example, Animaze by FaceRig [*1] recognizes user's facial movements from images to 2D and 3D avatars, including non-human avatars. However, image-based facial expression transfer applications process much information contained in images. Therefore, to capture facial features with low-dimensional data, photo reflective sensors were adopted for facial expression recognition [108]. Moreover, in [108], facial expressions could be reproduced by synthesizing several template facial expressions. In terms of haptic stimulation, by mapping hand haptic sensation to other body parts, virtual experience is improved. Therefore, even low-cost facial sensing and spatial inconsistency does not matter if input/output modalities are appropriate.

From the perspective of sensing and stimulation, it is considered to be important to detect complex facial expressions and stimulate the face with pinpointing direct cues for improving the virtual experience. This dissertation aims to build the embedded facial surface sensing and stimulation techniques with low dimensional data and localizable haptic sensation, which enables facial surface interaction.

[*1] https://store.steampowered.com/app/1364390/Animaze_by_FaceRig/?l=japanese

Chapter 3

Facial Surface Interaction in Virtual Environment

This chapter describes the core concept of facial surface interaction. In the advancement of facial surface interaction, skin surface interaction is described. Based on skin surface interaction, the differences between skin surface interaction and facial surface interaction are described to clarify the concept of this dissertation.

3.1 Skin Surface Interaction

We interact with physical environments through the skin. The skin deforms by our action and detects stimuli from the environment. The skin is always present on our body so that it is accessible anytime. Therefore, the concept of using skin surface as an interface, skin surface interaction, has been proposed. The skin surface enables a user to communicate with systems without external devices, such as a mouse, a keyboard, or an LCD display. Skin surface interaction contributes to a new concept of computers, like VR.

To realize skin surface interaction, there have been attempts to use changes on the skin as input and to use the skin as output. Especially, since Skinput presented by Harrison et al. [55], many skin-based input techniques have been developed [123]. Such techniques measured the changes on the skin in terms of spatial and temporal changes. To capture the changes, there were various sensing attempts using cameras [124,125], acoustic sensors [126], inertial sensors [127], ultrasonic sensors [128], and optical sensors [8,48,57,58,129,130]. Skin-based input focuses typically on localized skin because the skin surface moves in a simple manner.

For feedback to the skin, visual and haptic feedback has been adopted. Basically, to project visual stimuli, displays and projectors were employed. In such a case, the projected images are modified according to the target geometry. For example, Xiao et al. modeled the arm as a cone-like object to estimate the arm surface [130]. On the other hand, in terms of haptic feedback, skin geometry has not attracted too much attention. This is because most skin-based haptic feedback systems are directly mounted on target location and, if the systems are a stationary setup, the systems constrain the user's spatial location to provide feedback in an effective manner. In the case of ambient cues, such as ultrasonic, the target position is estimated to match stimuli to target [131], while the accurate skin position is usually not considered.

3.2 Facial Surface Interaction

The facial surface interaction focuses on a specified skin surface, while it is different from the conventional skin surface interaction. When the stimulation is provided to humans, the stimulation is likely to be encoded into head-centered reference frames [16]. Therefore, using the facial surface as an interface enables consistent interaction between the stimulus and the user's perception. Moreover, Penfield's cortical homunculus (Fig. 1.1) reveals that humans convey and receive much information through the face as well as hands. Targeting the face enriches interaction without using hands, which are important channels to sense and actuate the physical world. Moreover, by interacting with additional bodies and senses through the face, it is considered that humans augment their abilities in an effective manner. There were attempts to augment spatial awareness by presenting the forehead, which is a sensitive part of the entire body [69]. In other words, facial surface interaction has a potential to expand the range of interaction without preventing conventional hand-based interaction. Therefore, facial surface interaction contributes many research areas, such as telepresence, collaboration, affective computing, and human augmentation.

There were attempts to use facial expressions as input. Since humans form complex facial expressions owing to facial muscles, facial expressions enable a wide range of input. Masai et al. attempted facial gestures to apply to daily tasks, such as making a call, turning on a TV, and playing music [132]. Nakao et al. presented an EMG-based facial input technique that allows hands-free interaction [102]. Santis et al. used facial movements as input for disabled users [133]. In addition, facial expressions were leveraged to enable hands-free game control [134]. Mouth gestures have also been used for input [66, 67]. The above studies revealed that facial expressions were useful for human-computer interaction. To realize interaction using faces, it is essential to accurately recognize various facial expressions.

In stimulating the facial surface, skin surface estimation is rather important unlike skin surface interaction. In interaction, sensing and stimulation are important. However, the face should be provided stimulation with care because the face is easy to be injured. The head has many sensory systems, which are located on the face and are exposed. In addition, the skin around the eyes and lips is thin, thereby making it susceptible to injury. Therefore, several studies employed ambient stimulation, like wind [87]. The ambient cues provide the feedback

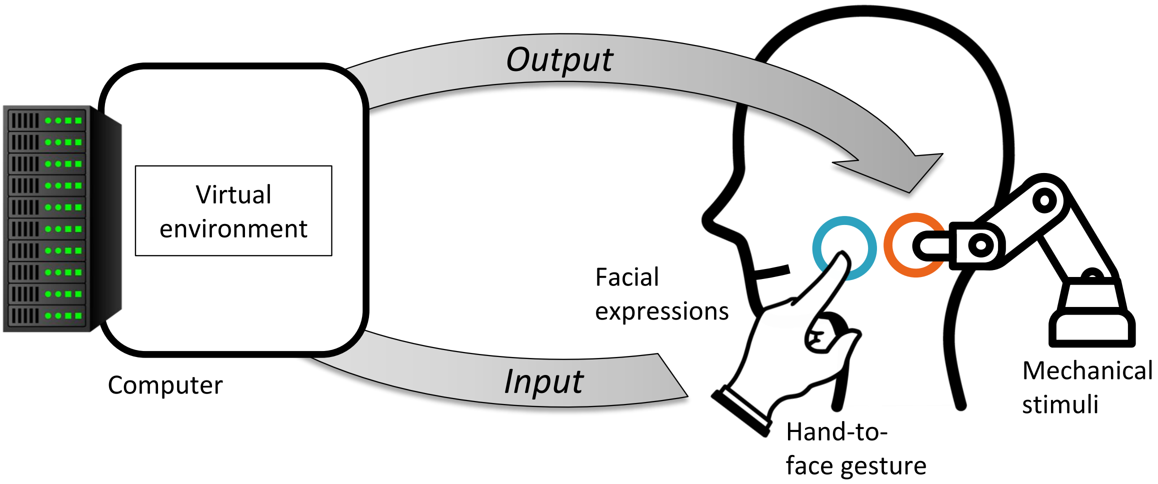


Figure 3.1: Facial Surface Interaction

in a safe manner, even without accurate skin surface information, while the ambient cues cannot stimulate to pinpoint despite the excellent sensitivity of the face. In another approach, vibrotactors were integrated into head-worn devices [84]. However, vibrotactile stimulation provides limited information. To exploit the potential of a face as an interface, it is essential to provide localizable stimulation. In such a case, direct contact is suitable, but it can injure the face. In addition, facial geometry and movements vary from person to person. The facial geometry can be controlled in a fine and complex manner, owing to the many facial muscles. Therefore, accurate surface information of each person is required to calibrate the mapping between the stimulation and surface.

In VR scenarios, facial surface interaction has been employed. Several studies used facial expression to control virtual third arms [17] and input to AR game [18]. Head gestures [98] and facial gestures [107] were also employed for interaction in VR environments. For haptic stimulation, haptic actuators, such as vibrotactors and thermal actuators, were adopted. Most HMD-based haptic feedback studies integrated haptic actuators on the contact area to the face [28–30, 113]. A jet-based tactile feedback device was embedded into the interior of an HMD [116]. In other approaches, haptic actuators were attached to the exterior of an HMD [28, 118–120]. Wilberz et al. leveraged robotic arms with soft tips on the end effector to present direct stimulation on the cheek [28]. However, the their study needed manual calibration. Especially, in the case of VR HMD users, the user cannot see the physical environment, and, thus, the other person is required for calibration. Therefore, it is important to automate the calibration of the mapping between the stimulation and surface, which allows users to easily interact with virtual environments.

Chapter 4

Mouth Shape Recognition with Embedded Optical Sensors

4.1 Introduction

VR systems allow their users to communicate via avatars in various scenarios. In such a case, the facial expression is essential because it conveys emotions and intentions, among others. Computer vision techniques can be used to capture facial expression in many cases; however, facial expressions under HMD cannot be easily captured due to facial occlusions caused by the display. Several solutions have been developed to overcome this issue. One remarkable method is using embedded optical sensors around the eye region and adopting machine learning techniques to recognize facial expressions [108]. However, given that the sensors are allocated around the eye region, the system recognizes only limited movements of the mouth. The mouth is important for understanding expressions [135], especially for Westerners.

This study introduces a system that recognizes the mouth shape of the HMD user with optical sensors. As the optical sensors, a photo-reflective sensor and a position sensitive detector (PSD) were adopted in this study. These sensors measure the distances between them and the skin surfaces surrounding the mouth. Optical sensor values are collected for each mouth shape and then labeled the sensor values using vowels, which are detected from speech. Classifiers are trained with these labeled sensor values. A prototype is built using the HMD to measure the mouth shapes. Also, an application was developed to transfer the user's mouth shape to an avatar. This application blends mouth shapes according to belonging probabilities for the classes. Figure 4.1 shows this application applying a blended mouth shape to an avatar.

Many studies on capturing facial performance use cameras or optical sensors. Camera-based techniques detect facial gestures accurately, but they have difficulty integrating to HMD-based systems because of limitations such as weight, hardware cost, and high computational cost. Meanwhile, optical sensors are lightweight, low-cost, and capable of recognizing gestures with low-dimensional data. Therefore, they are suitable for wearable devices, such as HMD. In particular, integrating machine learning with sensing with optical sensors enables powerful gesture recognition [75]. Many of such recognition requires collecting training data by hand, making the training process explicit and time-consuming. Therefore, labels are acquired by speech recognition to gather training data automatically. In interacting with VR systems, audio modality is often employed. If labels can be obtained from the audio modality, the training

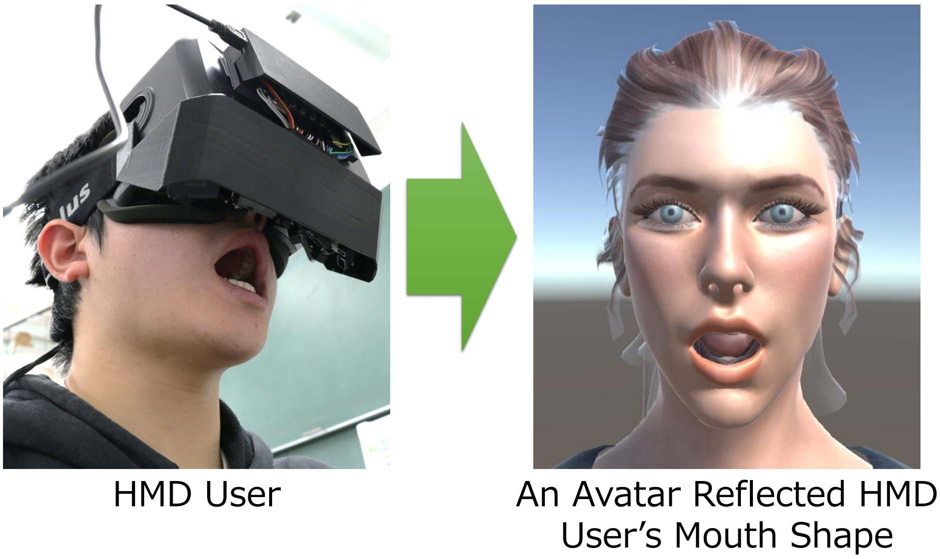


Figure 4.1: Reflecting Mouth Shape of HMD User to Avatar

process can be completed during interacting with the systems. Such an implicit training process leads to higher usability. This study investigates a speech-based sensor data labeling method to explore the possibility of automating the training process with audio information.

The main contribution of this study is as follows:

- This study developed a technique that recognizes mouth shapes while the user is wearing an HMD. A mouth shape sensing device was built to be lightweight, low-cost, and unaffected by the facial occlusion caused by HMD.
- The training data was automatically collected by speech recognition. Labels for training data was obtained by integrating vowel recognition with a mouth shape measurement technique.
- An application was built to project the user's various mouth shapes on avatars by synthesizing several bases of the mouth shape. The parameters of bases are blended according to belonging probabilities to reproduce multiple mouth shapes.

4.2 Related Work

This section reviews previous works on capturing facial performance and highlight wearable systems with embedded sensors. This section describes an overview of sensing approaches, such as audio, camera, contact sensors, and optical sensors.

4.2.1 Audio-based Approach

Speech has been used to produce an animation of lip movements. Speech consists of phonemes, the smallest unit of speech that makes sense as a language. A mouth shape and a tongue position determine a phoneme. A mouth shape corresponded to a lip position. Therefore, phoneme detection leads to estimation of lip movements [136]. Oculus Lipsync [137] recognizes the speech sounds of HMD wearers and matches the lip movements of avatars with the speech in real time. The audio-based approach is based on the speech, and thus does not work without any voice.

4.2.2 Camera-based Approach

One of the most popular approaches to capturing facial movement is the use of cameras. Focusing on HMD users' facial movement recognition, embedded camera approaches have been explored. For example, Hickson et al. [19] classified facial expressions from images of an eye camera embedded in HMD. Olszewski et al. [96] developed a system that reconstructs the facial geometry of the user using an HMD with both eye cameras and a mouth camera. However, this approach requires high computational power and expensive hardware because it targets high-fidelity avatars.

4.2.3 Contact-based Approach

Contact sensors can detect facial movements through muscles and skin deformation. For example, Gruebler et al. presented a wearable device to recognize positive facial expressions from electromyography [138]. Li et al. proposed a system that reconstructs facial geometry from both strain gauges and an RGB-D camera attached to an HMD [22].

Contact sensors are suitable for wearable devices because of their compactness and fulfillment of the required contact with surfaces. However, contact-based sensing techniques depend on the condition of the contact with a surface, and there are concerns about comfort while the device is mounted.

4.2.4 Measurement using Optical Sensors

Optical sensors have been deployed to wearable devices because of their capability to sense gestures. Some interfaces using optical sensors focus on the HMD wearer. Sakashita et al. developed a mask-type interface that transmits human action to puppetry [105]. Their system used optical sensors to detect the lower lip position and classifies three mouth states (closed, partly open, open). Suzuki et al. built an HMD-based system that recognizes five facial expressions of the HMD wearer [108]. Their system used machine learning to recognize various facial expressions. In their system, optical sensors detected the deformation around the eyes. However, this system has difficulties in measuring the mouth shape because it focuses on the eye region.

Combining machine learning to measuring by optical sensors enables the detection of various gestures [75] but requires a training process. To automate this process, Suzuki et al. requested individuals to imitate the facial expressions of avatars and collected training data [108]. However, their system can label training data incorrectly because users can make facial expressions that differ from those of avatars.

As mentioned above, previous studies have two limitations, namely, recognition methods of mouth shapes and labeling methods of training data. This study uses optical sensors to measure the mouth shape and machine learning to recognize various mouth shapes. In training, vowels are recognized to give correct labels to training data.

4.3 Mouth Shape Recognition by Embedded Optical Sensors in HMD

This section provides an overview of the mouth shape recognition system proposed in this study. The mouth shape recognition system consists of three techniques, namely, mouth shape classification, data labeling using vowel recognition, and mouth shape reproduction. Optical sensors detect the distances between them and skin surfaces. Optical sensor values are labeled with vowels recognized from speech and are learned to classify the mouth shapes. In reproducing the mouth shapes, multiple mouth shapes are blended according to the class membership probabilities. This reproduction approach is similar to that in a previous research [139]. Figure 4.2 shows the flow of mouth shape recognition by optical sensors and labeling of training data by vowel recognition. Section 4.3.1 describes the mouth shape measurement and the classification process. Section 4.3.2 describes the labeling technique of the training data using vowel recognition. Section 4.3.3 describes the blending method for the mouth shapes.

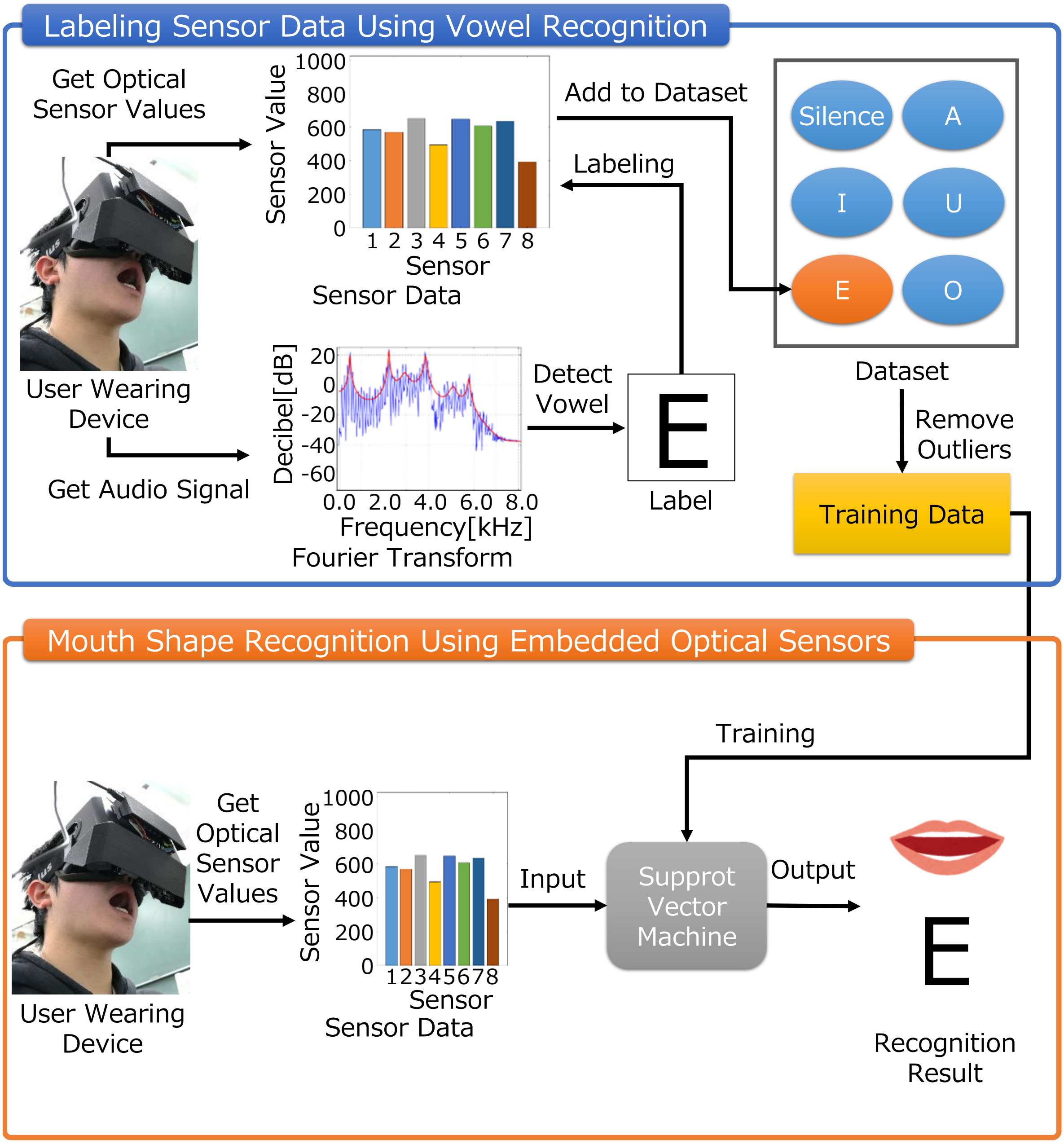


Figure 4.2: Mouth Shape Recognition using Optical Sensors and Labeling Method of Training Data using Vowel Recognition

4.3.1 Mouth Shape Classification by Embedded Optical Sensors

The mouth shape recognition system of this study measures mouth shapes by an approach similar to that in [75]. Embedded optical sensors are used to achieve an optimized sensor allocation and low computational cost. The skin deforms as the mouth muscles move. Optical sensors capture this deformation by detecting the distance between them and the skin surfaces. These distances are different for each mouth shape because the movement of mouth muscles varies depending on the mouth shape. Eight optical sensors are deployed to an HMD. The measurement points are the upper lip, upper cheek, lower lip, and cheek. These points are on both the left and right side.

This study adopts two kinds of optical sensors, namely, a photo reflective sensor and position sensitive detector (PSD), which differ in sensing target and measurable range. Photo reflective sensors detect the intensity of reflected light (Figure 4.3 Left), whereas PSDs detect the position where reflected light is received (Figure 4.3 Right). Most photo reflective sensors can measure from about 1 mm to 20 mm, while many PSDs can measure from approximately 10 mm to 200 mm. Hence, photo reflective sensors are suitable for measuring the upper lip and the upper cheek, which are relatively close to the HMD. By contrast, PSDs are suitable for measuring the lower lip and the cheek, which are relatively far from the HMD.

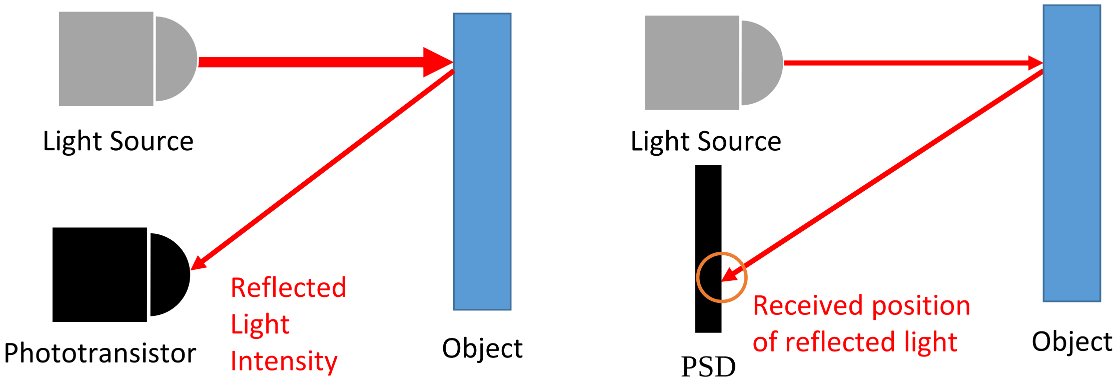


Figure 4.3: *Left*: Measurement Principle of Photo Reflective Sensor. Photo reflective sensors detect the light intensity reflected from the target object. *Right*: Measurement Principle of PSD. PSDs detect the position where the reflected light reaches.

This study applies machine learning to recognize mouth shapes. Our system uses the multiclass classifier support vector machine (SVM) using a linear kernel, which can predict belonging probabilities to each class. Our system learns the eight optical sensor values of each mouth shape to train a classifier. This approach leads lower computational cost than processing higher dimensional data such as a camera image. Given that SVM is a supervised model, it requires the correct assignment of labels for these sensor values in the training phase. Therefore, the optical sensor values should have correct labels.

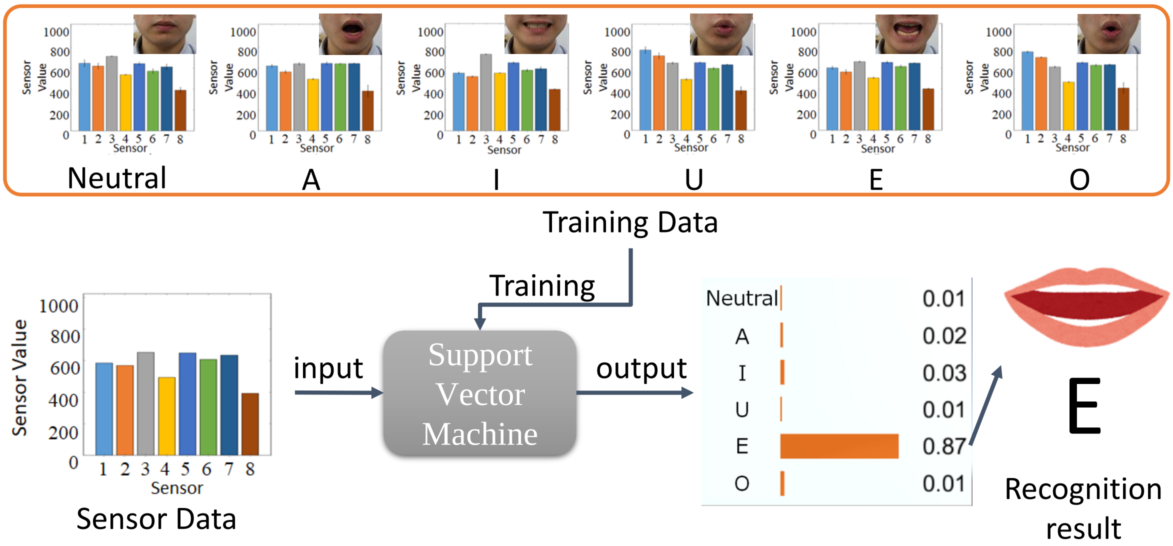


Figure 4.4: Learning and Classifying Mouth Shapes with Machine Learning

4.3.2 Labeling Training Data Using Vowel Recognition

This section describes the relation between speech and mouth shape. Speech consists of phonemes, which are the smallest units of speech. Phonemes are characterized by the resonant frequency of air in the vocal tract. The phoneme mainly consists of consonants and vowels. Consonants are generated by dynamic movements of mouth shape, such as changes in expiratory flow or friction. Meanwhile, vowels are generated by stable movements of mouth shape, such as lip circularity and jaw opening. Focusing on such stable movements, this study uses vowels to label the optical sensor values, thereby enabling automated dataset collection. Our previous study [108] expected users to make facial expressions accord with an graphical instruction timely during the training process. On the other hand, this study introduces auditory feature to label facial expression annotations to optical sensor values. However, the dataset can contain outliers if our system recognizes vowels incorrectly.

Outliers are removed from the dataset. A sample in the dataset consists of eight optical sensor values. For outlier removal, the Mahalanobis distance of a sample from the mean of each class are calculated. If the Mahalanobis distance of a sample is greater than a threshold, the sample is removed as an outlier. This removal is iterated until all samples contained in the dataset has less than or equal to the threshold (Formula 4.1).

$$
\begin{aligned}
&X = \{X | X = \{X_0, X_1, ..., X_n\}, D_m({}^{\forall}X) < \sqrt{D_{thr}}\} \qquad (4.1)\\
&X : SensorData\\
&D_m(x) : MahalanobisDistance\\
&D_{thr} : Threshold
\end{aligned}
$$

By investigating the effect of the threshold on mouth shape classification performance, the optimal threshold is decided. As the first step, a threshold is set to 0.0. With the threshold, outliers are removed to obtain training data. A classifier is trained using the training data and classifies the training data to evaluate classification accuracy. Then, the threshold is added to 1.0. The above procedure is iterated until the threshold is 150.0 to obtain the classification accuracy for each threshold. In this study, the proper threshold is defined as the threshold that achieved the highest classification accuracy among the set of classification accuracy. With the proper threshold, the best training data is obtained.

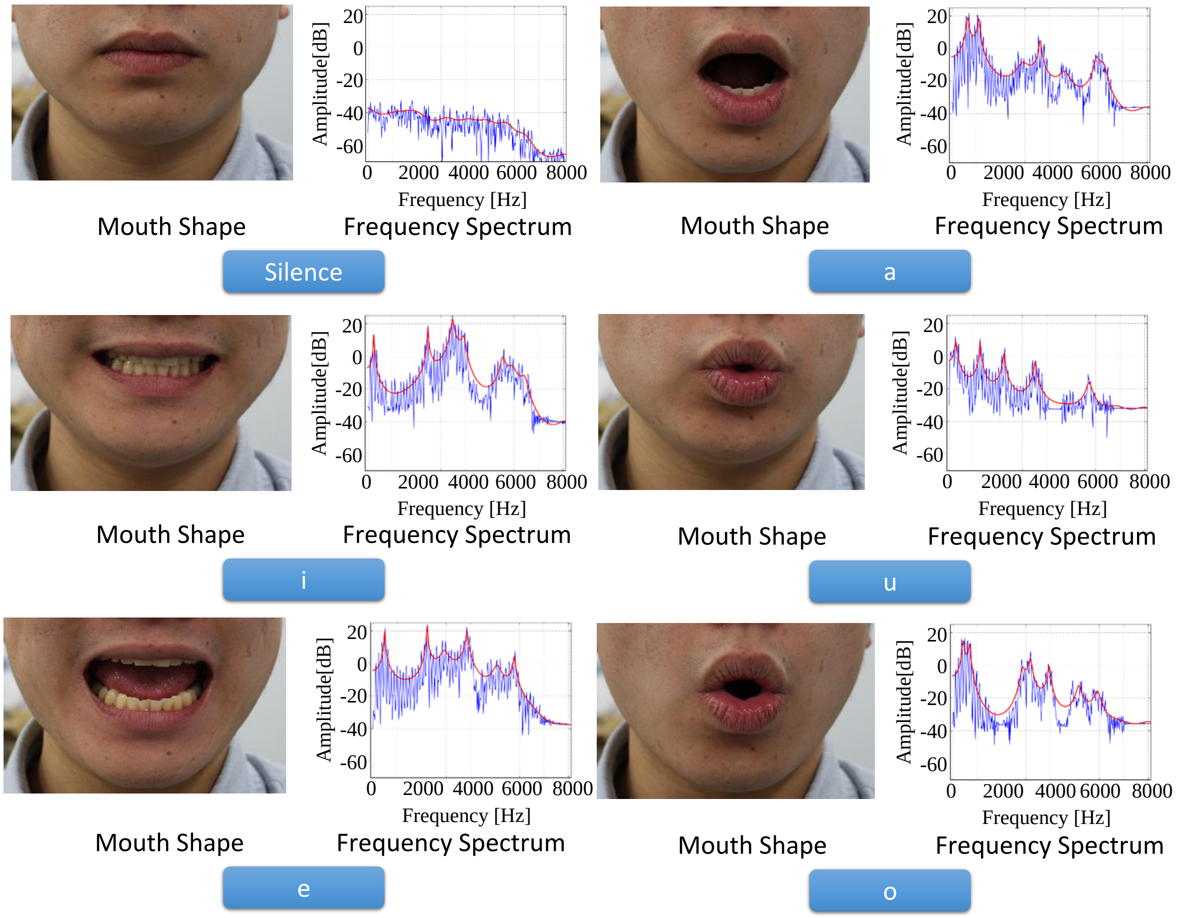


Figure 4.5: The Relation between Mouth Shapes and Vowels. There were six mouth shapes that were the closed mouth (silence) and the five vowels of Japanese. In the line graphs, the blue lines indicate the frequency of each vowel, and the red lines show the spectral envelopes.

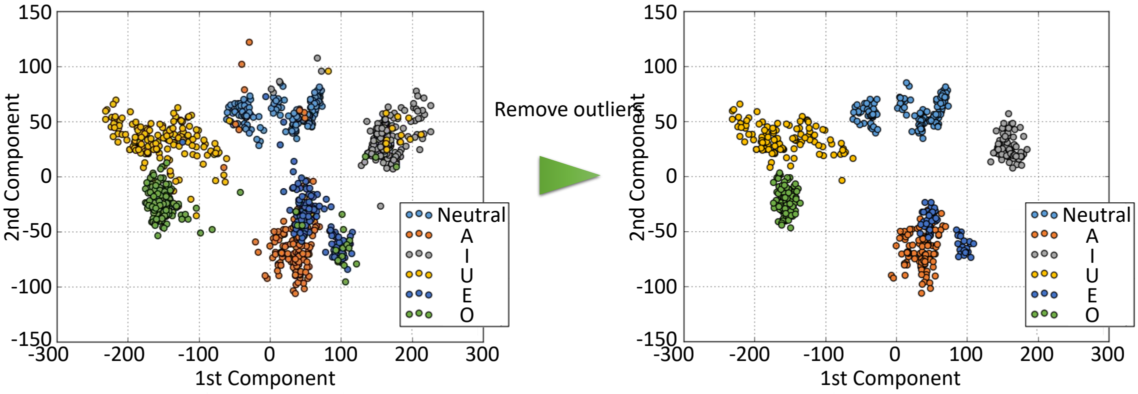


Figure 4.6: Outlier Removal Based on Mahalanobis Distance

4.3.3 Mouth Shape Reproduction

Various mouth shapes are reproduced from the optical sensors (Figure 4.7). The blending of several mouth shapes is assumed to reproduce various mouth shapes. In the preparation of the parameters of several mouth shapes $\vec{P}_i$, $(i = 1, 2, \cdots, m)$, mouth shape P can be expressed as

$$\vec{P} = \vec{P}_0 + \sum_{i=1}^{m} s_i(\vec{P}_i - \vec{P}_B) \tag{4.2}$$

where $\vec{P}_0$ is the parameter of the mouth shape during silence, $\vec{S} = (s_1, s_2, \cdots, s_m)$ is the belonging probability to each mouth shape class.

This approach is the same as that in a previous research [108], which synthesizes facial expressions according to five facial expressions, namely, neutral, happy, angry, surprised, and sad.

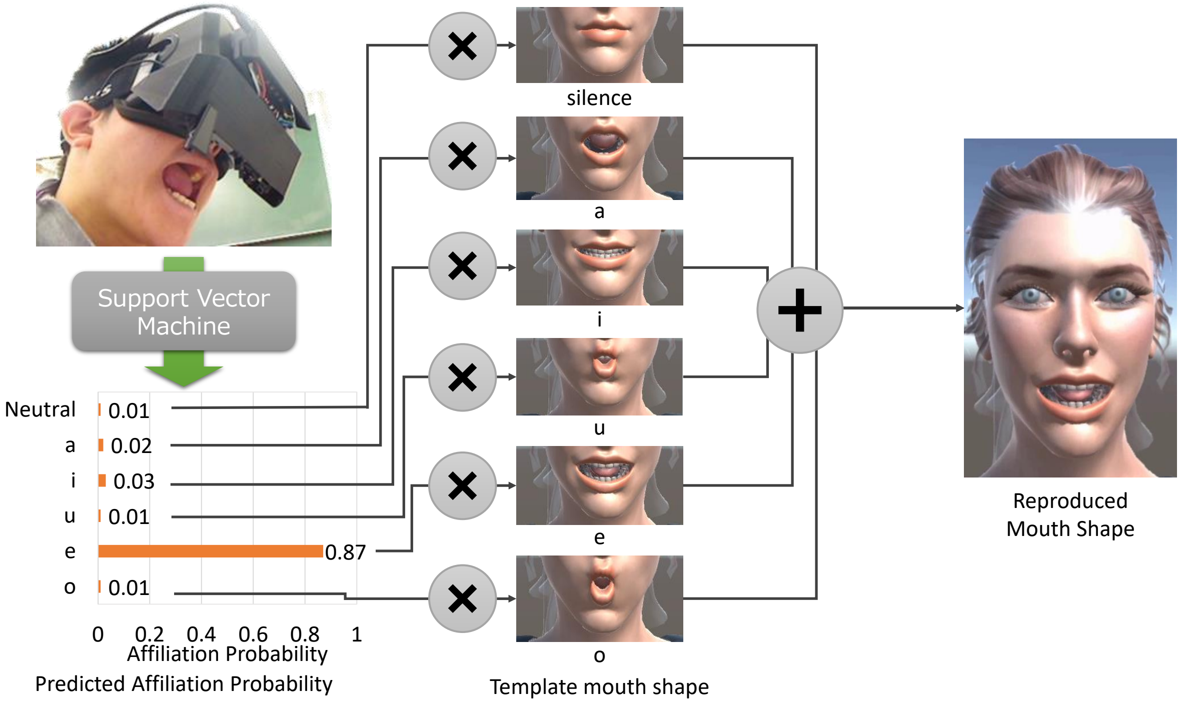


Figure 4.7: Mouth Shape Reproduction

4.4 Implementation

Our system consisted of a computer and a device that measures mouth shape and audio. The device sent the measured sensor values and audio signals to the computer, which then learned the sensor values and recognized the mouth shapes. In training, the computer recognized vowels from the audio signals, labeled the sensor values, and used these labeled sensor values to train an SVM. In recognition, the SVM recognized mouth shapes from the optical sensor values. Section 4.4.1 describes our hardware and Section 4.4.2 is dedicated to the software.

4.4.1 Hardware

A prototype was developed by modifying an HMD to measure the mouth shape and audio (Fig. 4.8). The prototype had four components, namely, photoreflectors (LBR-127 HLD), optical distance measuring units (SHARP GP2Y0A21 YK), a microphone (Audio-Technica AT9904), and a microcomputer (Akitsuki Densho AE-ATMEGA 328-MINI). The photoreflectors and the optical distance measuring units were optical sensors attached to the bottom of the HMD (Oculus Rift DK2 [5]). The microphone was attached to the right side of the mounting surface and connected to the computer through an amplifier (Audio-Technica AT-MA2). The audio signal was directly sent to the computer. Meanwhile, the microcomputer was attached to the front of the HMD and connected to the computer with a USB cable. The microcomputer sent the sensor values of the photoreflectors and the optical distance measuring units. Covers for the sensors and circuits were created using a 3D molding machine.

Figure 4.9 illustrates the arrangement of the photoreflectors and the distance measuring units. The photoreflectors (Nos. 1 to 4) measured the upper lip and the upper cheek. The optical distance measuring units (Nos. 5 to 8) measured the lower lip and the cheek (Fig. 4.10). The sensitivity of the photoreflectors was adjusted for each group (upper mouth and upper cheek) to detect the deformation of the parts. Figure 4.11 shows the relation between the optical sensor values and the distance.

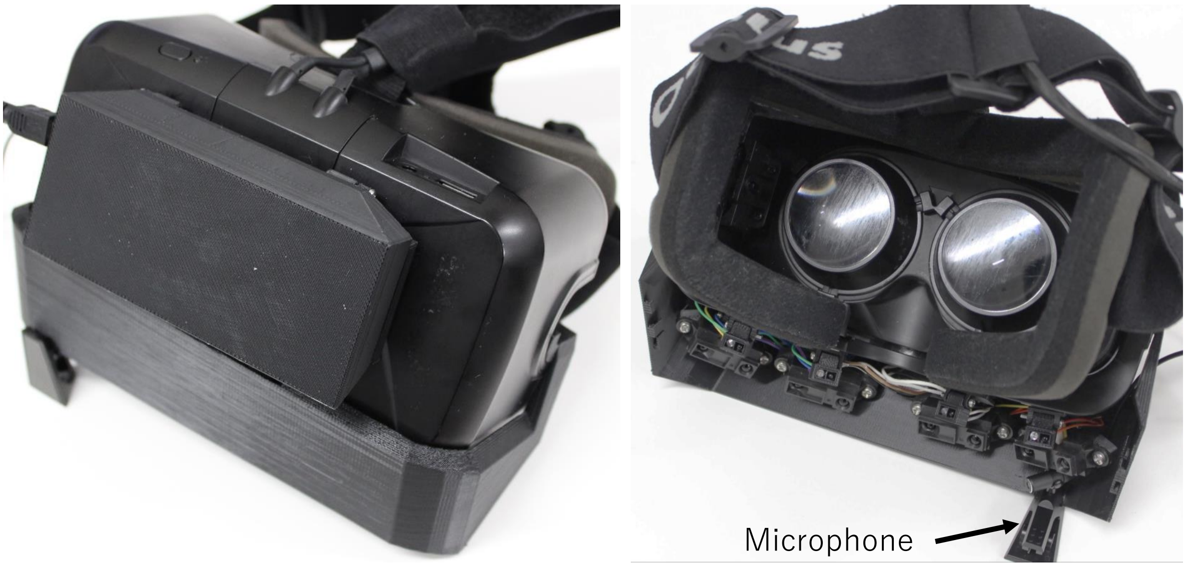


Figure 4.8: Prototype

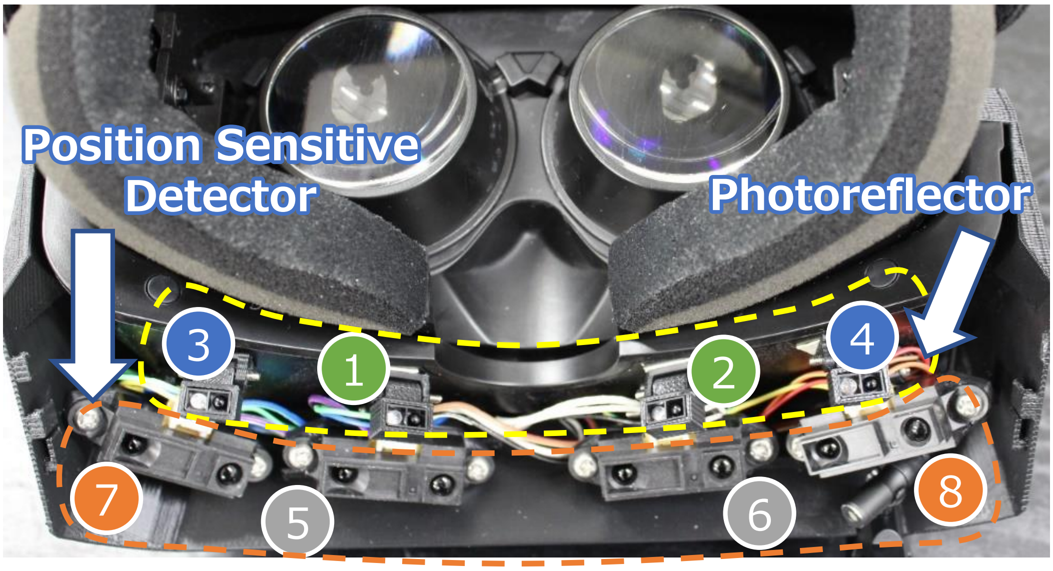


Figure 4.9: Placement of Sensors

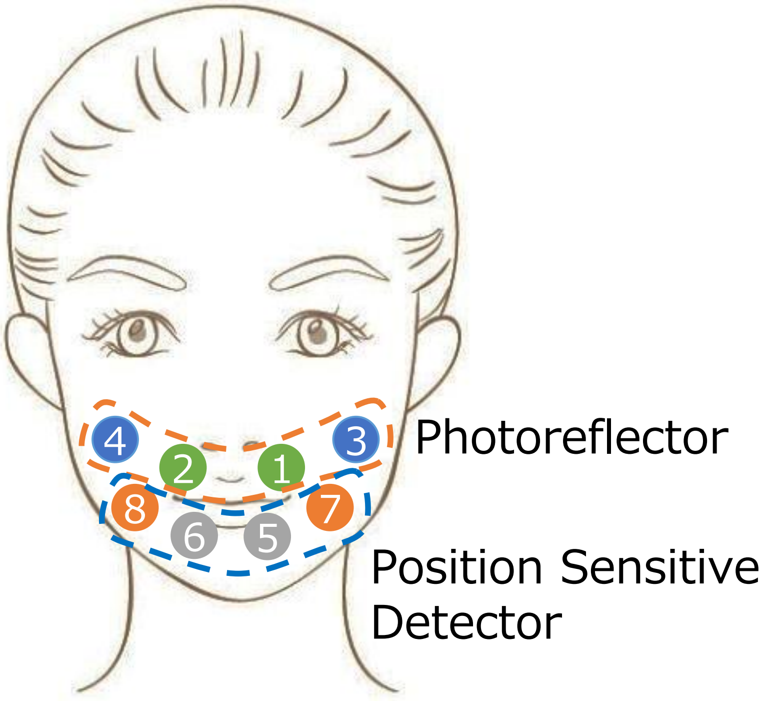


Figure 4.10: Points of Measurement

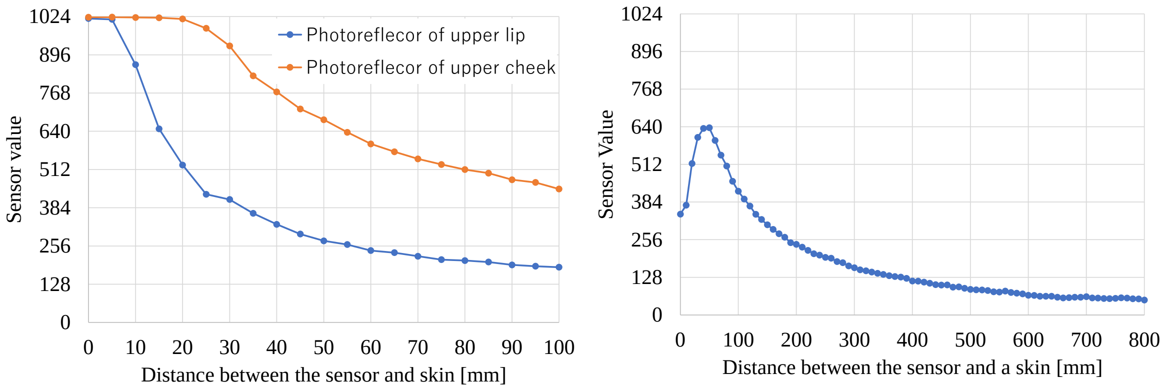


Figure 4.11: The relation between optical sensor values and distance. *Left*: The relation between the distance and photoreflector values. The sensitivity of each photoreflector group was adjusted to the deformation of the part. *Right*: The relation between the distance and PSD values.

4.4.2 Software

Mouth shape recognition was implemented with a machine learning technique. Our software had two processes, namely, training and recognition. In the training process, the computer collected training data by vowel recognition. The computer recognized vowels from the audio signal and labeled optical sensor values with the recognized vowels. In the recognition process, the computer recognized the mouth shape from the optical sensor values. The computer predicted the belonging probabilities of each class and synthesized the mouth shape using these probabilities. Section 4.4.2 and Section 4.4.2 describe the training and recognition processes, respectively.

Training Process

The dataset was collected automatically by vowel recognition. The computer received eight optical sensor values and audio signals. At first, the computer recognized a vowel from the audio signals. Then, the computer labeled the eight optical sensor values with the vowel, thereby enabling the collection of the dataset of vowels. The computer provided a waiting period before such dataset is collected. During this period, the computer acquired the optical sensor values, which were labeled by the computer with "silence." Thus, the dataset was obtained.

Ourliers were removed from the dataset to obtain the training data. The computer calculated the Mahalanobis distance of the samples for each class in the dataset and then eliminated the samples whose Mahalanobis distances were higher than the threshold. After elimination, the training data was obtained for training the SVM.

Recognition Process

The mouth shapes were recognized with eight optical sensor values. The computer inputted the eight optical sensor values it received to SVM, which then predicted the belonging probabilities to each mouth shape class. The computer detected the class that had the highest probability of these classes. The computer then regarded this class as the recognition result.

4.5 Experiment

Experiment 1 evaluated the recognition accuracy of six mouth shapes by the embedded optical sensors. Five of six mouth shapes were of mouths speaking five Japanese vowels. The remaining shape was one during silence. Training data was collected manually and called the method of learning mouth shapes in this experiment "manual learning." Experiment 2 evaluated the recognition accuracy of mouth shapes by automated labeling using vowel recognition. In particular, the effect of automated labeling method on the recognition accuracy was examined. The recognition accuracy of six mouth shapes was compared with the results of Experiment 1. Then, training data was collected automatically using vowel recognition and called this method of learning mouth shapes "automatic learning." In Experiment 3, user-independent classification accuracy was evaluated to investigate the possibility of building a generalized classifier. The inter-participant classification accuracy was compared with intra-participant one.

4.5.1 Experiment 1: Mouth Shape Recognition by Manual Learning

This experiment investigated the recognition accuracy of the following mouth shapes by the embedded optical sensors: "silence," "a", "i", "u", "e", and "o". "Silence" had a closed mouth and no facial expression. "a", "i", "u", "e", and "o" were the mouth shapes for five Japanese vowels. The subjects of the experiment were five Japanese males in their twenties, none of whom had speech disorders or abnormalities in peripheral shapes, including the mouth. The experimental procedures were as follows.

1. The experimenter explained the six mouth shapes ("silence," "a", "i", "u", "e", and "o") to the subjects. After the explanation, the experimenter instructed the subjects to wear our prototype.
2. The experimenter instructed the subjects to make the six mouth shapes and to hold them until the experimenter provided next instruction. The order of instruction was as follows: "silence," "a", "i", "u", "e", and "o". The experimenter manually operated the keyboard to collect 50 samples for each mouth shape.
3. The experimenter iterated Step 2 three times.
4. The experimenter instructed the subjects to make the six mouth shapes again and to hold them until the experimenter gave additional instruction. The order of instruction was the same as in Step 2. The experimenter manually operated the keyboard to collect 200 samples for each mouth shape.

For the training, 900 samples (50 samples * 6 mouth shapes * 3 iterations) were collected in Step 2. For the test, 1200 samples (200 samples * 6 mouth shapes * 1 iteration) were collected in Step 4.

4.5.2 Result of Experiment 1

Table 4.1 shows the result of Experiment 1. The average recognition accuracy of five subjects was approximately 99.9%. Therefore, our method recognized the six mouth shapes with high accuracy.

Table 4.1: Result of Mouth Recognition Accuracy using Optical Sensors

Subject	A	B	C	D	E
Recognition Accuracy	100.0 %	99.3 %	100.0 %	100.0 %	100.0 %

Compared with a previous study on facial expression recognition [108], our system achieved higher recognition accuracy. It is considered that this is because the deformation around the mouth is larger than that around the eyes. This large deformation leads to a wide variance in sensor values, which enables the accurate classification of mouth shapes.

4.5.3 Experiment 2: Mouth Shape Recognition by Automatic Learning

This experiment investigated the influence of the automated labeling method on recognition accuracy. The dataset was collected by combining vowel recognition and mouth shape measurement. Then, to obtain the best training data, this dataset was analyzed. In the analysis, the Mahalanobis Distance of sensor data in the dataset was calculated, and an optimal threshold of Mahalanobis Distance was decided for outlier removal. This optimal threshold was leveraged to obtain the training data, which was then used to evaluate the recognition accuracy of mouth shapes. Finally, the results of this experiment were compared with those of manual learning.

Experiment 2a: Decision of Optimal Threshold

This experiment investigated the optimal threshold of Mahalanobis Distance for outlier removal. The recognition accuracy of mouth shapes was evaluated under the threshold values of 1.0-150.0, with 1.0 change interval. The subjects were the same as those in Experiment 1.

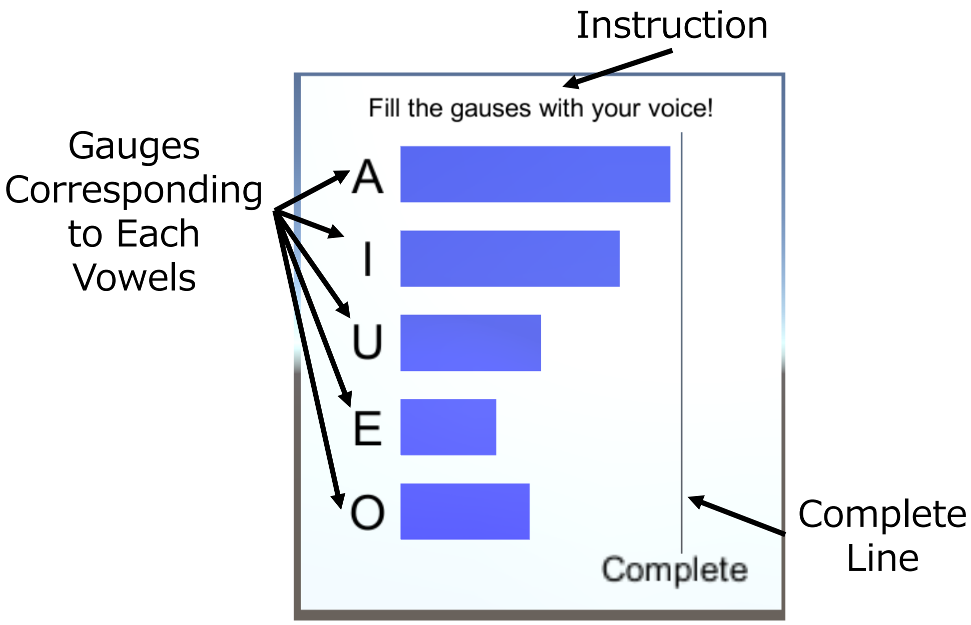

Instruction
Fill the gauses with your voice!
Gauges Corresponding to Each Vowels
A
I
U
E
O
Complete
Complete Line

Figure 4.12: Interface to Collect Dataset Automatically

A dataset labeled by using vowel recognition was collected for each subject. The following is the procedure for collecting the dataset.

1. The experimenter explained six mouth shapes ("silence," "a", "i", "u", "e", and "o") and a user interface for automatic dataset collection displayed on the HMD to the subjects. The experimenter asked the subjects to speak the five Japanese vowels clearly during this experiment. Then the experimenter instructed the subjects to wear our prototype.
2. The experimenter displayed the user interface for learning mouth shapes through the HMD and used it to instruct the subjects to wait. During this time, the experimenter collected 50 samples for "silence."
3. Our system instructed the subjects to speak vowels through the user interface. Speaking instructions and gauges were displayed on the user interface. During this period, The experimenter gathered 50 samples for the five classes ("a", "i", "u", "e", and "o"). The order of speaking the vowels was arbitrary for the subjects.
4. Our system provided a 3 s break to the subjects by displaying a "waiting" instruction.
5. Our system iterated thrice from Step 2 to Step 4.

900 samples (50 samples * 6 mouth shapes * 3 iterations) were collected in Step 2 and Step 3.

Figure 4.12 shows the interface for automatic data collection. The user interface had three components, namely, gauges, an instruction, and a completion line. The upper part of the interface displayed instructions, such as waiting and speaking. The completion line was on the right side of the interface. Arrival of the gauges at this line indicated the completion of sample collection. The center of the interface had the gauges, which indicated the number of sensor data collected for each mouth shape. As the sensor data increased, these gauges extended to the right and eventually reached the completion line. The color of these gauges indicated whether our system completed sample collection: blue meant unfinished, and red meant finished.

The dataset was analyzed to remove only outliers. In the analysis, the threshold of Mahalanobis Distance was explored to separate outliers.

The following is the procedure of analyzing the dataset.

1. Our system set the threshold to 1.0.
2. Our system obtained the dataset filtered with the threshold using formula 4.1.

3. The filtered dataset was divided into training and test data (even and odd).
4. Our system evaluated the recognition accuracy of the threshold. Our system learned training data and calculated classification accuracy on test data.
5. If the threshold was lower than or equal to 150.0, our system added the threshold to 1.0 and return to Step 1. If not, our system finished the analysis.

Thus, the recognition accuracy of each threshold, which was 1.0-150.0 with 1.0 change interval, was calculated. Among this set of recognition accuracy, our system detected the threshold which achieved the highest recognition accuracy.

Result of Experiment 2a

Figure 4.13 shows the experiment results where the recognition accuracy was 80%-100% and the threshold was 0.0-60.0. The recognition accuracy for classes without samples was 0.0%, as our system could not recognize the mouth shape. Table 4.2 shows that several thresholds approximately between 10.0 and 25.0 achieved the highest accuracy.

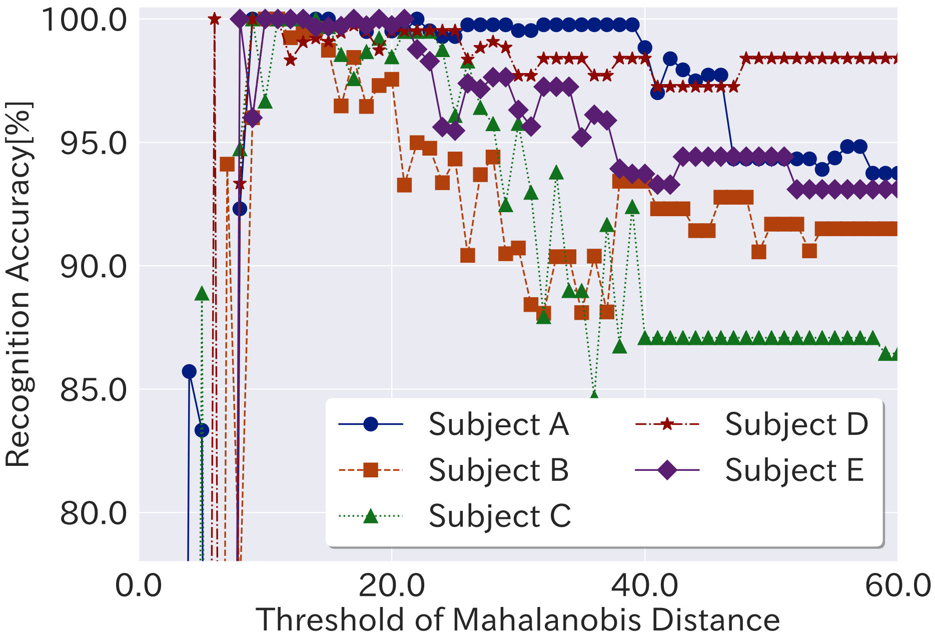


Figure 4.13: Variance in Classification Accuracy of Mouth Shape at Each Threshold

Table 4.2: Maximum Recognition Accuracy and Thresholds of Each Subjects

Subject	The Highest Recognition Accuracy	Threshold
A	100.0 %	9.0, 10.0, 11.0, 12.0, 13.0, 14.0, 15.0, 17.0, 19.0, 22.0
B	100.0 %	10.0, 11.0
C	100.0 %	8.0, 9.0, 10.0, 11.0, 12.0, 13.0, 14.0
D	100.0 %	6.0, 9.0, 10.0, 11.0
E	100.0 %	8.0, 10.0, 11.0, 12.0, 13.0, 17.0, 19.0, 21.0

Figure 4.14 shows the graph of the threshold and the number of datasets after outlier removal. Figure 4.15 is the graph of the number of datasets after outlier removal where the recognition accuracy was 80%-100%.

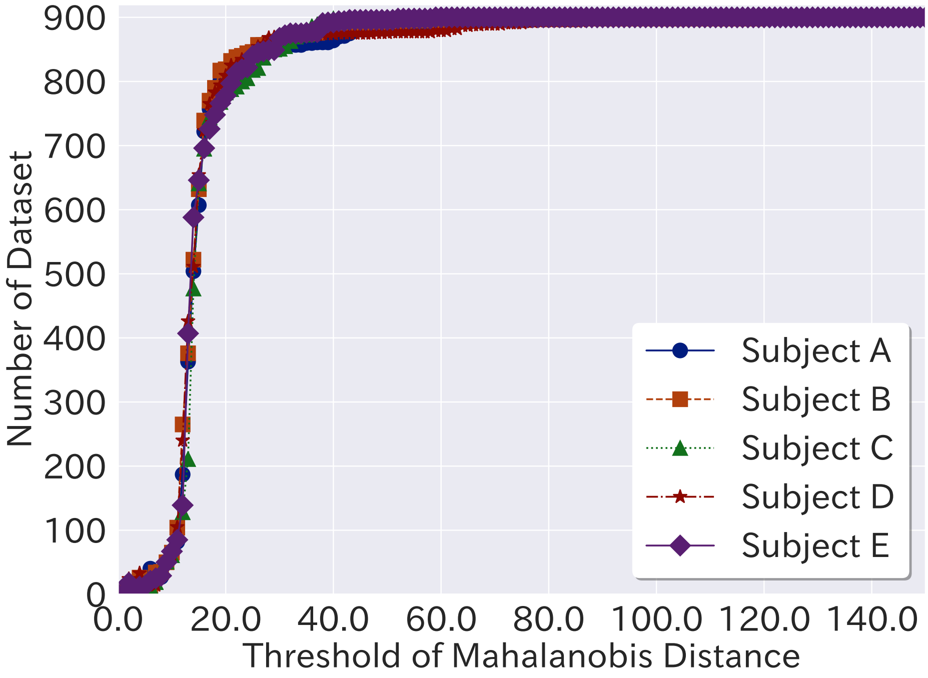


Figure 4.14: Threshold and Number of Dataset

Figure 4.14 reveals that the number of datasets significantly fluctuated between approximately 100 and about 800 when the threshold was between 10 and 25. According to Fig. 4.15, the recognition accuracy decreased when the number of the dataset was out of the 100-700 range. This finding implies that the shortage of the dataset and insufficient removal of outliers resulted in the decreased recognition accuracy. Therefore, it is considered that the optimal threshold was the maximum value among Table 4.2. For example, the optimal threshold for subject A was 22.0.

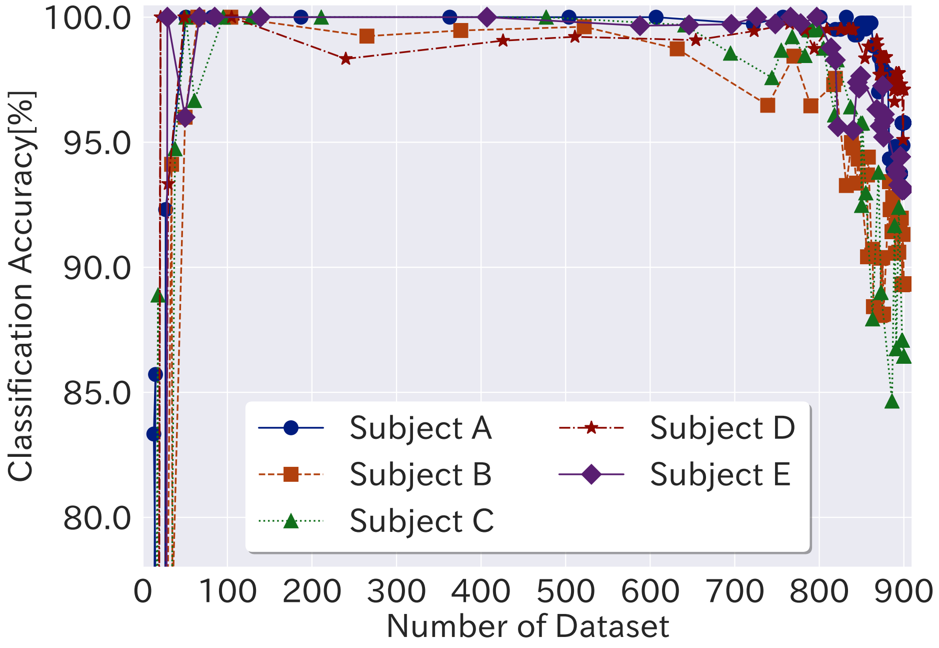


Figure 4.15: Number of Dataset and Recognition Accuracy

Experiment 2b: Mouth Shape Recognition in Automatic Learning

This experiment evaluated the recognition accuracy by the automated labeling method by comparing the recognition accuracy in this experiment with those of Experiment 1. This experiment was conducted after Experiment 2a without removing the HMD, and the subjects were the same as those in Experiment 1.

Outliers were removed from the dataset to obtain the training data. The maximum threshold in Table 4.2 was used and 1,200 samples (200 samples * 6 mouth shapes * 1 iteration) were collected as the test data for each subject. The procedure of data collection was the same as in Step 4 of Experiment 1.

Result of Experiment 2b

Table 4.3 shows the result of Experiment 2b. The average recognition accuracy of the five subjects was approximately 96.3%. Figure 4.16 indicates the comparison of the result of Experiment 2b with that of Experiment 1. According to the 4.16, compared with the recognition accuracy in manual learning, that in automatic learning decreased by about 3.6%. Nonetheless, our automatic labeling method classified the six mouth shapes accurately.

Table 4.3: Result of Mouth Recognition Accuracy using Optical Sensors in Automated Labeling Condition

Subject	A	B	C	D	E
Recognition Accuracy	100.0 %	95.7 %	86.1 %	100.0 %	100.0 %

The recognition accuracy of subject C decreased significantly compared with that in Experiment 1. For analysis of this result, subject C's training data was visualized with principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) in Fig. 4.17 and Fig. 4.18, respectively. Also Table 4.4 shows the subject C's confusion matrix of mouth shape recognition. According to Fig. 4.17 and Fig. 4.18, several samples of "e" were close to clusters of "i" and "o". This implies that, in collecting the dataset, his mouth shape of "e" differed at each trial. Therefore, outliers of subject C's "e" increased. This large number of outliers led to an increase in the Mahalanobis distances of all samples and thus insufficient outlier removal. Table 4.4 indicated that our system mispredicted 53.5% of "i" as "e" and mispredicted 30.0% of "e" as "o". These findings suggest that our mouth shape recognition accuracy depended on the stability of the reproduction of the user's mouth shape.

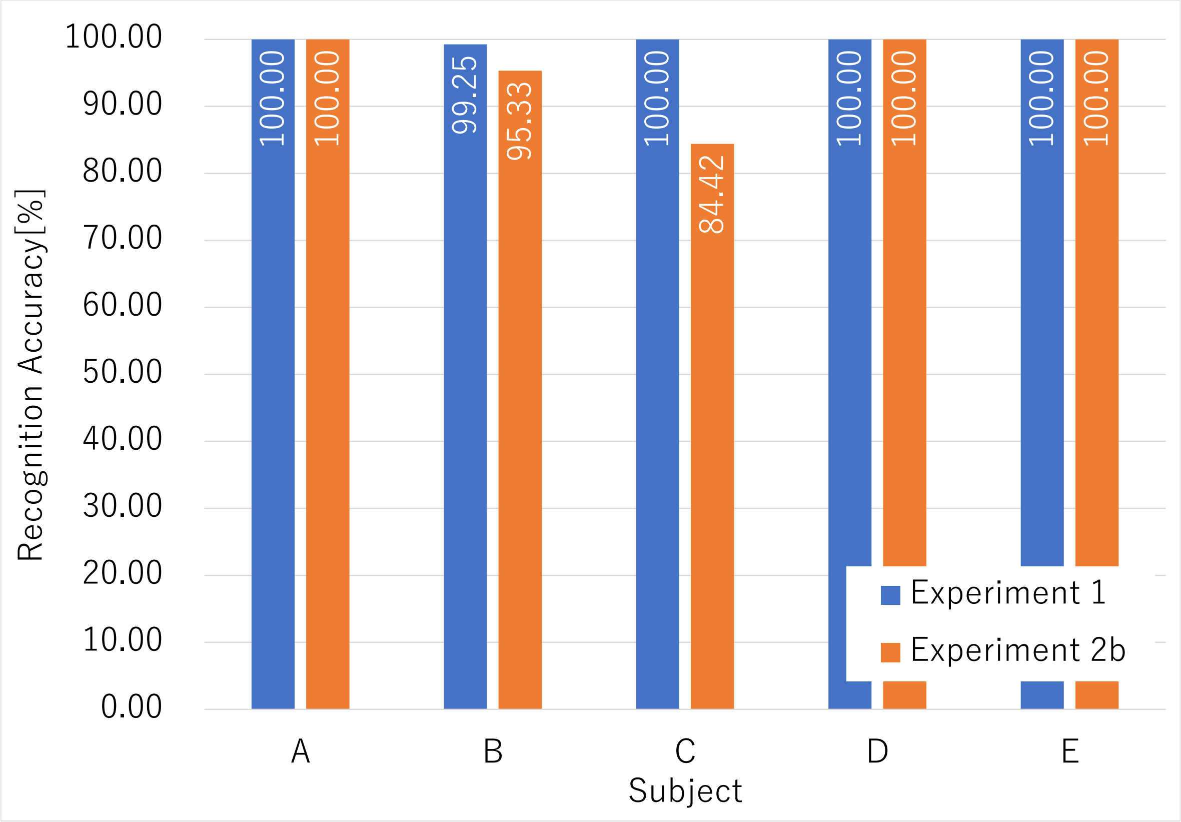


Figure 4.16: Comparison of the mouth shape recognition accuracy between Experiment 1 and Experiment 2b

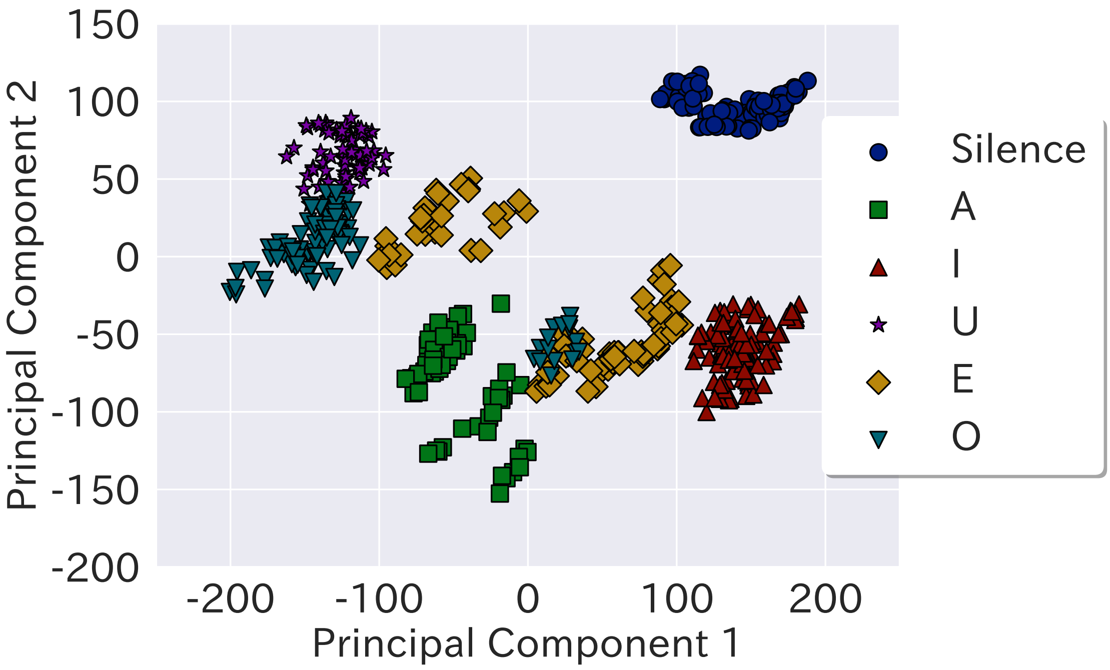


Figure 4.17: PCA Result of Subject C's Training Data

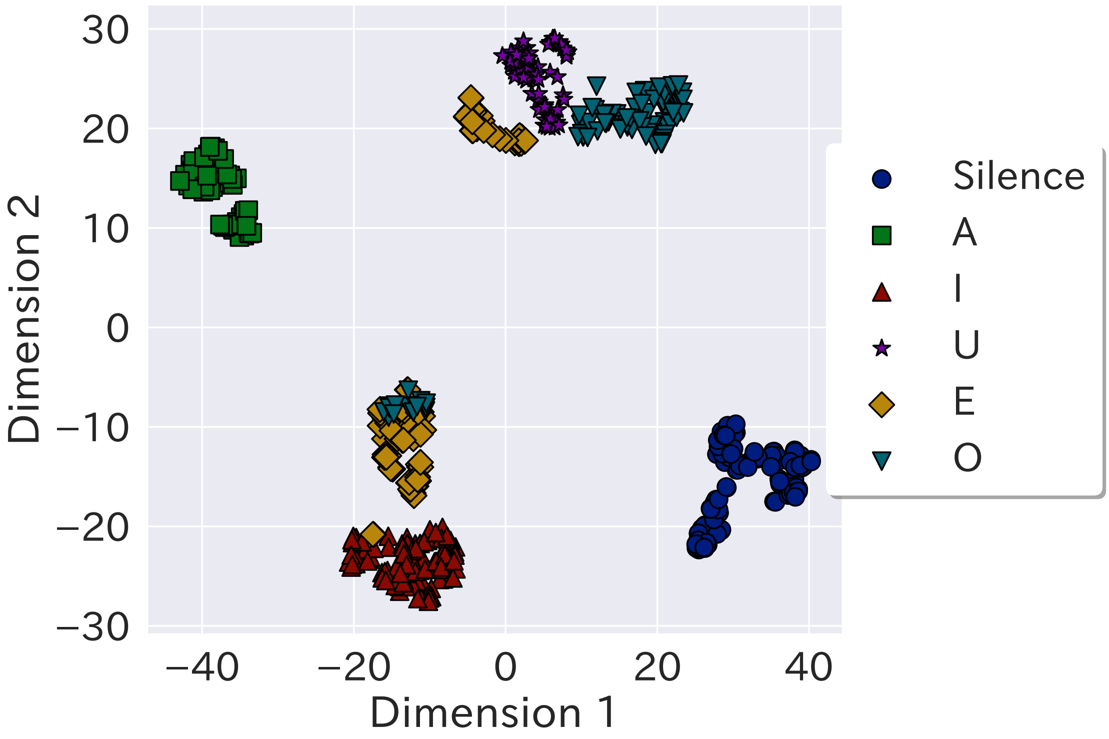


Figure 4.18: t-SNE Result of Subject C's Training Data

Table 4.4: Subject C’s Confusion Matrix of Mouth Shape Recognition

		Predicted Label					
		Silence	A	I	U	E	O
Correct Label	Silence	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	A	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%
	I	0.0%	0.0%	46.5%	0.0%	53.5%	0.0%
	U	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
	E	0.0%	0.0%	0.0%	0.0%	70.0%	30.0%
	O	0.0%	0.0%	0.0%	0.0%	0.0%	100%

4.5.4 Experiment 3: User-Independent Accuracy

This experiment investigated the user-independent classification performance of our mouth shape recognition method. In photo reflective sensor-based facial expression classification, there was high user-dependency [140]. This experiment compared inter-participant classification accuracy with intra-participant one to evaluate the user-dependency of our recognition method. There were ten participants (8 males, 2 females, in their twenties).

A dataset was collected as follows.

1. The experimenter explained the six mouth shapes ("silence", "a", "i", "u", "e", "o"), which were classification targets. Then, the experimenter instructed the participants to sit on a chair and wear the mouth shape recognition device tightly fixed to their heads. At that time, the experimenter showed no visual images on display.
2. The experimenter instructed the participants to form the "silence" mouth shape exaggeratedly and keep it. While the participant kept the shape, the experimenter collected 200 samples by hand.
3. The experimenter instructed to make the other five mouth shape ("a", "i", "u", "e", "o") with step 2. The instruction order was "a", "i", "u", "e", and "o".
4. The experimenter iterated the steps from 1. to 3. three times.

Through the above procedure, 3600 samples (200 samples * 6 mouth shapes * 3 iterations) were collected as the dataset for each participant. A classifier was trained with the dataset of one of the participants and, by using the trained classifier, classified the dataset of each participant to evaluate mouth shape classification accuracy.

4.5.5 Result of Experiment 3

The result is shown in Table 4.5.

The user-dependent average classification accuracy was 99.4%, while the user-independent average classification accuracy was 30.7%. Since the highest user-independent accuracy was 88.1%, several participants had similar mouth shapes. However, most of the user-independent accuracy was very low. This suggests that our current mouth shape recognition algorithm has high user-dependency as well as the previous study [140]. Therefore, in our current system,

Table 4.5: Cross-Participant Classification Accuracy Matrix

		Training data									
	Participant	A	B	C	D	E	F	G	H	I	J
Test data	A	100.0%	16.7%	24.8%	60.6%	40.5%	48.4%	39.4%	16.7%	52.9%	56.6%
	B	25.0%	99.9%	16.7%	17.8%	17.1%	16.7%	16.7%	34.1%	31.8%	23.2%
	C	33.1%	20.2%	95.9%	18.5%	16.7%	16.8%	16.7%	29.9%	24.8%	17.5%
	D	50.1%	28.1%	21.6%	100.0%	32.9%	42.3%	49.9%	22.7%	45.9%	88.1%
	E	38.9%	44.9%	37.8%	22.2%	100.0%	16.7%	16.8%	40.7%	16.9%	17.4%
	F	32.8%	16.7%	16.8%	33.8%	16.7%	100.0%	63.1%	16.7%	53.8%	68.3%
	G	16.7%	16.7%	16.7%	25.0%	16.7%	68.9%	100.0%	16.7%	38.6%	52.1%
	H	19.6%	49.1%	33.4%	29.0%	16.7%	22.7%	16.7%	100.0%	16.7%	16.7%
	I	37.7%	17.2%	33.3%	33.3%	16.7%	55.6%	33.3%	22.1%	99.6%	33.3%
	J	32.4%	16.7%	16.7%	66.8%	16.7%	40.3%	50.0%	16.7%	37.4%	100.0%

it is required to train the classifier individually.

4.6 Application: Reflecting Mouth Shape on Avatar

An application that transferred the HMD user's mouth shape to an avatar was developed (Fig. 4.19). This application blended the parameter of the six mouth shapes ("silence," "a", "i", "u", "e", and "o") to reproduce the mouth shapes.

The procedure by which this application reflected the mouth shape on the avatar is described. Our system predicted the probabilities of each mouth shape from the optical sensor values. By using these probabilities, the application calculated the parameter of the mouth shape blended on the basis of Formula 2 and then applied this parameter to the avatar. Figure 4.20 shows that the avatar reflected the blended mouth shape. This technique enabled the reflection of the various movements around the mouth, which were not limited to six states.

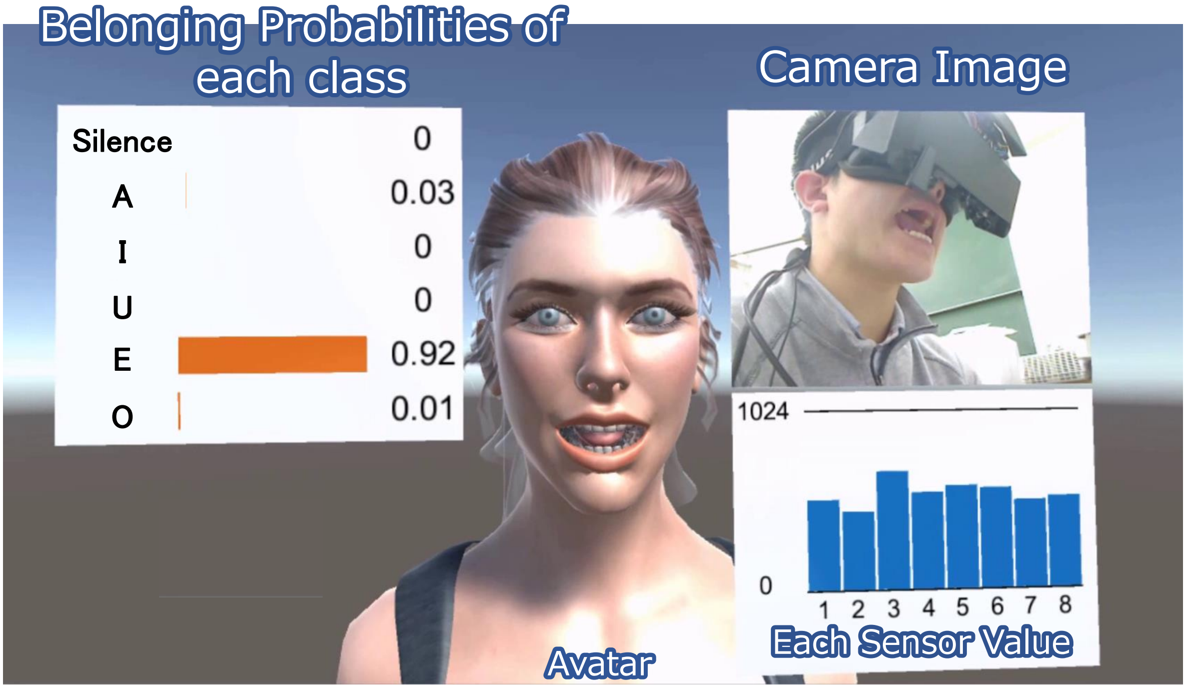


Figure 4.19: Application Reflecting User's Mouth Shape on Avatar

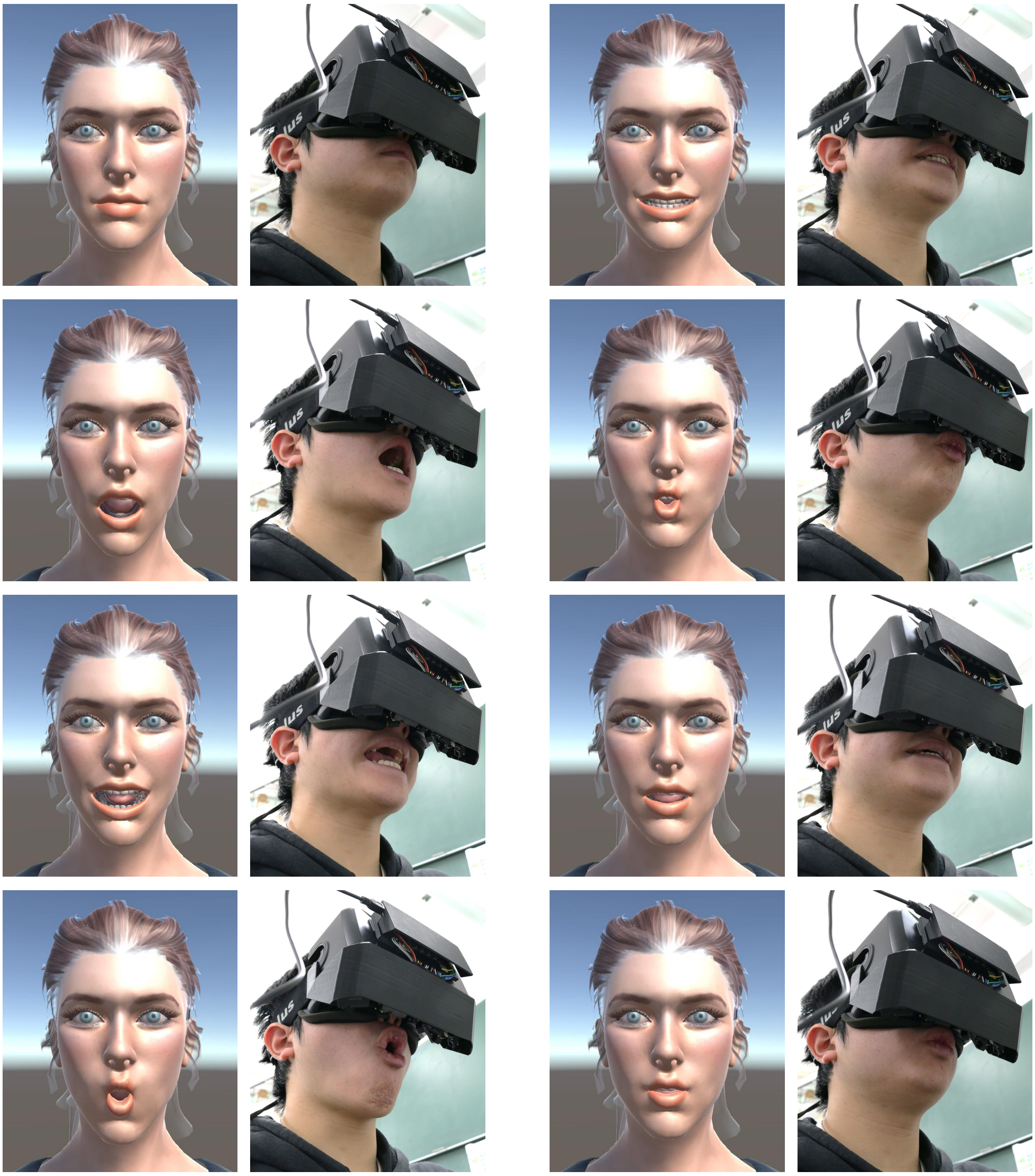


Figure 4.20: Reflecting Animation of User’s Various Mouth Shape to Avatar

4.7 Limitation

Although our system reproduced mouth shapes based on six template mouth shapes of silence and Japanese vowels, there were difficulties in reproducing specific mouth shapes (e.g., lip biting, cheek swelling) that were not formed in speaking. This study focused on the mouth deformation that occurred with speaking Japanese vowels. Our current implementation was trained with mouth shapes without considering temporal information. However, when humans speak, mouth shapes continuously deform. Consonants are produced by dynamic mouth shape changes. By taking into account the dynamics of mouth expressions, our system can reproduce more various mouth shapes.

Although various mouth shapes were reproduced by blending six mouth shapes, it is not discussed whether the class membership probabilities is suitable as weights of blending or not. Therefore, the system may not have blended the mouth shape accurately. In the future, the suitable blending method will be explored by analyzing the geometry deformation of mouth shapes.

The result of Experiment 3 showed that our current system failed to construct a cross-user classifier. Since the previous study on photoreflector-based facial expression classification [140] also failed to build a cross-user classifier, it is difficult to achieve a generalized classifier by classifying sensor values with a traditional SVM. Therefore, to build a generalized classifier, preprocessing method of sensor data and machine learning models, such as deep learning, should be investigated.

This study used only optical sensor values for recognizing mouth shapes after the training process. Auditory information may be used as an additional feature to recognize mouth shapes more robustly.

This study employed young Japanese participants in the experiments and most of the participants were males. The experiments aimed to investigate how accurately our technique recognized the mouth shapes of each participant. The experimental result showed that our method accurately classified the mouth shapes for each participant. In terms of gender bias, there were two female participants in Experiment 3. Therefore, to investigate the gender effect on our method further, additional female participants should be recruited. Since this study targeted facial skin movements caused by speaking Japanese vowels, our system was designed for Japanese speakers. However, the participants were recruited within a limited age group and races. Facial muscles decline by aging. In addition, facial movements vary from language to language when speaking. However, our system collected training data for each user and used an SVM, which is a supervised learning model. Therefore, our method will be applicable to elderly people. In the case of languages other than Japanese, our system will work by taking into account vowels for labeling.

4.8 Conclusion

This study proposed a system that recognizes the mouth shapes of HMD users with optical sensors. An HMD-based prototype were developed with four photoreflectors, four optical distance measuring units, a microphone, and a microcomputer. This prototype measured mouth shape and audio signals. The photoreflectors and optical distance measuring units detected the movement of eight points (upper lip, upper cheek, lower lip, and cheek). The microphone acquired audio signals. Meanwhile, our system detected the vowels from audio and used them to label the optical sensor values.

From the manual learning experiment, our system achieved an average accuracy of approximately 99.9% for the five subjects. The automated labeling method achieved an average accuracy of about 96.3% for all subjects. The recognition accuracy of automatic learning was lower by about 3.6% than that of manual learning. Nonetheless, it is considered that our system could label training data properly through our experiments. In addition, the result of Experiment 3 indicated that our method achieved the user-independent classification accuracy of 30.7%, which was lower by 68.7% than the user-dependent one. Therefore, our current system has difficulty building a generalized classifier.

An application that projected the mouth shapes to an avatar were developed. The application predicted the class membership probabilities to each mouth shape class, and blended each mouth shape on the basis of the class membership probabilities to reproduce various mouth shapes. This application showed that our system could reflect various mouth shapes on the avatar.

Chapter 5

Cheek Haptic-Based Spatial Directional Guidance

5.1 Introduction

This study presents Virtual Whiskers, a spatial directional guidance system attached to a Head-Mounted Display (HMD) with facial haptic stimuli to cheeks.

Spatial guidance cues are essentials in Virtual Reality (VR) applications. Most of them rely solely on visual perception, but VR already has a lot of visual information to process (arguably, more than in real-life), and too much visual information can cause visual overload [141]. Some rely instead on audio-visual perception to reduce the visual workload, but they also have an effect on the workload since both the visual/audio coordinates need to be transformed to the body coordinate [142] and the head/eyes typically need to look at the visual/audio target. Another approach is to use haptic cues [143], which are less likely to be overloaded, are already mapped to the body coordinate, and do not require to look at the target.

Previous works already used haptic-based guidance successfully [12, 13], mostly by relying on vibrotactile stimulation. A vibrotactile system is easy to use but needs to be placed over the targeted zone and provide limited information.

Moreover, in a haptic-based guidance task, where precision matters, not only the type of stimulation (vibration, pressure, wind, etc.), but also the positioning of the stimuli need to be considered. Indeed, if the head is stationary, the position of haptic stimuli tends to be perceived relatively to a body-centered reference frame [144] (i.e., the body midline); else if the head is not stationary (like in a VR application), the stimuli tend to be perceived relatively to an eye-centered reference [145]. If the haptic stimuli are on the torso when the head moves, there is a shift of the perceived position of the stimulation [16].

With facial haptic stimuli at the eye-level, humans can avoid having two reference frames when the head is non-stationary. A previous study on facial haptics revealed that checks' facial haptic was better in localization perception compared to the forehead and to above the eye-brow [89]. Several facial-based systems have been developed, such as winds [121] and ultrasounds [89].

With the advent of consumer VR Head Mounted Display (HMD), consumer VR Haptic devices have been released to the market [*1] [*2]. Those devices typically stimulate hands (e.g., haptic gloves), waist (e.g., haptic belt), or the body (arms, legs, chest, etc.; e.g., haptic suit). Nevertheless, there are not many devices targeting the user's face, despite the excellent sensitivity of the facial region [146].

This study proposes an HMD-based facial haptic system that provides stimulus to the cheeks. It consists of two robotic arms attached to the bottom side of an HMD (c.f., Fig. 5.5) with proximity sensors. This study investigates how haptic cues on the cheek provide directional information; and how facial haptic cues allow spatial guidance in a Virtual Environment (VE). Our contributions are as follow:

- Cheek stimulation by robot arms integrated with HMD to provide spatial navigation. The robot arms are moved with proximity sensing to control contact with our cheek surface;
- Investigating how cheek stimulation affects directional guidance. Our experiment on directional guidance showed that haptic cues on the cheek provided accurate direction two cues in VR space, but that azimuthal angular accuracy was better than elevational one.
- Investigating how cheek stimulation guide users; The experiment investigated how haptic cues on the cheek improve task performance. In the target searching task, our guidance technique shortens task completion time than only visual information and enhanced spatial directional perception.

[*1] bhaptics, https://www.bhaptics.com/tactsuit/

[*2] HaptX | Haptic gloves for VR training, simulation, and design, https://haptx.com/

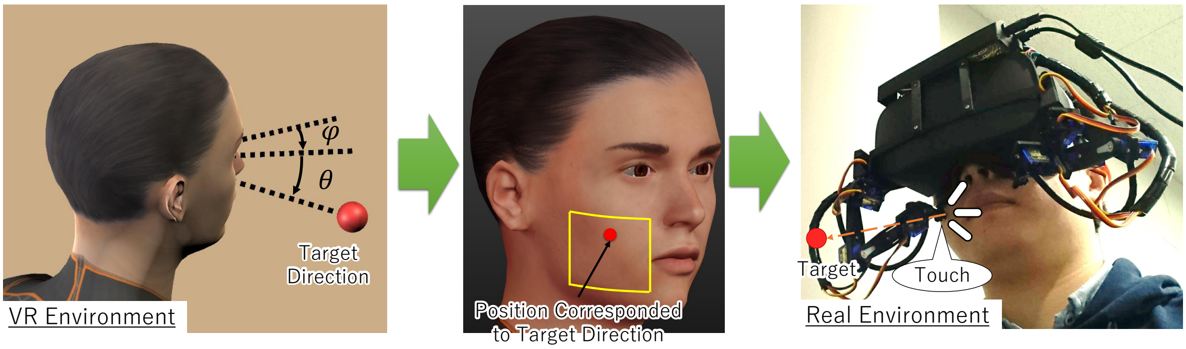


Figure 5.1: Spatial directional guidance with cheek haptic stimulation. *Left*: The target (red sphere) is located in a virtual space. The direction of the target is represented by the azimuthal angle φ and the elevation angle θ in spherical coordinate. *Center*: The azimuthal and angle (θ, φ) are mapped to a point on a cheek surface (x_s, y_s, z_s). The spatial direction in the virtual space is mapped to a cheek position in real space. *Right*: The robot arm moves to the point on the cheek and touches to cheek surface. Thus our system presents the direction of the target in virtual space to a user.

5.2 Related Work

Haptic feedback plays an important role in VR and computer-human interaction. Haptic receptors are distributed throughout the body, while receptors for other major senses: vision, audio, taste, and smell are located in specific facial location. Haptics includes senses for various type of stimuli, such as touch, temperature, and pressure. In previous studies, haptic stimulation has been leveraged to stimulate various body parts and has been integrated other modalities such as visual and audio cues in order to enhance VR experiences as a multi-/cross-modal stimulation.

Our torso has a large surface area. As such, numerous previous studies presented haptic stimuli on this area. Delazio et al. developed a force feedback device that used airbags located on the side of a vest to improve the VR experience [62]. Our hands can be considered as one of the most sensitive regions for haptic stimuli. Günther et al. used a tactile glove that had multiple embedded tactors to navigate 3D space by encoding spatial information into vibration patterns [13]. Chen et al. developed a handheld pin-array haptic display to present a direction to the palm [12]. Ion et al. presented an arm-worn skin drag display to let users recognize a tactile shape [61]. A quadcopter was utilized as a haptic display in 3D space to render a touchable surface [112]. One remarkable system is the wearable robotic arm approach. Shen et al. designed a neck augmentation system using a robotic arm attached to the top of the user's head [147]. AI-Sada et al. developed a wearable robotic arm that provided haptic feedback to a VR user [110]. They attached multiple haptic actuators to the end effector of the arm to enhance the VR experience. However, it was reported that, when the head was not stationary, haptic stimuli to the body were encoded in an eye-centered reference frame [145]. The superiority of the eye-centered frame in directional perception has been reported.

Equally important, our head is another of the most sensitive parts capable of being used for haptic stimuli as indicated in Penfield's cortical sensory homunculus. So, various interactive haptic devices for the head region have been proposed in previous studies. Tseng et al. attached physical widgets to the HMD to enable tangible interaction [32]. By controlling the widgets physically, they interacted with the virtual environment. Cassinelli et al. built a prototype to detect the surroundings and provide haptic feedback to the head [83]. Berning et al. encoded distance information about the user's surrounding objects into pressure values and presented it to the head, allowing the user to perceive spatial information [85]. Not only 2D- but also 3D- directional interaction techniques have also been proposed. Tsai et al. presented a 2.5D instant impact on an HMD using the impact devices attached to the front side of the HMD [31]. Matsuda et al. developed a necklace-type device to indicate directions via vibration patterns for remote collaboration [65]. Beren et al. placed multiple vibrotactile motors around the head to guide the user even at different heights [84]. Oliveira et al. designed a haptic guidance system with vibrotactile motors around the forehead [28]. They translated the azimuthal direction into vibrational position and the elevational direction into vibrational frequency. Moreover, thermal feedback was leveraged for providing spatial directional cues to the forehead [29, 30]. As the above previous studies indicated, facial haptics is useful for spatial interaction, such as receiving the feedback from the virtual environment, spatial guidance, and spatial awareness.

As a part of the face, the cheek is also sensitive to haptic stimulation, but there are few approaches that leverage it for providing feedback. In the field of brain computer interface, cheek stimuli potential was explored [148]. The cheeks' properties for haptic stimulation were investigated. A previous study showed that in-air ultrasonic haptic cues on the cheek allow users to perceive stimulus location well [89]. Liu et al. found that synchronizing haptic stimulation to the cheek and visual oscillation by user's footstep reduced VR sickness [120]. Theo et al. integrated visual stimulation with the cheek haptic feedback and vestibular stimulation by the electronic current to present weight sensation [149]. The above approaches presented haptic stimulation to specific locations on the cheek, while Wilberz et al. mounted a robot arm to an HMD to provide haptic stimulation around the mouth in fully localizable positions [121]. They attached some actuators to provide multiple haptic feedback and found that users could judge directions from wind cues. They also showed the multi-modal haptic feedback improved the overall VR experience.

Haptic systems are useful for augmenting a human's ability. Primarily, the cheeks and mouth are more sensitive than the other facial areas. Previous studies revealed that the cheek's potential of directional perception was superior to other areas and that fully localizable stimulation improved the VR experience. Most mouth stimulation approaches presented ambient information such as wind [121] and investigated horizontal directional cues. By utilizing cheeks' directional potential, this study developed a cheek haptic-based guidance system in 3D space. Lip skin is thin and, as such, susceptible to various damage; direct stimulation to the lips can cause discomfort. Therefore, this study explored the potential of haptic feedback to only the cheeks, minus the lips. This study investigated how "direct" haptic stimulation to the cheek could guide in 3D space. Also, through our experiments, this study investigated how vertical movements on the cheek can guide in the elevational plane. In order to provide haptic cues on cheeks, this study developed a robot arm-based haptic stimulation system because the robot arm can stimulate at a free point on the cheeks.

5.3 System Design

This section describes the principle of our guidance method. This study leverages the cheeks for presenting directional cues to an HMD user. Our face has a dense distribution of haptic receptors as indicated in Penfield's cortical homunculus (Fig. 1.1). Especially, the mouth and cheeks are sensitive to mechanical stimuli, so a human can recognize the stimulation position precisely. However, our lips are more susceptible to even slight stimulation due to their thin skin layer. In addition, the lip engages in essential activities such as eating and speaking. Therefore, this study avoids stimulating the lip. Strong force toward the teeth can injure the mouth because of the teeth's hardness. Therefore, both sides of the cheeks are touched with weak to moderate force.

For stimulating the cheek, two robotic arms are attached to an HMD. A human can recognize the stimulation position on the cheek, and a haptic device that can stimulate the precise position is required. Therefore, a robotic arm with several linkages is utilized to stimulate a localizable position on the cheek. The robot arms are attached to an HMD to present the stimulation even while walking around. However, if only one robot arm is used, it increases the overall weight because it requires a linkage extension and high-torque motors for the joint. Therefore, two robot arms are employed for stimulating the left and right sides of the cheeks.

When touching the cheek with the robot arms, the surface geometry information of the cheek is required. A camera-based approach is popular for measuring facial geometry. However, an HMD and a robot arm can occlude the face. When a camera is mounted on the robot arm, the sensing would be difficult because of a motion blur. Therefore, this study uses photoreflectors as proximity sensors. A Photoreflector has a light emitting diode and a phototransistor. A light emitting diode emits the light, and a phototransistor detects the intensity of the light reflected from a skin surface. The light intensity varies with the distance between the photoreflector and the object, so the phototransistor value can be converted to a distance. Photoreflectors can detect the distance in a close range, so they are attached to the end effector of the arms. On each four sides of the end effector, a photoreflector is arranged to detect the distance to the skin surface so that the shape around the end effector can be obtained. Photoreflectors detect points on the cheek surface and track the cheek surface (Fig. 5.2). Thus, the points on the cheek surface are collected. This study assumes that this local facial part can be represented

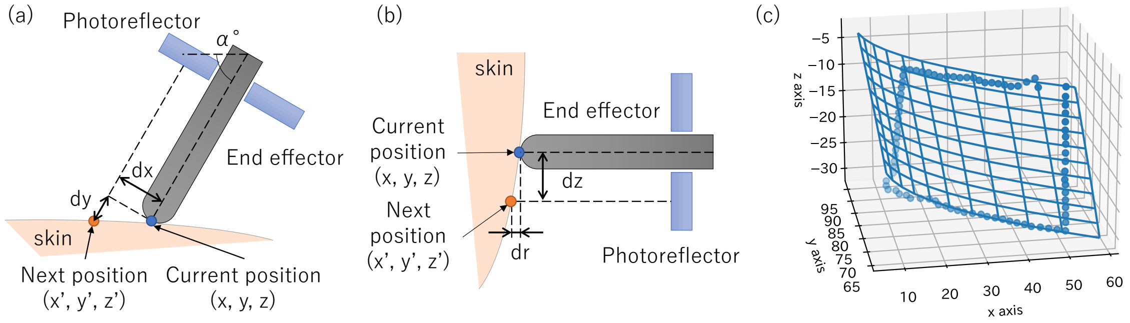


Figure 5.2: Cheek surface calibration. Left: Cheek surface tracking in horizontal direction. Center: Cheek surface tracking in vertical direction. Right: Fitted quadratic surface to 3D points on cheek surface. Blue markers were the actual points on the cheek surface. Wireframe was estimated surface.

as a quadratic surface. The quadratic surface is fit to the points on the cheek in order to obtain the cheek surface.

When presenting the haptic stimulation indicated a target direction, the direction is needed to map to a point on the cheek (Fig. 5.3). The target direction (azimuthal and elevational angle) is gotten by converting the Cartesian coordinate system to a spherical coordinate system. The azimuthal and elevational angles are transformed into a line equation. The elevational angle is converted to an offset of the height position, and the azimuthal angle is converted to the slope of the line. An intersection between the quadratic cheek surface and the direction represented as the line are calculated. This intersection is used as a stimulation point. However, the frontal direction cannot be presented because this study avoids touching our lips, which are located on the center of our face. Therefore, when the target positions in front of the user, two robot arms stimulated both cheeks at the same time. Thus, a directional cue is presented by touching on the position corresponding to the spatial information of the target.

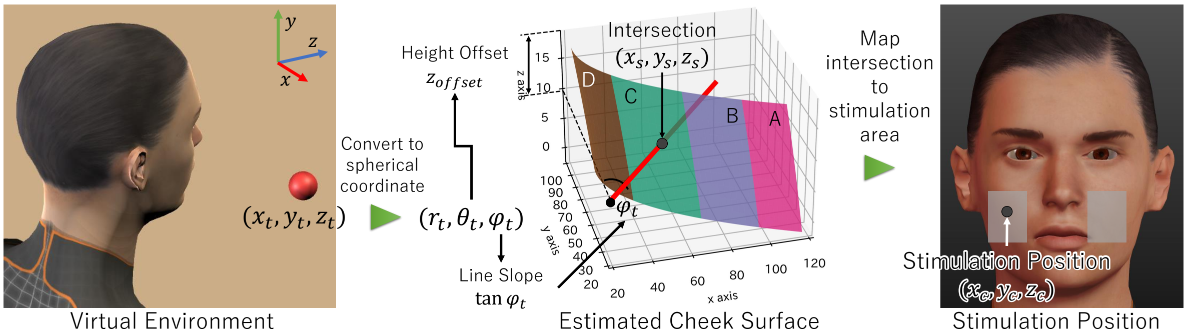


Figure 5.3: Haptic stimulation flow. The target position was converted to an azimuthal and an elevational angle. These two angles were translated into a line. The intersection between the line and the cheek surface was calculated and was mapped to the stimulation area (transparent orange area) to obtain the stimulation position.

5.4 Implementation

This study developed a system to provide directional cues on the cheek (Fig. 5.4). Our system consisted of a haptic stimulation device (Fig. 5.5) and software. The device presents haptic directional cues on the cheek using two robotic arms. Section 5.4.1 describes the details of the device. The software controlled robot arms, detected the cheek surface, translated a direction to a point on the cheek. Software and hardware were integrated to VE developed with Unity. These functions were described in Section 5.4.2, Section 5.4.3, and Section 5.4.4 respectively.

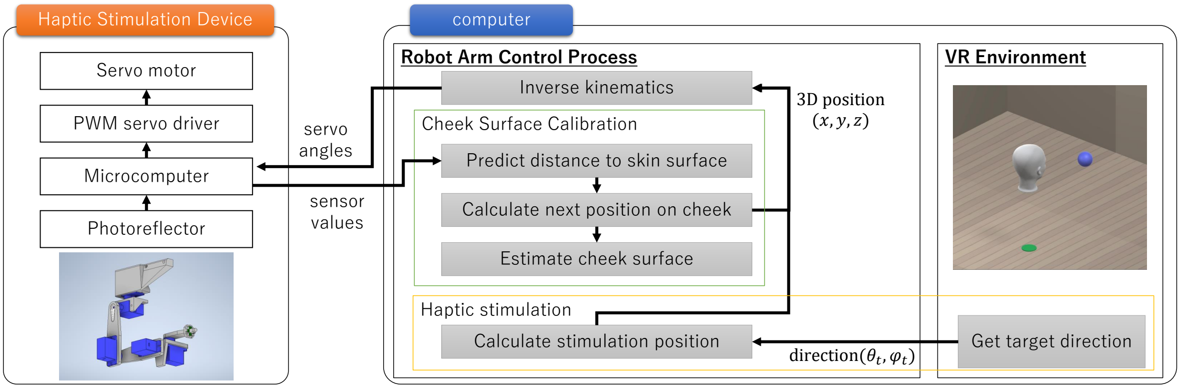


Figure 5.4: System Configuration

5.4.1 Haptic Stimulation Device

A haptic stimulation device was built by modifying an Oculus Rift CV1 (Fig. 5.5). Two robotic arms and a circuit were attached. The robotic arms were attached to the left and right sides of the bottom of the HMD so that they stimulated each side of the cheeks. The robotic arms were fixed by using brackets created with a 3D printer. The circuit was attached to the HMD's head strap and mounted on the top of the HMD.

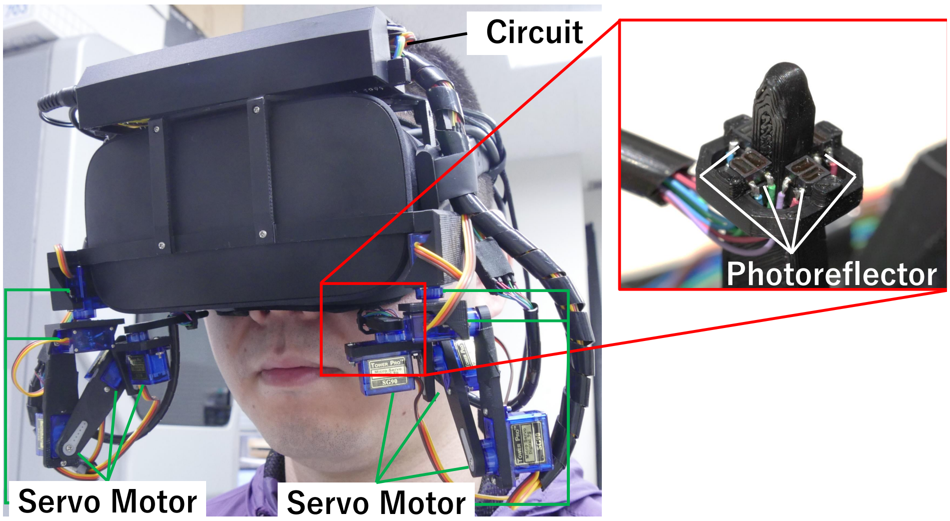


Figure 5.5: Haptic Stimulation Device

The robotic arm was designed with 5 DoF (degree of freedom), allowing to stimulate the cheek in 3D space. Fig 5.6 illustrates the configuration of the robot arm attached to the left side of the HMD. The left and right robotic arms were constructed symmetrically. The 3 Dof of the robotic arm controlled a position in the 3D space, and the rest 2 Dof controlled the end effector posture. Controlling the end effector posture allowed the end effector to adjust the angle approaching to the cheek and the photorefletors on the end effector to measure the cheek within close range. Five servo motors (SG-90, Tower Pro) were used on the joints for each robot arm and placed one at each joint. Each servo angle was $\theta_0, \theta_1, \theta_2, \theta_3, \theta_4$ in Fig. 5.6. The links of the robotic arm were created by a 3D printer. The one side of the link was fixed to a servo motor with servo horn; the other side was fixed to another servo motor with screws. The tip of the arm was rounded to avoid hurting the cheeks. The end effector had four photoreflectors on the right, left, up, and down sides (Fig. 5.5).

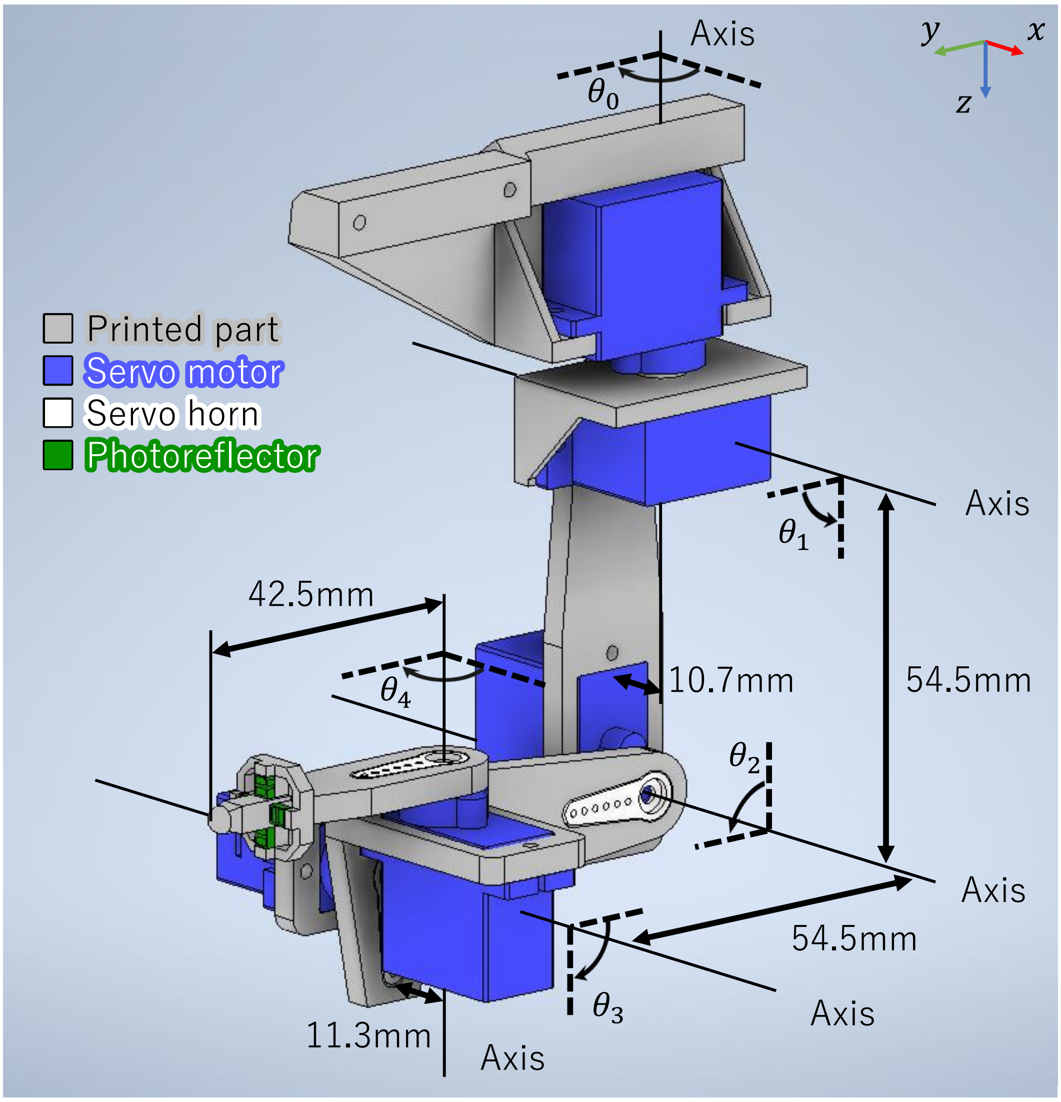


Figure 5.6: Robotic Arm Configuration

The circuit had a microcomputer (Arduino Nano) and a pulse width modulation (PWM) servo driver (PCA9685, NXP). The microcomputer was connected to a USB cable. The microcomputer received instructions from the software via serial communication, rotated servo motors, and sent sensor values of photoreflectors. The microcomputer communicated the PWM servo driver with I2C. The servo driver control servo rotation. Servo motors supplied the power by 5V AC adapter. The microcomputer got photoreflector values via an analog multiplexer (TC4052BP, TOSHIBA). The microcomputer was fixed to a bracket created by a 3D printer and attached to the top of HMD through the headset strap. The cables were covered by black tubes.

The robotic arm weighted 70g (SG-90 9g × 5 + link 25g). The total weight of two robotic arms and the circuits including cables was 298g (robotic arm 70g × 2 + circuits including cables 158g), which was lighter than the system presented in previous study [121] (405g). Therefore, our system was light but allowed to stimulate the left and right sides of the cheek simultaneously.

5.4.2 Robot Arm Control

Inverse kinematics was leveraged to move the end effector of the robot arms to a 3D position (x_g, y_g, z_g). Three servo angles, which corresponded to a 3D position of the specific arm part, were calculated. Based on the three angles, the rest two servo angles, which controlled the end effector posture, were calculated. By controlling five servo motors as the above, the end effector was allowed to face a constant direction to the head. This calculation enables the photoreflectors to detect the reflection intensity according to the distance from a constant direction when measuring the cheek surface. First, the azimuthal angle of the end effector θ_E and the elevational angle of the end effector θ_A were used to calculate the third servo motor position (x_2, y_2, z_2) with Equation 5.1.

$$\begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} = \begin{pmatrix} x_g \\ y_g \\ z_g \end{pmatrix} - (\mathbf{R_y}(\theta_E) \cdot \mathbf{l_3} + \mathbf{R_x}(\theta_A) \cdot \mathbf{l_4}) \tag{5.1}$$

$\mathbf{l_3}, \mathbf{l_4}$ was the third link length and the fourth link length, respectively. $\mathbf{R_x}(\theta), \mathbf{R_y}(\theta)$ represent the rotation matrices that rotate around x axis and y axis by θ degrees, respectively. The servo angles $\theta_0, \theta_1, \theta_2$ corresponding to the position (x_2, y_2, z_2) were computed with Inverse Kinematics. Then, the rest two servo angles θ_3, θ_4, that controlled the end effector posture, was calculated with Equation 5.2 and Equation 5.3.

$$\theta_3 = \theta_1 + \theta_2 - \theta_E \tag{5.2}$$

$$\theta_4 = -\theta_0 + \theta_A \tag{5.3}$$

In our implementation, θ_E and θ_A were set to 0 degree and 150 degrees, respectively.

5.4.3 Cheek Surface Detection

The mapping between real and VR space was calibrated to encode a virtual target into the position on the cheek surface. For the cheek surface calibration, the 3D points on the cheek surface were detected and collected. By fitting a quadratic surface to the collected 3D points, the cheek surface was estimated. Calibration was performed for each arm because of the asymmetry of the facial geometry. Also, only the width and height of the stimulation area was set because the facial geometry depends on each person.

In collecting the point on the cheek surface, each robot arm traced the edge of a rectangular area for stimulation. At first, the robot arm touched the cheek surface in a straight line from the initial position. While the arm is moving, the photoreflectors measured the distance from the arm tip to the cheek surface and converted the sensor values into a distance in the real world with a linear regression model. Then, the arm moved along the cheek edge of the stimulation area in the left, down, right, and up directions in sequence. When the arm moved to the left, a photoreflector on the left side of the arm tip measured the distance to the cheek surface and our system computed a point on left side of the arm tip with formula 5.4 (Fig. 5.2, left).

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} -\sin\alpha & \cos\alpha & 0 \\ \cos\alpha & \sin\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} dx \\ dy \\ 0 \end{pmatrix} \tag{5.4}$$

dx was 1.5 because the photoreflector width was about 3.0mm, and α was 60 degree as described in Section 5.4.1. The arm iterated to move the detected point until the arm reached the left edge of the stimulation area. When the arm moved down, a photoreflector located on the bottom side of the arm tip measured the distance to the cheek surface, and our system calculated the position of a point on the lower side of the arm tip using formula 5.5 (Fig. 5.2, center).

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} \cos\beta & -\sin\beta & 0 \\ \sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ dr \\ dz \end{pmatrix} \tag{5.5}$$

dz was decided to 1.5 thanks to the photoreflector width, and β was the angle of first joint of the arm. The arm iterated to move the detected point until the arm reached the bottom edge of the stimulation area. A similar procedure as above was performed when the arm moved to the right and up direction. Thus, the arm moved on the stimulation area. By gathering the points of the arm trajectory, 3D points of the cheek surface were acquired.

The cheek surface was estimated by fitting a quadratic surface to collected 3D points (Fig. 5.2, right). The quadratic surface was fitted to collected points to compute the parameters of the quadratic surface with the least-square method (equation 5.6). This way, the cheek surface was obtained as the quadratic surface. The cheek surfaces were estimated for both left and right arms, and each side of the surface was estimated separately.

$$ax^2 + by^2 + cx + dy + e = z \tag{5.6}$$

5.4.4 Haptic Stimulation

The haptic directional cues were presented with the robotic arm, changing the stimulation according to the spatial information of the target (Fig. 5.3). Directional information was obtained from the target position (x_t, y_t, z_t) and computed the corresponding cheek position (x_s, y_s, z_s). Based on the cheek position (x_s, y_s, z_s), the position (x_c, y_c, z_c) and the method of the cheek stimulation were determined.

In calculating the point on the cheek (x_s, y_s, z_s) corresponding to the target's position (x_t, y_t, z_t), the position information was converted into directional information in VR space. The target's position (x_t, y_t, z_t) was converted to the position from user's view (x_{tu}, y_{tu}, z_{tu}). By converting the position in Cartesian coordinate (x_{tu}, y_{tu}, z_{tu}) to the position in Spherical coordinate $(r_t, \theta_t, \varphi_t)$, an azimuthal angle φ_t and an elevation angle θ_t were calculated. The azimuthal and elevation angle (φ_t, θ_t) was mapped to the point on a curved surface (x_s, y_s, z_s). The sine of the elevation angle θ_t was mapped to the height of the cheek coordinates z_{offset} (formula 5.7).

$$z_{offset} = \begin{cases} z_{min} & (\theta_t < 0) \\ \frac{\sin\theta_t}{sin\theta_{max}} * (z_{max} - z_{min}) & (0 \leq \theta_t < \theta_{max}) \\ z_{max} & (\theta_t \geq \theta_{max}) \end{cases} \tag{5.7}$$

In our implementation, θ_{max} was 45 degrees, z_{min} was 0, and z_{max} was 30. The azimuthal angle φ_t was translated to a line representation and added the height offset z_{offset} (formula

5.8, t is constant).

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = t \begin{pmatrix} \tan \varphi_t \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ z_{offset} \end{pmatrix} \tag{5.8}$$

The intersection between the estimated cheek surface (formula 5.6) and the line (formula 5.8) was computed. In this way, the cheek position (x_s, y_s, z_s), that mapped to the target position in VR space (x_t, y_t, z_t), was calculated.

Since the calculated point (x_s, y_s, z_s) could be out of the stimulation area, the stimulation method and the stimulation position (x_c, y_c, z_c) were decided based on x_s. x_s was classified into four areas; area A was $x_f < x_s$; area B was $x_{c_{max}} \leq x_s < x_f$; area C was $x_{c_{min}} \leq x_s \leq x_{c_{max}}$; area D was $x_s < x_{c_{min}}$. x_f was the threshold of the facial center area. $x_{c_{max}}$ and $x_{c_{min}}$ were the maximum and minimum value of the stimulus range respectively. Based on which area x_s was, the stimulation position was calculated as follows:

$$(x_c, y_c, z_c) = \begin{cases} (x_{c_{max}}, y_{x_{c_{max}}, z_s}, z_s) & (x_s \text{ in area A or B}) \\ (x_s, y_{x_s}, z_s) & (x_s \text{ in area C}) \\ (x_{c_{min}}, y_{x_{c_{min}}, z_s}, z_s) & (x_s \text{ in area D}) \end{cases} \tag{5.9}$$

$y_{x_{c_{max}}, z_s}$ was the y value in equation 5.6 when $x = x_{c_{max}}$ and $z = z_s$. $y_{x_{c_{min}}, z_s}$ was the y value in equation 5.6 when $x = x_{c_{max}}$ and $z = z_s$. If x_s was in area A, both arms were used for the stimulation and calculated the stimulation positions on both sides. If not, the single robot arm was leveraged. The stimulus of azimuthal direction was discrete, but that of elevational direction was continuous. It was expected that the continuous stimulation in the elevational plane would help users to recognize the height information accurately. Finally, the stimulation position (x_c, y_c, z_c) was decided and was touched with the robot arms.

5.5 Experiment1: Directional Guidance

This experiment investigated how haptic cues on the cheeks provided directional information in VR space. In this experiment, participants looked for and pointed the invisible target in 3D space with haptic cues on cheeks. The targets appeared at fifteen-degree intervals of 180 degrees in the azimuthal plane and of 90 degrees in the elevation plane, so the appeared location was totally 91 (13 (azimuth) * 7 (elevation)). When the target appeared, the robot arms began to stimulate the participant's cheeks. The stimulation was provided depending on the gap between the participant's direction and the target direction. In the VR system, the only head rotation was reflected, while the head position was fixed. Participants pointed the position by turning to the target direction and by pressing a controller button. A reticle was placed in front of participants to help them to select the target. Instruction texts were shown to participants in the VR space. This experiment evaluated the absolute azimuthal and the absolute elevational angular error and task completion time. There were six participants (six males; age was M=23.5, SD=0.96), and they had no disability related to tactile organs. All participants had experienced using VR (> 10 times), and two of them played VR less than an hour a week and the rest of them played VR more than an hour a week.

Our experiment had two sessions, namely, a rehearsal session and an actual performance session. In the rehearsal session, participants looked for the target with visual and haptic feedback. The rehearsal session had 45 trials, and one target appeared in each trial. There was a five seconds interval between each trial. Participants were instructed to look at the front during the intervals. During the intervals, the stimulation stopped and the robot arms returned to the initial position. When participants selected the correct target, they heard the audio feedback for the correct answer. In the actual performance session, participants searched the invisible target with haptic feedback. The session had 91 trials. When participants selected the direction, they heard the audio feedback. The participants were instructed to wear earplugs and a noise canceling headphone and let the participant hear white noise during the experiment to reduce the noise from robot arms and to hear only the sound from the VE.

5.5.1 Experimental Protocol

The experimenter first explained the task and the procedure of the experiment to the participants. The experimenter instructed the participant to put on the device and earplugs and sit him on a chair. Then, the experimenter calibrated the cheek surface of the participant. The experimenter instructed the participant not to move during the calibration. After calibration, the experimenter displayed a VE and presented visual and haptic information of the target position to the participant. Here, the experimenter instructed the participant to move the head up, down, left, and right to demonstrate haptic stimulation when the target was located in front. After the demonstration, the experimenter put a noise canceling headphone on the participant and the rehearsal session started. In the rehearsal session, the participant searched for the visible target 45 times. Immediately after the rehearsal session, an actual performance session started. In the actual performance session, the participant searched for the invisible target 91 times. After the actual performance session, the participant removed the earplugs, the headphone, and the device. The participants sit to the seat during the experiment.

5.5.2 Result

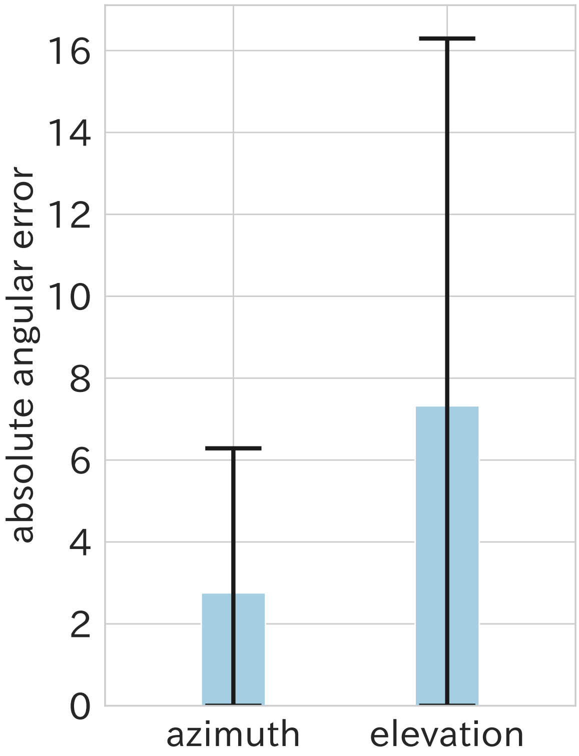


Figure 5.7: Absolute azimuthal and absolute elevational angular error in Experiment 1.

The result is shown in Fig. 5.7 and Fig. 5.8. The absolute azimuthal angular error was M=2.76 degrees, SD=3.53 degrees. On the other hand, the absolute elevation angular error was M=7.32 degrees, SD=8.97 degrees (Fig. 5.7). The task completion time was M=7.42 seconds, SD=7.64 seconds. Also, Fig. 5.8 shows histograms of both angular error respectively. Fig. 5.9 indicates an example of temporal changes of the gap between the target direction and head one. According to the histograms of both directions, almost azimuthal angular errors were around 10 degrees. It is because both cheeks were stimulated when the target was within 10 degrees of the azimuthal angle in our implementation.

On the other hand, elevation angular error was larger than azimuthal one. It was expected the elevation error was similar to the azimuthal one because continuous stimulation was provided on the cheek as described in Section 5.4.4. However, it seemed to be difficult to recognize the elevation angle as the result showed. As shown in Fig. 5.9, the participants repeatedly shook

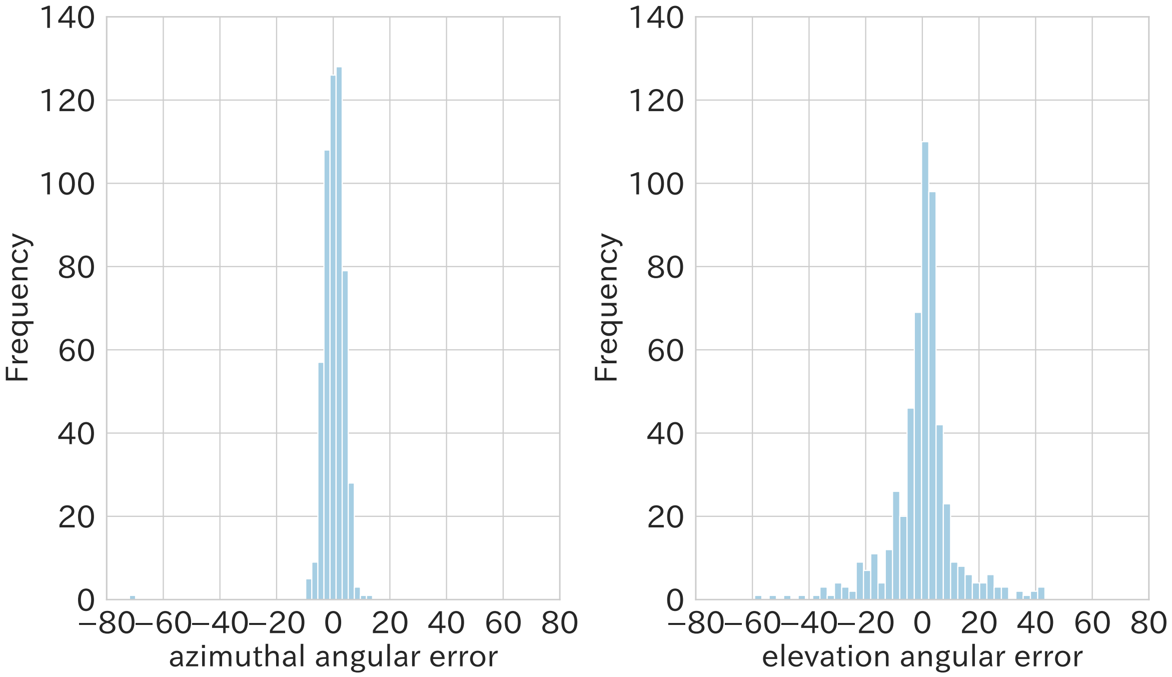


Figure 5.8: *Left*: A histogram of azimuthal angular error. A bar width indicated the four degree. *Right*: A histogram of elevational angular error. A bar width indicated the four degree.

their heads in elevational direction. After the experiment, all participants reported that they had difficulty finding the elevational direction. Furthermore, one participant reported that he lost the origin of height direction of haptic feedback during the experiment. It is considered that our feedback of elevational direction has room for improvement, as the simple continuous feedback could result in ambiguity.

After the experiment, some participants mentioned that they sometimes did not feel the haptic feedback. This was caused by a slight change in the position of the device as the participant moved the head. In our system, the robot arms were fixed to the HMD, so when the HMD moved, the distance to the cheek also changed. When the participants looked down, the HMD could move due to the effect of gravity. Therefore, a method to adaptively estimate the cheek surface is required.

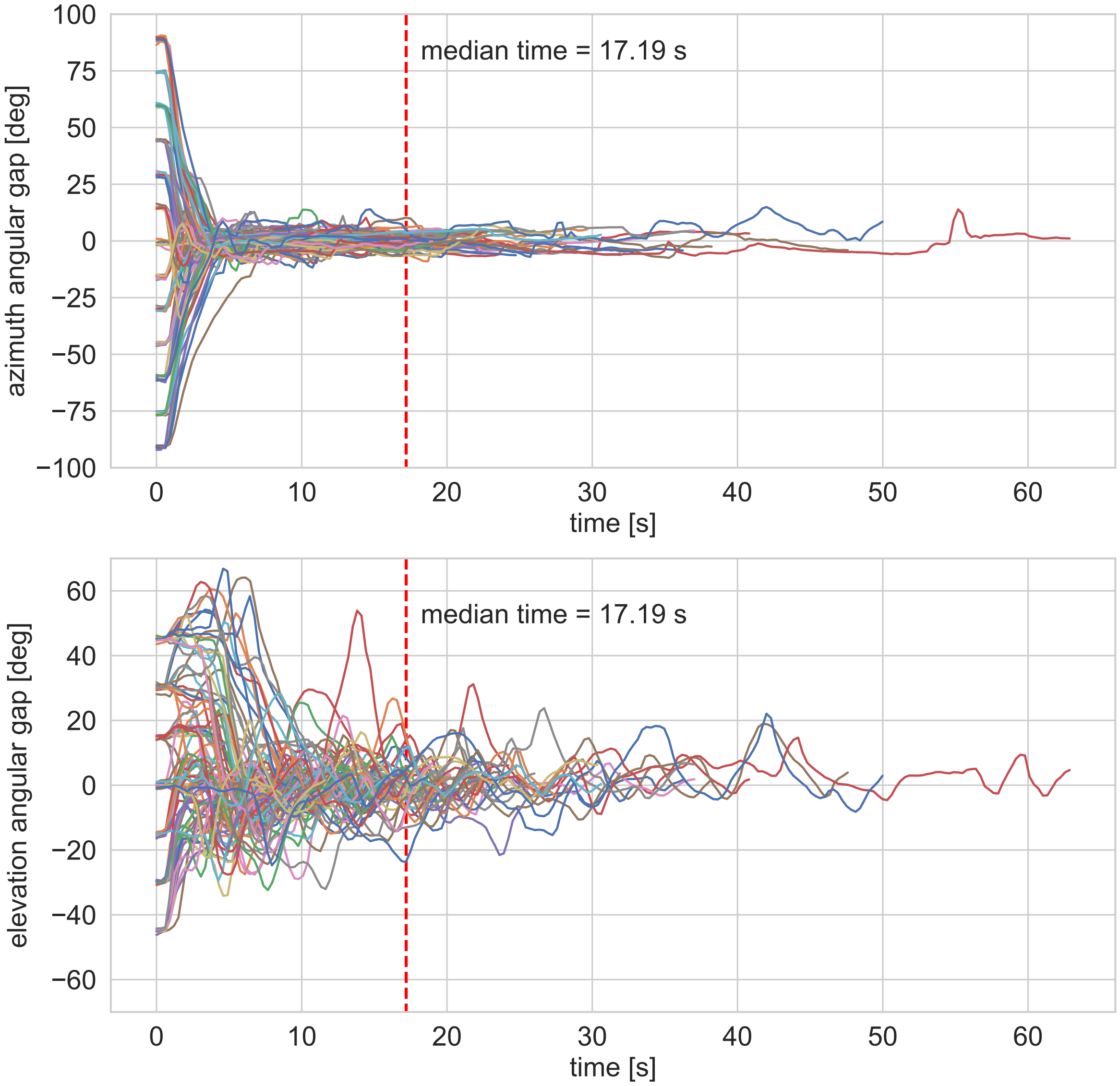


Figure 5.9: Example of temporal change of azimuthal and elevation angular gap between the center point of a participant view and target position in Experiment 1. These contained 91 trials. The red line indicated a median of task completion time.

5.6 Experiment 2: Spatial Guidance

This experiment investigated how our spatial guidance technique was effective in VR space. Participants had to look for and touch the visible target in 3D space. This experiment compared task completion time in visual + haptic condition with in visual condition. In visual condition, participants searched the target with only visual information. In the visual + haptic condition, participants detected the target with visual information and haptic feedback to cheeks. There were six participants (six males; age was M=23.5, SD=0.96), and they had no disability related to tactile organs and vision. All participants had experienced using VR (> 10 times), and two of them played VR less than an hour a week and the rest of them played VR more than an hour a week.

Participants searched for and found targets of a specific color among several spheres. Participants performed 80 trials and took a break after every 20 trials. Each target ball was placed in a room (3 m x 3 m x 2 m) with 50 balls in it. The target positions were placed in randomly chosen positions. The target positions were used to generate 40 random locations and then randomly sorted to generate a set of counterbalanced positions totaling 80 points. The other 49 fakes were randomly placed at the beginning of each trial. In this case, each ball was placed so that they did not overlap. The colors of the targets and fakes were randomly set when the trial started. The color of each ball was set so that the distance between its colors in HSV space was greater than 0.25. When the balls appeared, the robot arms began to stimulate the participant's cheeks. The stimulation was provided depending on the gap between the participant's direction and the target direction. In this experiment, the position and rotation of the head were reflected in the VR environment. The example of the room for the experimfent is shown in Fig 5.10.

5.6.1 Experiment Protocol

Our experiment had two sessions, namely, a rehearsal session, an actual performance session. Before the actual performance session, the experimenter let participants train to find the target in each condition. In the rehearsal session, participants looked for the target with each condition. The session had 20 trials and one target appeared in each trial. There was a 5 seconds interval between each trial. During the interval, the cheek stimulation stopped and the robot arms returned to their initial position. Participants were instructed to look in front of them while waiting for the next trial. When participants selected the correct target, they heard the audio feedback for the correct answer. In the actual performance session, participants searched the visible target with haptic feedback. The session had 80 trials and the participants took a break after every 20 trials. The experimenter instructed the procedure to the participants before the experiment.

This experiment flow is shown below:

1. The participant was instructed to go to the center of the room. The center was indicated as a circle.
2. The participant was shown the target to find and was heard a sound.
3. The participant looked for the target. The target to find was located in the front of participants.
4. When the participant touched the target, the trial was finished.

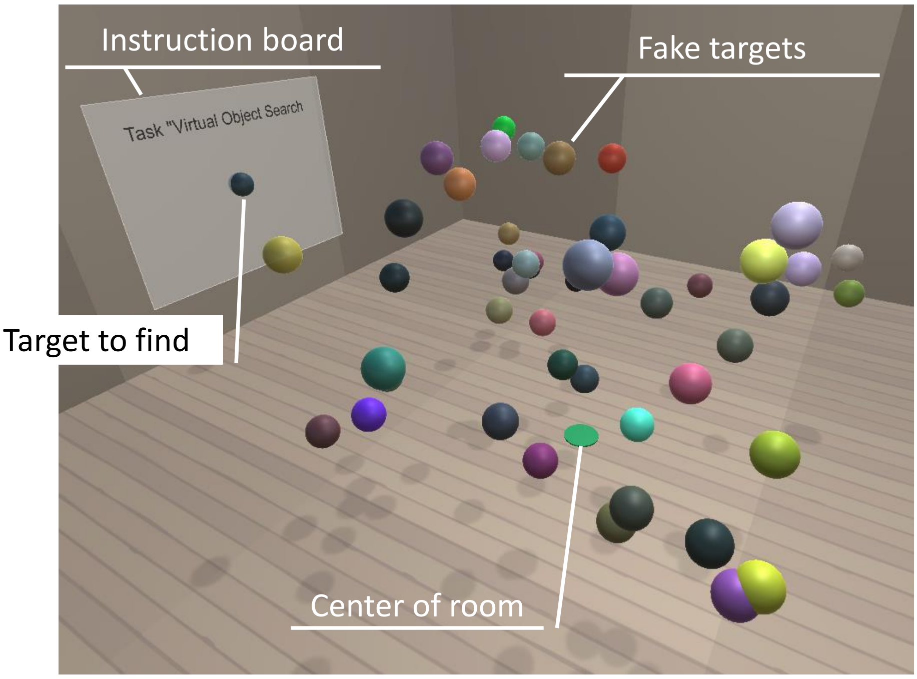


Figure 5.10: Virtual Environment of Experiment 2

5.6.2 Result

The overall average result is shown in Fig. 5.18 and the individual results are shown in Fig. 5.12. The task completion time in visual condition was M=12.45 s, SD = 14.51 s. The task completion time in visual and haptic condition was M = 6.91 s, 5.48 s. Friedman test revealed that there was a significant difference (p=0.014) between visual condition and visual+haptic condition in terms of the task completion time.

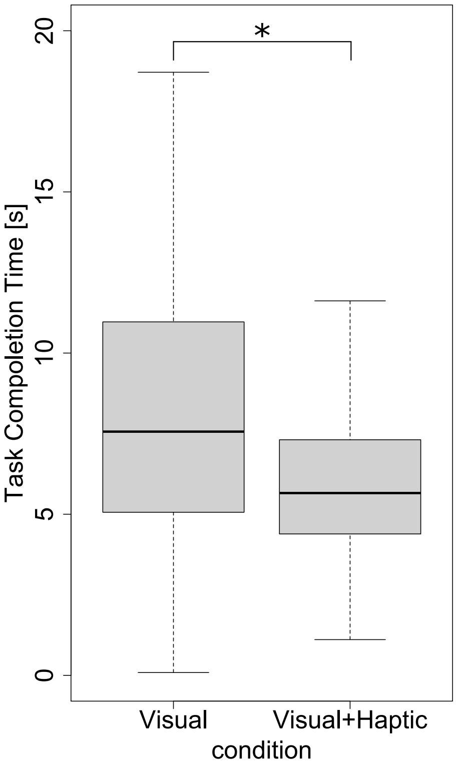


Figure 5.11: Box Plot of Task Completion Time in Visual and Visual+Haptic condition

Looking at the participants' behavior during the experiment, the Visual+Haptic condition resulted in faster behavior when exploring. However, the search took longer when the target was at the foot of or above the target. In the Visual + Haptic condition, participants searched horizontally from their field of view first, and therefore could not locate the target located under their feet. They acted like he was looking around, so they were able to find them quickly. This is because the azimuthal direction was the easiest to find in our direction presentation.

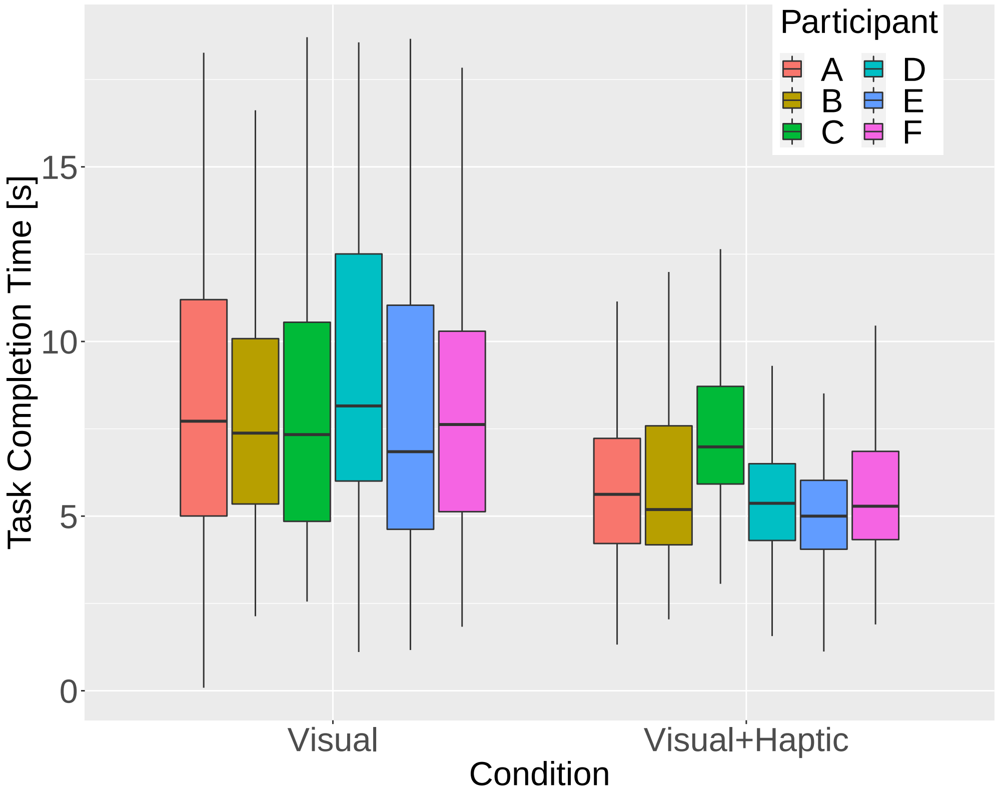


Figure 5.12: Box plot of Task Completion Time in Each Condition for Each Participant

Also, participants were confused when the targets were hidden by fakes from the participant view. Since our system does not present information about the depth direction, this may be because of an inconsistency between the visual and haptic information when occlusion occurs.

After the experiment, some participants reported that they changed their behavior in the Visual and Visual+Haptic conditions. In the Visual condition, the participants learned the target's color without making mistakes, but in the Visual+Haptic condition, they learned the rough color of the target and looked for it by haptic stimulation. This is because the Visual condition provided no directional information and some colors were difficult to identify, while the Visual+Haptic condition provided directional cues. Also, in the Visual+Haptic condition, some participants did not feel the stimulation when they changed their facial expressions after successfully locating the target. This is an issue with our haptic presentation method using cheek surface estimation. The calibration was performed only once before stimulating the cheek surface and was not do so afterwards. However, the cheeks change their facial shape, such as when making facial expressions. Especially in the region near the lips, the haptic stimulation seemed to disappear because the tactile stimulus position was around there when the target was in front.

5.7 System Improvement

Our system was improved in terms of device, cheek surface estimation, and experimental design to investigate the potential of cheek haptic-based spatial guidance method further. In the device improvement, the HMD was replaced and the robotic arms was reconstructed to increase the rigidity. As for the cheek surface estimation, the end effector posture was controlled to face perpendicular to the cheek surface in collecting points on the cheek to measure the surface more appropriately. Also, the protocol of the experiment on directional guidance was modified and the evaluation of the experiment on spatial guidance was added.

5.7.1 Improvement of Device

The cheek haptic stimulation device was improved by modifying an Oculus Quest 2 and reconstructing robotic arms (Fig. 5.13). Oculus Quest 2 had built-in cameras to capture physical environment. Therefore, robotic arms and circuits were attached out of the camera view.

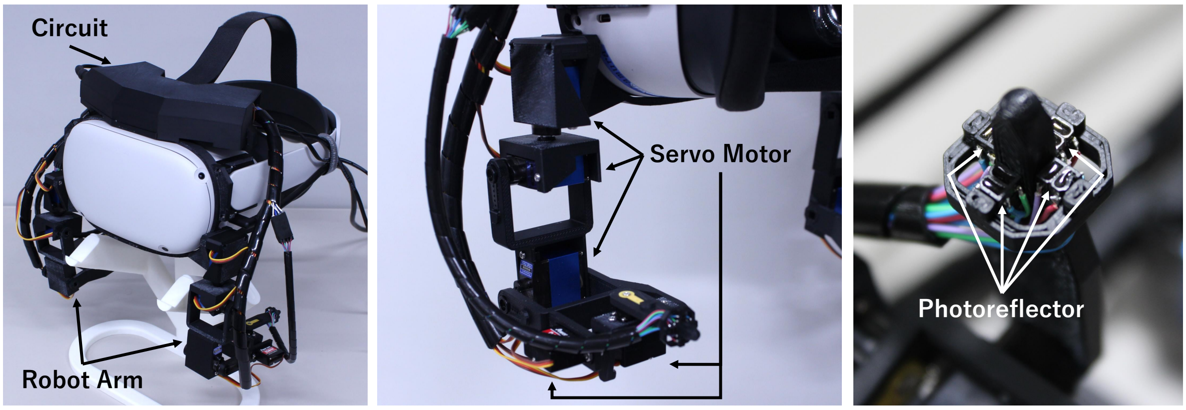


Figure 5.13: Improved Cheek Haptic Stimulation Device. *Left*: Device Overview. *Center* Robotic Arm. *Right* Photoreflectors on the End Effector.

Also, the robotic arms, which attached to the HMD's left and right sides, were reconstructed. Although previous robotic arms (Fig. 5.6) could stimulate the cheek, the arm had issues on the durability such as weak link supports and plastic-geared servo motors. Therefore, the structure was strengthened and the plastic-geared servo motors were replaced to metal-geared servo motors. Fig. 5.14 illustrates the improved robotic arm configuration and θ_n indicates the rotation angle of *n-1*th joints (e.g. if it is 1st joint, the rotation angle is θ_0). Five servo motors of two kinds of metal-geared servo motors were used. Three of five were Tower Pro MG92B servo motors that had a high torque, the other two were PowerHD DSM44 servo motors that were lighter than MG92B servo motor. MG92B servo motors were installed on the first, second, and third joints, requiring high torque performance. DSM44 were installed on the fourth and fifth joints that controlled the end effector posture in azimuthal and elevational plane because the joints required low torque. By combining two kinds of servo motors according to the requirements of joints, the overall weight of the robotic arm was reduced. On the arm tip, four photoreflectors were located as well as the our previous robotic arms. The circuit was the same as the previous device but the bracket was remade with a 3D printer.

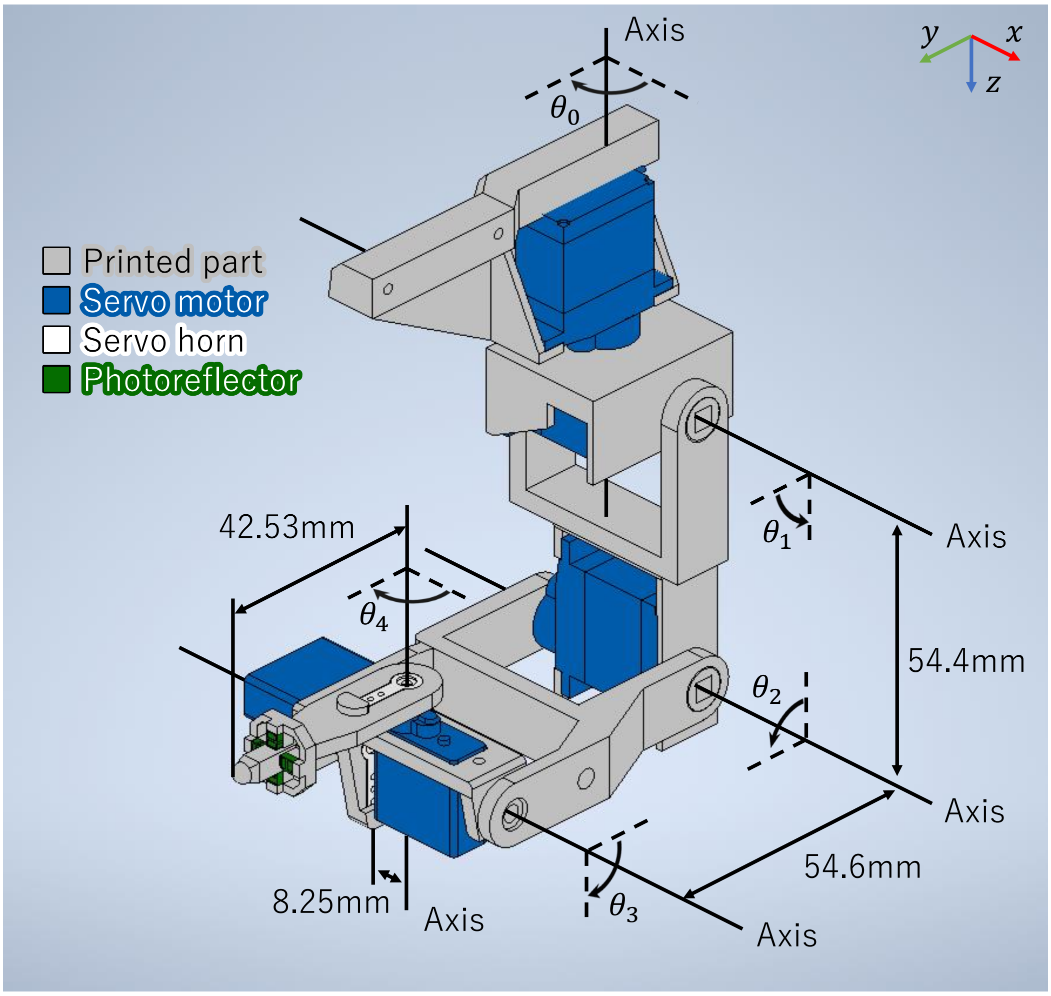


Figure 5.14: Improved Robotic Arm Configuration

The improved robotic arm weighted 82g (MG92B servo 14g × 3 + DSM44 servo 6g × 2 + link 28g) and the circuit weight was 165g. So, the total weight of two robotic arms (82g × 2) and the circuit (165g) was 329g. In providing haptic stimulation, the arm stimulated the cheek with a weak force (around 0.4N). There was a latency of about 140 ms from the control signal generated in the virtual environment to robotic arm actuation.

5.7.2 Improvement of Cheek Surface Estimation Method

The distance prediction method, horizontal cheek surface tracking method, and tracking path were modified to make the cheek surface estimation better.

Photoreflector detects the light intensity, and the sensor values correspond to the distance. In our current implementation, a linear regression model was used in converting the values to the distance in the physical world. This is because, within the limited range, the relation between the values and the distance can be regarded as a linear regression. However, the values increase or decrease nonlinearly according to the distance and the previous study [108] showed the relation between the values and the distance can be modeled a power-law model. Therefore, a power-law model ($y = ax^b$) was adopted to translate the sensor values to the distance.

The horizontal cheek surface tracking was improved to measure various facial geometry. Facial geometry varies from person to person. Especially, the cheek surface has a large curvature in a horizontal plane. In our current implementation, the end effector faces at a constant angle during the tracing of the cheek surface, and the photoreflectors can be too far from the cheek skin. It can cause the photoreflectors to mispredict the distance to the cheek due to their close sensing range. Therefore, the horizontal normal direction was detected, and turn the end effector was turned perpendicular to the cheek surface to keep the distance within an appropriate range while tracing the cheek (Fig. 5.15).

Here, the improved collecting procedure of points on the cheek is described. In collecting points on the cheek surface, the robot arm traced the cheek surface. At first, the robot arm moved straight from its initial position until it touched the cheek surface. While the arm was moving, the photoreflectors detected the distance from the arm tip to the cheek surface by converting the sensor values into an actual distance with a power-law model ($y = ax^b$). Then, the arm moved along the cheek surface in the following order: rightward, leftward, downward, rightward, and upward. At that time, the arm moved within a range $x_{s_{min}} < x < x_{s_{max}}, z_{s_{min}} < z < z_{s_{max}}$. When the arm moved to the right, the photoreflector on the right side of the arm tip measured the distance to the cheek surface, and the next position to move was computed using equation 5.4 (Fig. 5.2, left). The photoreflectors on the left and right sides detected the distance to the cheek surface, and then the gap between the normal and the current arm tip direction was calculated as follows: $\theta_{diff} = \arctan((dy_l + dy_r)/2 * dx)$.

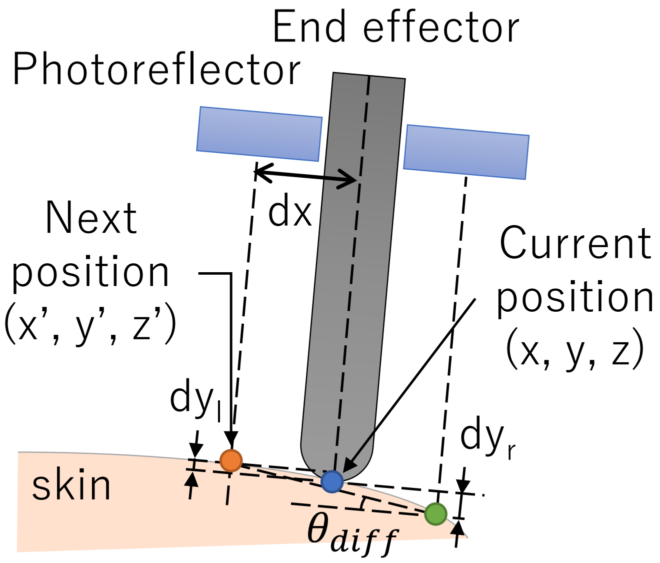


Figure 5.15: Improved Cheek Surface Tracking in a Horizontal Direction. The robotic arm detects the horizontal normal directions and faces to the cheek to measure the distance between the photoreflectors and cheek skin accurately.

The tip of the arm was rotated by the difference from the horizontal normal direction of the face. Then, the arm moved to the next position. The arm iterated the above procedure until the arm's position x reached the right edge of the stimulation area ($x_{s_{max}}$). Next, in case that the arm moved to the left direction, a similar procedure as described above was performed until the arm's position x reached the left edge of the stimulation area ($x_{s_{min}}$). When the arm moved downward, the photoreflector located on the bottom side of the arm tip measured the distance to the cheek surface, and the next position to move was calculated using formula 5.5 (Fig. 5.2, center). The arm iterated to move the detected point until the arm's position z reached the bottom edge of the stimulation area ($z_{s_{min}}$). A similar procedure as above was performed when the arm moved to the right and up direction. Thus, the arm moved along the cheek surface. By collecting the points on the arm trajectory, the points of the cheek surface were acquired.

5.8 Experiment 3: Directional Guidance with Improved Protocol

This experiment evaluated our improved system in a similar experiment as Experiment 1 in terms of directional guidance. The rehearsal session of the experimental protocol was changed. Although the rehearsal session in Experiment 1 showed the target visibly, the participants might learned the relation between the visible target and the cheek haptic stimulation. Therefore, invisible targets were shown with cheek haptic stimulation to let participants learn the cheek haptic stimulation that was encoded the direction.

This experiment investigated how accurately haptic cues on the cheeks provided directional information in VR space. In this experiment, participants searched and pointed invisible targets in a virtual environment with haptic cues on cheeks. The targets appeared at fifteen-degree intervals of 180 degrees in the azimuthal plane and of 90 degrees in the elevation plane, for a total of 91 locations (13 (azimuth) * 7 (elevation)). When the target appeared, the robot arms started to stimulate the participant's cheek. The robot arms provided the stimulation according to the participant's head posture and the target location. In the virtual environment, the head position was fixed, while the head rotation was reflected. Participants turned to the target direction and pressed a controller button to point the target direction. A reticle was placed in front of participants to help them to select the target. Instruction texts were shown to participants in the virtual environment. This experiment evaluated the absolute azimuthal and the absolute elevational angular error and task completion time. There were 18 participants (16 males, 2 females; age was M=24.28, SD=2.68), and they had no disability related to tactile organs. The participants were recruited from the students of the Faculty of Science and Technology of our university. Out of 18 participants, 14 had experienced using VR (>10 times), and the others had experienced using VR (>once). Nine of 18 participants played VR less than an hour a week, and the rest of them played VR more than an hour a week. The participants provided written consent form before the experiment.

Our experiment had two sessions, namely, a rehearsal session and an actual performance session. In the rehearsal session, participants looked for the invisible target with the help of haptic stimulation. The rehearsal session consisted of 45 trials, and one target appeared in each trial. There was a five seconds interval between each trial. During the interval, participants were instructed to gaze at the front, and the robot arms stopped stimulating. Participants selected the direction until they selected the correct one. If the error between the selected direction and the actual target direction was within 10 degrees in azimuthal and elevational angles, it was regarded as correct. When participants selected the correct target, they heard the audio feedback for the correct answer and showed the actual target for three seconds. Then, the trail moved on the next. In the actual performance session, participants looked for the invisible target by relying on the haptic stimulation as in the rehearsal session. The actual performance session consisted of 91 trials. When participants selected the direction, they heard the audio feedback, and the trial moved on to the next. The participants were instructed to wear the haptic stimulation device, earplugs, and a noise canceling headphone. During the experiment, white noise was played through the headphone to reduce the noise from the robot arms so that participants hear only the sound from the VE. This experiment followed the guideline provided by the research ethics committee at the Faculty of Science and Technology, Keio University.

5.8.1 Experimental Protocol

First, the participants were explained the task and the procedure of the experiment. The participant were instructed to put on the haptic stimulation device and earplugs and sit him on a chair. Then, the cheek surface of the participants were estimated. At that time, the participants were instructed not to move during the calibration. After the cheek surface estimation, the participants were shown the virtual environment and were demonstrated visual and cheek haptic information of the target position. Here, the participants were instructed to move the head up, down, left, and right to present haptic stimulation when the target was located in front. After the demonstration, the participants were put a noise canceling headphone. Then, the rehearsal session was performed. In the rehearsal session, the participant searched for the invisible target 45 times. Immediately after the rehearsal session, an actual performance session was performed. In the actual performance session, the participant searched for the invisible target 91 times. After the actual performance session, the participant removed the earplugs, the headphone, and the device. The participants sit to the chair during the experiment. This experiment took about an hour.

5.8.2 Result

The result is shown in Fig. 5.16. Some participants made unintended selections by touching the button. There were three such mistakes, and they were removed. The absolute azimuthal angular error was M=2.54 degrees, SD=2.19 degrees. On the other hand, the absolute elevation angular error was M=6.54 degrees, SD=9.34 degrees (Fig. 5.16 Left). The histograms of both angular errors respectively in Fig. 5.16 Center, Right. The task completion time was M=13.01 seconds, SD=6.78 seconds. An example of temporal changes of the azimuthal and elevational angular gap between the target direction and head one is shown in Fig. 5.17. Compared with Experiment 1, the absolute elevational angular error was less. This improvement was brought by the modification of the rehearsal session. The histograms of both directions indicate that almost azimuthal and elevation angular errors were within 10 degrees. It is considered because, in the rehearsal session, the gap within 10 degrees of both angles was defined as the correct answer.

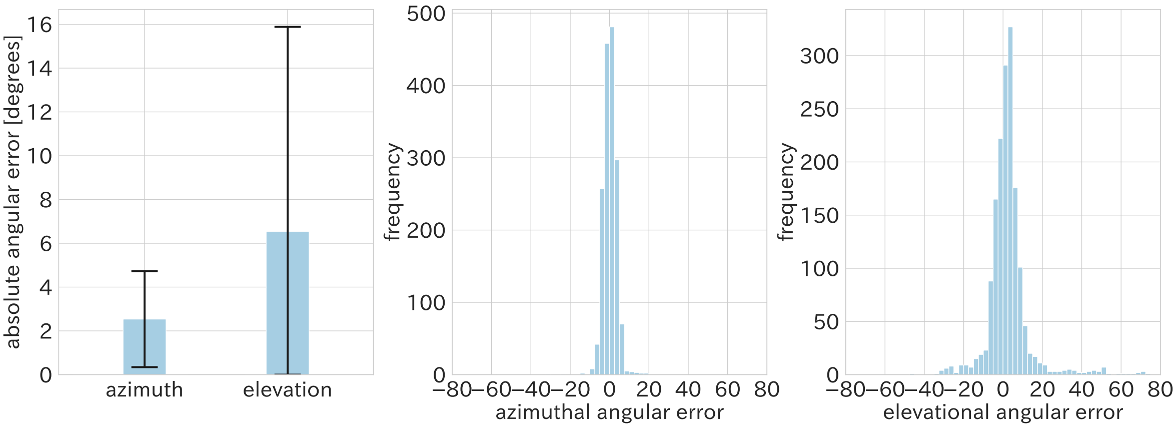


Figure 5.16: The result of directional guidance pointing accuracy. Left: absolute azimuthal and absolute elevational angular error. Center: a histogram of azimuthal angular error. A bar width indicated 2.5 degrees. Right: a histogram of elevational angular error. A bar width indicated 2.5 degrees.

Despite of the improvement of the absolute elevational angular error, similarly as the Experiment 1, the standard deviation of elevational angular error was larger than the azimuthal one. Also, according to Fig. 5.17, the participant, it seemed to be difficult to recognize the

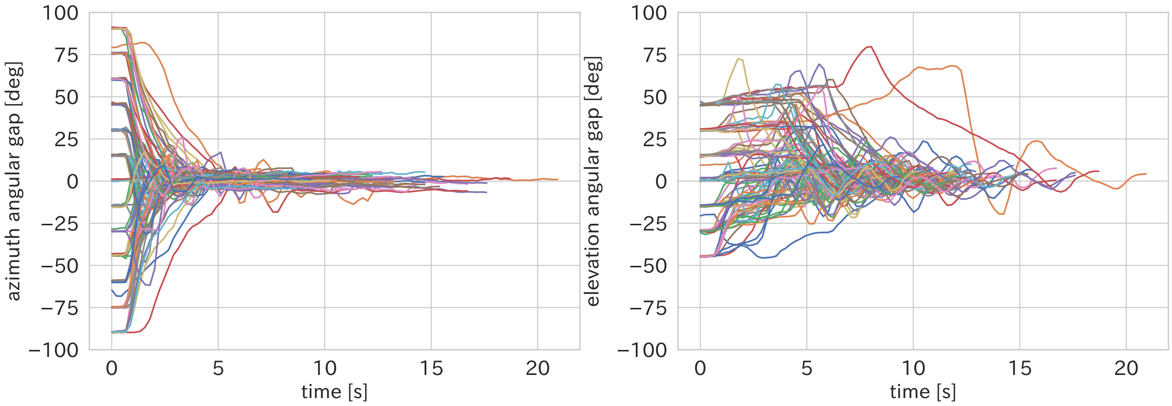


Figure 5.17: Example of temporal changes of azimuthal and elevation angular gaps between the participant's front and target direction in Experiment 1. 91 trials were plotted. Left: The temporal changes of the azimuthal angular gap. Right: The temporal changes of the elevational angular gap.

elevation angle intuitively. Fig. 5.17 Right indicates that the participants found the azimuthal direction soon. On the other hand, Fig. 5.17 Right shows that the participants repeatedly shook their heads in the elevational direction to find the accurate target direction. After the experiment, all participants reported that they had difficulty finding the elevational direction. Furthermore, one out of the participants reported that he could not recognize the base of the height of the stimulation, so he judged whether the stimulation stopped or not. Therefore, our haptic stimulation in elevational angle has room for improvement, as the simple continuous stimulation could result in unclearness.

5.9 Experiment 4: Spatial Guidance with Additional Evaluation

To evaluate our spatial guidance technique further, this experiment investigated how our technique affected the task performance, usability, and workload of spatial guidance in VR space. Participants had to look for and touch the visible target in virtual space. This experiment compared task completion time, System Usability Scale (SUS), and NASA-TLX score in visual+haptic condition with those in visual condition and visual+audio condition. In the visual condition, participants searched the target with only visual information (color). In the visual+audio condition, participants looked for the target with visual and audio cues. This experiment used a 440 Hz tone as the basis for the audio cues. The azimuthal and elevational angles were encoded into the difference between left and right sound and the sound frequency, respectively, since several studies modulated elevational angular information into auditory spectral cues to let users localize directions [150]. The sound frequency was transformed based on the difference between the target direction and the participant's frontal direction as follows: $f_t = -440 * abs(\theta_{diff_e})/90 + 440$, where f_t was sound frequency, θ_{diff_e} was the elevational angular difference between the target direction and the participant's frontal direction. In the visual+haptic condition, participants found the target with visual information and haptic stimulation to cheeks. There were 18 participants (16 males, 2 females; age was M=24.28, SD=2.68) who had no disability related to tactile organs. The participants were recruited from students of the Faculty of Science and Technology of our university. The participants were the same at those who took part in Experiment 1. Of 18 participants, 14 had experienced VR more than ten times, and the rest had experienced VR more than once. Of the 18 participants, nine played VR less than an hour a week, and the others played VR more than an hour a week. In advance of the experiment, consent was gotten from the participants.

Participants looked for and detected targets of a specific color among many spheres (Fig. 5.10). Participants performed 80 trials and took a break after every 20 trials. In each trial, 50 spheres, including a target and 49 fakes, appeared in a room (3 m x 3 m x 2 m). The targets were placed in randomly chosen positions. 40 random positions were generated, and the position components (x, y, z) were sorted randomly to obtain 80 counterbalanced positions.

The other 49 fakes randomly appeared at the beginning of the trial. At that time, each sphere was placed without overlap. The sphere colors were chosen randomly so that the distance between each color was greater than 0.25 in HSV space when the trial started. When the spheres appeared, the robot arms started to stimulate the participant's cheeks. The robot arms provided the stimulation according to the participant's head posture and the target location. The head movements, including the position and posture, and the hand movements were reflected in the virtual environment. An protocol of this experiment followed the guideline provided by the research ethics committee at the Faculty of Science and Technology, Keio University.

5.9.1 Experiment Protocol

The experiment had two sessions, namely, a rehearsal session and an actual performance session. First, the rehearsal session was conducted to let participants train to find the target. In the rehearsal session, participants searched a target with each condition's cues. The session consisted of 20 trials, and a target appeared in each trial. There was a five seconds interval between each trial. During the interval, participants were instructed to look at the front, and the robot arms stopped stimulating. When participants selected the correct target, they heard the audio feedback. In the actual performance session, participants looked for the target with each condition's cues. The session had 80 trials, and there was a break after every 20 trials.

This experiment flow was shown as follows:

1. The participant was instructed to go to the center of the room. The center was indicated as a circle on the floor.
2. The participant was shown the target to find on the instruction board and was heard a sound.
3. The participant looked for the target. The target to find was shown on the instruction board.
4. When the participant touched the target, the trial was finished.

After each condition, the participants answered the SUS and the NASA-TLX questionnaires. At first, the participants answered ten questions about the SUS by rating a 5-point Likert scale, from 1 "Strongly Disagree" to 5 "Strongly Agree." Next, the participants rated six subjective subscales about the NASA-TLX scores, namely, "Physical Demand," "Mental Demand," "Temporal Demand," "Performance," "Effort," "Frustration." And then, for every pair of six subjective subscales (15 pairs), the participants answered which subscale was important for them in the task. Each condition was randomly conducted for each participant.

When this experiment was conducted, the order of the conditions was randomized for each participant to take a counterbalance of the condition order. The experimental procedure took about one and half hours.

5.9.2 Result

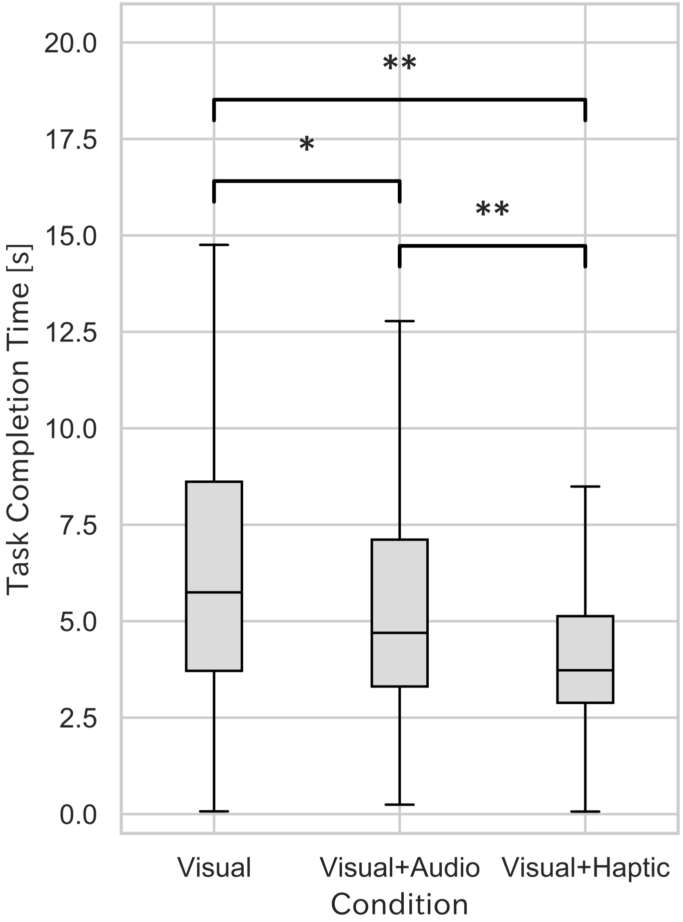


Figure 5.18: Box plots of task completion time in visual, visual+audio, and visual+haptic condition. * and ** indicates $p < 0.05$ and $p < 0.01$, respectively.

The overall average result of task completion time is shown in Fig. 5.18 Left and the individual results are shown in Fig. 5.18 Right. The task completion time in the visual condition was M=6.39 s, SD=3.34 s. The task completion time in the visual+audio condition was M=5.62 s, SD=3.12 s. The task completion time in the visual+haptic condition was M=4.35 s, SD=2.26 s. Friedman test revealed that there was a significant difference (p = 0.0000) in three conditions in terms of the task completion time. Post-hoc tests with Bonferroni correction found that there were significant differences between visual and visual+audio

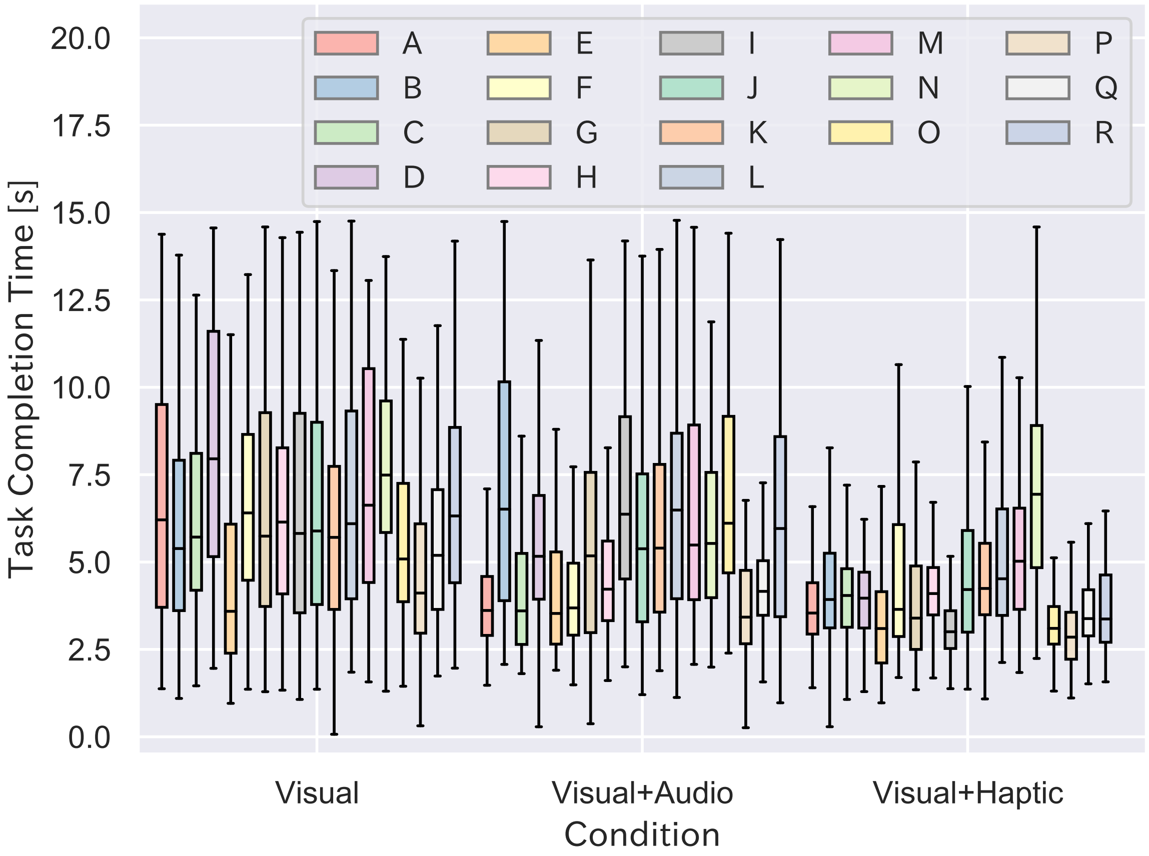


Figure 5.19: Box plots of Task Completion Time in Each Condition for Each Participant.

condition (p = 0.0300), visual and visual+haptic condition (p = 0.0000), and visual+audio and visual+haptic condition (p = 0.0012). The SUS score is shown in Fig. 5.20. In the visual condition, the SUS score was M=55.83, SD=20.40. In the visual+audio condition, the SUS score was M=47.78, SD=20.09. In the visual+haptic condition, the SUS score was M=80.42, SD=10.99. Friedman test showed a significant difference (p = 0.0002) in the three conditions in terms of the SUS score. Post-hoc test with Bonferroni correction found siginificant differneces between visual and visual+haptic condition (p = 0.0002), and visual+audio and visual+haptic condition (p = 0.0000). The visual+haptic condition using our guidance technique achieved the highest SUS score among all conditions, and there were statistical significant differences between the other conditions. Also, the SUS score in visual+haptic condition reached the highest grade according to the guideline on SUS score interpretation. The NASA-TLX overall score is shown in Fig. 5.21, and each weighted NASA-TLX score

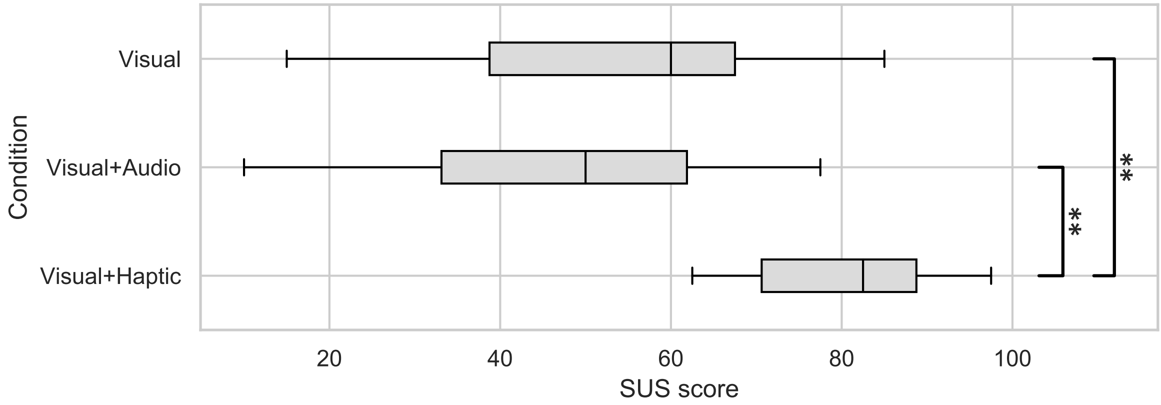


Figure 5.20: Box plots of SUS scores in each condition. ** indicates p < 0.01.

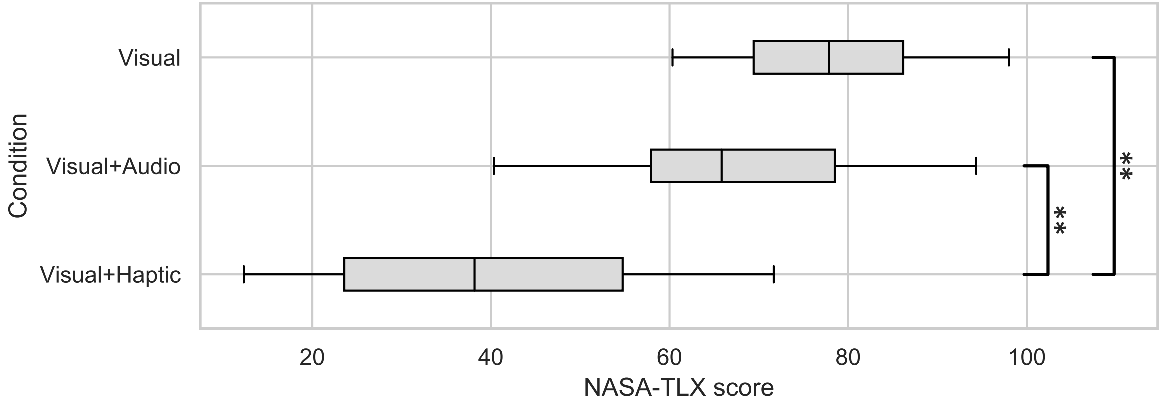


Figure 5.21: Box plots of total NASA-TLX scores in each condition. ** indicates p < 0.01.

are shown in Fig. 5.22. In the visual condition, NASA-TLX total score was M=75.81, SD=16.89. In the visual+audio condition, NASA-TLX total score was M=67.57, SD=14.96. In the visual+haptic condition, NASA-TLX total score was M=38.83, SD=18.52. Friedman test indicated a significant difference (p = 0.0000) in the three conditions regarding the workload. Post-hoc test with Bonferroni correction revealed significant differences between visual and visual+haptic condition (p = 0.0000), and visual+audio and visual+haptic condition (p = 0.0000). Mental Demand scores were M=16.07, SD=10.44 and M=19.33, SD=7.85 and M=9.31, SD=7.14 in the visual, the visual+audio, and the visual+haptic condition. Physical Demand scores were M=13.87, SD=9.58 and M=8.13, SD=7.00 and M=11.46, SD=8.49 in the visual, the visual+audio, and the visual+haptic condition. Temporal Demand scores

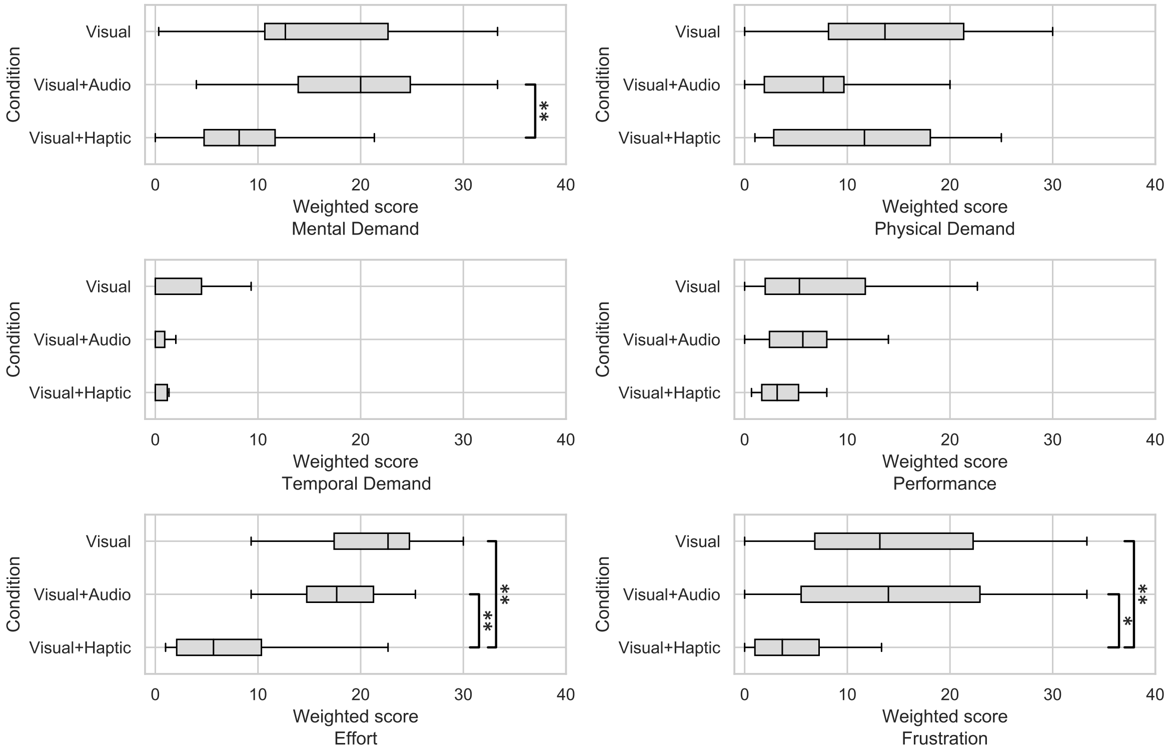


Figure 5.22: Box plots of each NASA-TLX weighted subjective subscale score in each condition. * and ** indicates p < 0.05 and p < 0.01, respectively.

were M=2.76, SD=4.62 and M=1.19, SD=2.50 and M=1.70, SD=3.24 in the visual, the visual+audio, and the visual+haptic condition. Performance scores were M=8.04, SD=8.26 and M=7.07, SD=6.90 and M=3.61, SD=2.15 in the visual, the visual+audio, and the visual+haptic condition. Effort scores were M=20.22, SD=6.94 and M=17.50, SD=5.22 and M=7.61, SD=6.82 in the visual, the visual+audio, and the visual+haptic condition. Frustration scores were M=14.85, SD=11.30 and M=14.35, SD=11.23 and M=5.13, SD=5.80 in the visual, the visual+audio, and the visual+haptic condition. As for the Mental Demand score, Friedman test indicated that there is a significant difference between the three conditions (p = 0.0098). Post-hoc tests with Bonferroni correction showed a significant difference between the visual+audio and the visual+haptic condition (p = 0.0010) in the Mental Demand score. In the Physical Demand score, Friedman test indicated that there is a significant difference between the three conditions (p = 0.0030). However, post-hoc tests with Bonferroni correction indicated no significant difference between any of the three conditions. As for the Effort score, Friedman test indicated a significant difference in the three conditions (p = 0.0000).

Post-hoc tests with Bonferroni correction revealed significant differences between the visual and the visual+haptic condition (p = 0.0079) and the visual+audio and the visual+haptic condition (p = 0.0117) in terms of Effort score. In the Frustration score, Friedman test showed a significant difference in the three conditions (P = 0.0049). Post-hoc tests with Bonferroni correction indicated a significant difference between the visual and the visual+haptic conditions (p = 0.0117), and visual+audio and visual+haptic conditions (p = 0.0079) in terms of Frustration score. Therefore, our spatial guidance technique achieved shorter task completion time, higher usability, and less workload in the target searching task.

Looking at the participants' behavior during the experiment, in the visual+haptic condition, the participants tended to turn to the target direction more quickly when searching. On the other hand, the participants seemed to have difficulty detecting the target around the foot or above the head. In the visual+haptic condition, participants first were likely to search the target horizontally from their view. Therefore, it is considered that participants had difficulty finding the target that was located out of their view and at different heights. If the target was around the eye-level, the participants identified the target easily and quickly. This implies that our technique presented the azimuthal direction appropriately. Participants were confused when fakes occluded the targets from the view of participants. Since our system does not provide depth information, such a lack of information might cause the participants to feel an inconsistency between the visual and haptic information when the targets were occluded by fakes.

After the experiment, some participants reported that they changed their behavior in visual+haptic conditions. In the visual condition, the participants made a great effort to learn the target's color without making mistakes. In the visual+haptic condition, they learned the audio cues, but they had difficulty learning the cues, so they eventually learned both color and audio cues. However, in the visual+haptic condition, they took less effort to learn the target's color and searched it relying on haptic stimulation. This is because the visual condition offered no directional information and the visual+audio condition made the participants learn the cues, while the visual+haptic condition provided directional cues intuitively. The Effort score of NASA-TLX also shows less effort of the visual+haptic condition in Fig. 5.22. In addition, according to the Frustration score of NASA-TLX in Fig. 5.22, the visual+haptic achieved less frustration than the other conditions. In the Mental Demand score, visual+audio condition achieved significantly lower score (5.22). Therefore, in the visual+haptic condition, less effort, less frustration, and less mental demand brought to the higher SUS score. Also, some participants reported that they lost the stimulation when they changed their facial expressions after successfully detecting the target in the visual+haptic condition. This loss of stimulation was caused by our cheek surface estimation method. The cheek surface was estimated only once before the stimulation. However, the cheek changes its shape, for example, when making facial expressions. Especially near the lips, the shape deforms significantly. Our system stimulated the region around the lips when the target was in the front, so it seems to lose the haptic stimulation that indicated the front direction.

5.10 Demonstration

In SIGGRAPH Asia 2021, our spatial directional guidance technique was demonstrated in a virtual environment. In the demonstration, two applications were implemented.

1. Users found a target image among 50 icons with the help of cheek haptics. The users were shown a target image on their right side for five seconds. Then, 50 icons appeared around the users. The users looked for the target among 50 icons relying on the cheek stimulation corresponding to the target direction and select the target with the controller. This application let the users perceive the directional information while sitting and concentrating on the cheek stimulation.

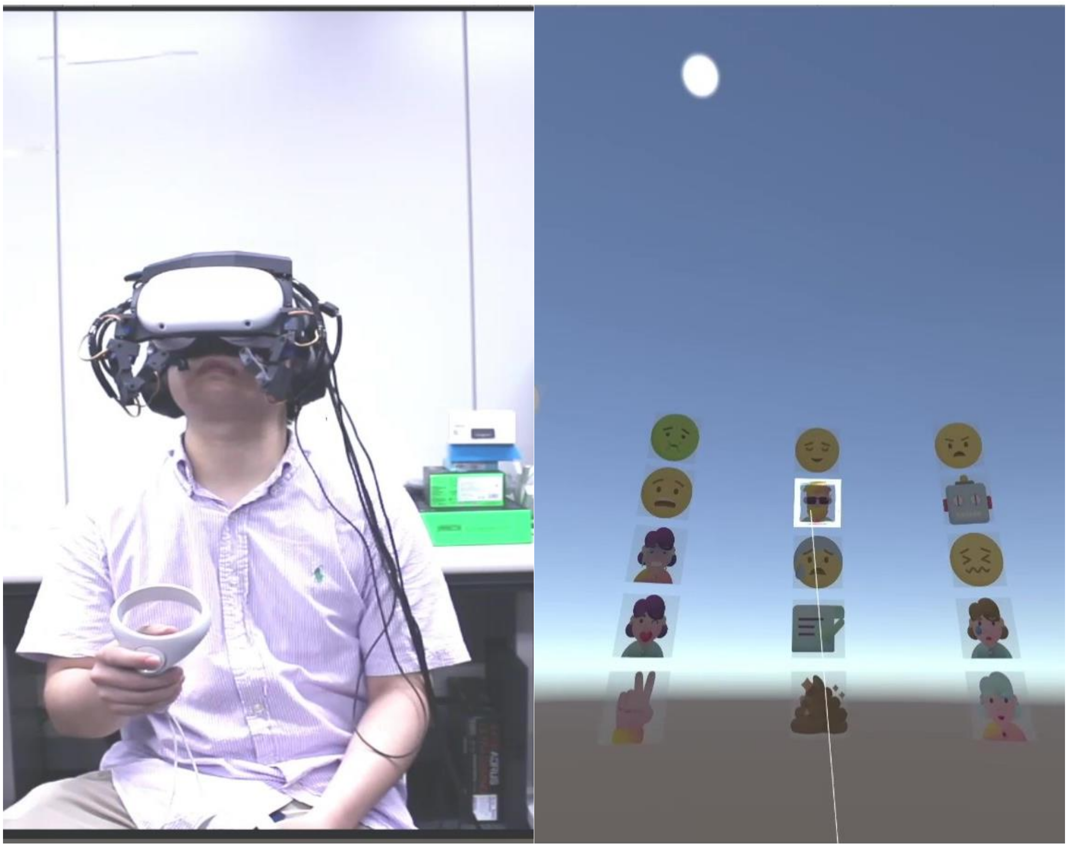


Figure 5.23: Image Finding Application. Users look for a target icon among many icons with the help of cheek haptic stimulation.

2. Users touched approaching targets relying on cheek haptic stimulation. A target appeared and approaches the users. The users searched the target with the cheek haptic stimulation and touch it. The controller position was represented as a white sphere in the virtual environment. Through this application, the users intuitively understood the target direction while standing and moving actively.

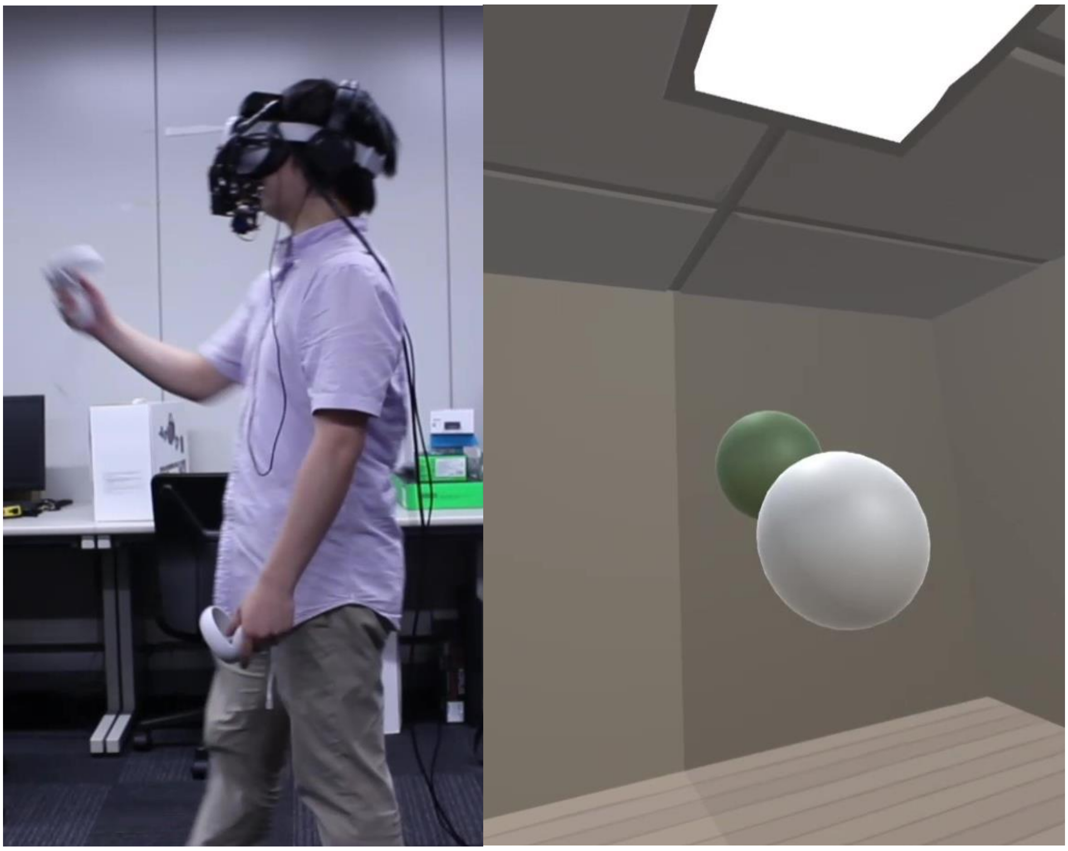


Figure 5.24: Target Touching Application. Users search and touch approaching targets relying on cheek haptic stimulation.

In SIGGRAPH Asia 2021, about 80 people tried our demonstration. Through the demonstration, it was shown that our cheek sensing and stimulation technology worked well on various people, such as people with beards and people wearing makeup.

5.11 Limitation

Our technique estimated the cheek surface as static to present the target direction. However, the cheek deforms when speaking, eating, and forming facial expressions, which can cause a robot arm not to reach the cheek. In Experiment 2, some participants mentioned that they lost cheek cues by changing facial expressions. Especially, the geometry around the mouth deforms complexly thanks to several facial muscles, making it difficult for the robotic arms to reach. In addition, if an HMD position is shifted, the estimated surface also is shifted so that the robot arm may not touch the cheek. Similar issues caused by the misalignment between the HMD and the face have been mentioned in previous studies on photoreflector-based sensing. Therefore, it is important to detect cheek surface in real-time to make the surface estimation robust to the positional drift and capture the dynamic facial deformation.

However, in case of real-time sensing, servo delays significantly affect the system responsiveness. Our system calculated a point of cheek surface based on the distance predicted with photoreflectors and robotic arm position. However, if our system get sensor data before arm actuation, our system mispredicts the positions, which causes wrong cheek surface estimation. Therefore, our current system waited for robot arm actuation for accurate cheek surface sensing. However, if sensing and stimulation are tried at the same time in our current system configuration, the stimulation is delayed by the latency of the sensing at least. In our current implementation, the used servo motors are controlled by PWM control width of 20ms. To actuate robotic arms more quickly, the high responsive servo motors are required.

In our spatial guidance experiment, elevational angular information were modulated into audio frequency to compare visual condition and visual+haptic condition. The elevational angle were translated into the pitch of audio cues. However, as the other method to modulate elevational information into frequency, head related transfer function-based methods were adopted [151, 152]. Therefore, by employing the head related transfer function-based approaches, the result may be changed.

There were young participants in both experiments. The sensory systems of human, such as hearing and haptics, decline as we age. However, in the haptic sensation, tactile capability is attenuated and, especially, in the lower limbs. Therefore, the experimental results may be changed by elderly participants.

Our implementation provided slight stimulation on the cheek to present spatial directional information in a safe manner. However, it is still unclear whether the slight stimulation is appropriate in offering spatial guidance or not. Therefore, it is important to investigate comfortable force for using the cheek as a haptic display.

In the demonstration of SIGGRAPH Asia 2021, it is confirmed that our facial sensing and stimulation technology worked well on most participants. However, in case of small faces, our cheek sensing failed because the calibration area was set with constant values in our implementation. The human face is different for each individual, and the calibration area should be set according to the face size of each person. Therefore, by considering face size, facial surface estimation could be applicable to more various people.

Our system employed servo motors for the robot arm. However, the angular step of the servo motor was 1.8 degrees. In the future, it will be tested whether using a high-resolution servo can make the stimulation more accurate.

A robotic arm with some linkages was leveraged as a stimulation device. However, the arm with linkages requires multiple joints with servo motors, which increases the weight of the arm. This weight increase makes the users tired. Therefore, it is necessary to construct a lightweight robot arm.

Facial surface was measured using photoreflectors to provide slight touch on the cheek. According to several experimental participants, provided haptic stimulation was very light. Therefore, even our current configuration is sufficient to present safe stimulation. By introducing force probes on the tip of robotic arms, the force on the cheek can be detected the actual force more precisely, which makes the haptic presentation safer.

The result of experiment 1 indicated that the elevational angular error was larger than the azimuthal one. While participants seemed to have difficulty moving their heads slightly, there is a possibility that the limitation of our robot arm-based haptic stimulation led to this result. However, it is unclear how precisely a human can control our heads in the elevational direction by ourselves. In the future, the controllable angle of the head by ourselves will be investigated.

Generally, continuous haptic stimulation can lose the sensation by sensory adaptation, while the participants did not report that in our experiments. It will be investigated if such adaptation occurs when stimulating the cheek with our method for a longer duration.

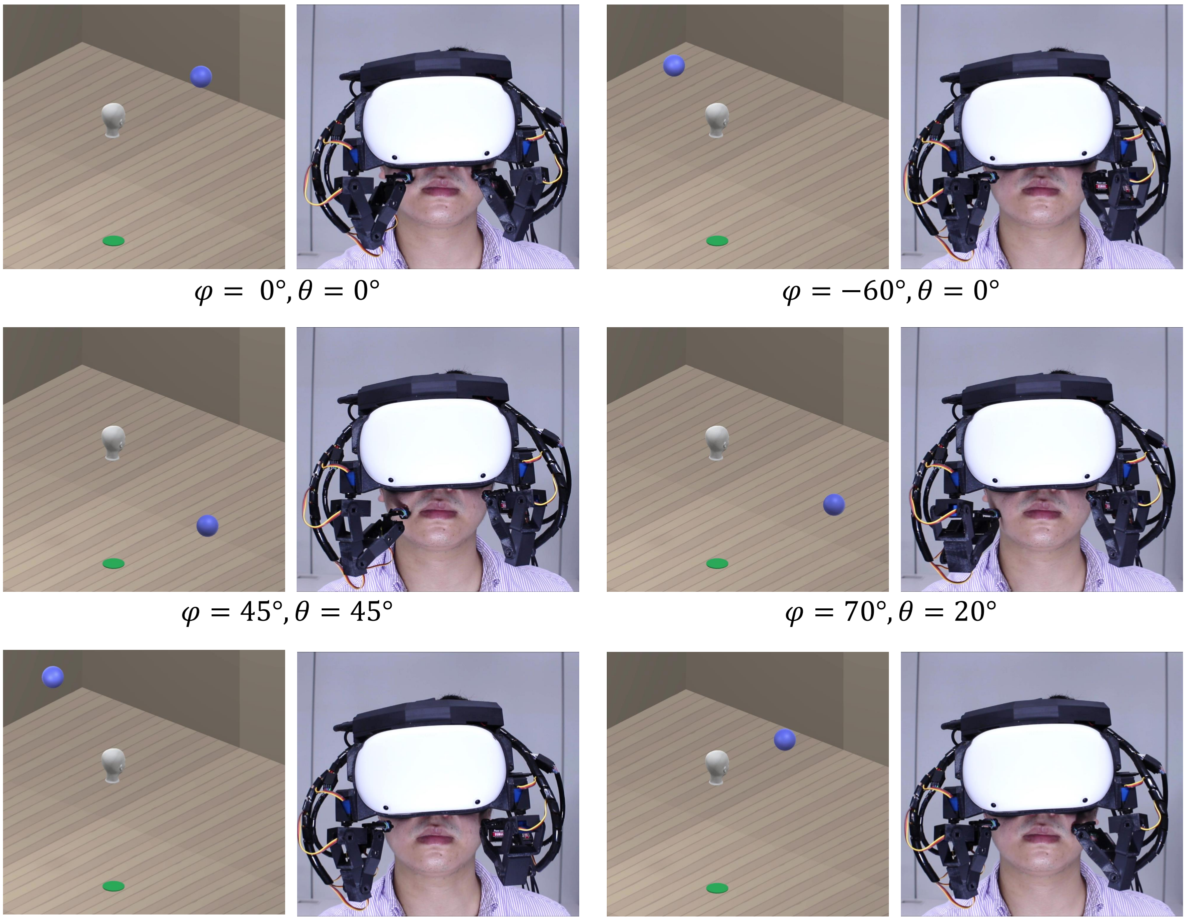
5.12 Conclusion

This study proposed Virtual Whiskers, a spatial directional guidance system that provided haptic cues on the cheeks with robot arms attached to an HMD. Robot arms were placed to the left and right sides of the bottom of an HMD, allowing to offer haptic cues on both sides of the cheek. Points on the cheeks' surface were detected using photo reflective sensors located on the robot arm tip and fitted a quadratic surface to the points to estimate the cheek surface. In stimulating the cheek, the point on the estimated surface that was encoded the targets' azimuthal and elevational angles in the VR space were calculated and the point was mapped to the stimulation position. According to spatial information of the target and user's head, the arms touched the cheeks to present directional cues.

The experiment on the directional guidance accuracy was conducted by haptic directional cues on the cheeks and evaluated the pointing accuracy. Our method achieved the absolute azimuthal pointing error of M=2.36 degrees, SD=1.55 degrees, and the absolute elevational absolute pointing error of M=4.51 degrees, SD=4.74 degrees. Also, the experiment on task performance in a spatial guidance task was conducted using our guidance technique. The experiment compared the task completion time, SUS score, and NASA-TLX score in the target searching task in the three conditions; only visual information was presented; visual and audio cues were presented; visual and haptic information was presented. The result showed that the averages of task completion time were M=6.68 s, SD=3.45 s and M=6.20 s, SD=3.30 s and M=4.22 s, SD=2.05 s in visual and visual+audio and visual+haptic condition, respectively. The averages of the SUS score were M=44.17, SD=21.83, and M=44.44, SD=23.24, and M=86.11, SD=8.76 in visual and visual+audio and visual+haptic condition, respectively. The averages of NASA-TLX scores were M=75.22, SD=22.93 and M=71.03, SD=12.13 and M=35.00, SD=17.16 in visual and visual+audio and visual+haptic condition, respectively. Statistical tests showed significant differences in task completion time, SUS score, and NASA-TLX score between the visual and visual+haptic condition and visual+audio and visual+haptic condition.

Through our experiments, this study showed the effectiveness of cheek haptics on spatial guidance. Our technology could be applied to collaboration, extended body control, and affective interaction. In collaboration with other people, it is important to share attention and interest. In such a case, our technique can indicate attention to users. In terms of human augmentation, by mapping extended body control and feedback to facial gestures and stimulation, the embodiment can be improved. As for affective interaction, communication via cheek haptics enriches interaction since the cheek haptic interaction improved remote communication quality [153], modified stress [87]. Our system is integrated into an HMD and makes the user's hands free. Therefore, cheek haptic-based interaction allows users to collaborate effectively.

In our system, the robot arms were used to present a single direction by stimulating the cheek. However, the several robot arms attached to the HMD enable tangible interaction on a cheek for immersive HMD users. In the future, it will be explored that the potential application and the interaction technique with virtual objects.



$\varphi = \ 0^\circ, \theta = 0^\circ$

$\varphi = -60^\circ, \theta = 0^\circ$

$\varphi = 45^\circ, \theta = 45^\circ$

$\varphi = 70^\circ, \theta = 20^\circ$

$\varphi = -70^\circ, \theta = -20^\circ$

$\varphi = -10^\circ, \theta = 10^\circ$

Figure 5.25: Haptic directional cue presentation with robotic arms attached to the HMD. In this figure, there are six example as pairs of the virtual environment (VE) and the haptic stimulation. The image on the left side in the pair was the VE. The 3D head model was the user's head position and the blue sphere was a target. The image on the right side was the user that stimulated the cheek with the robot arms attached to the HMD.

Chapter 6

Conclusion

This dissertation developed two HMD-based systems that enable interaction using a face. In a virtual environment, mouth shape recognition with photo reflective sensors embedded into an HMD and cheek haptic-based spatial directional guidance were presented.

In mouth shape recognition, the mouth shapes of HMD users were recognized with optical sensors. By integrating four photoreflectors, four optical distance measuring units, a microphone, and a microcomputer into an HMD, the prototype captured mouth shapes and audio signals. The photoreflectors and optical distance measuring units detected the movement of eight points (upper lip, upper cheek, lower lip, and cheek). A mouth shape classifier was built by training an SVM with the sensor values for each mouth shape. Also, five Japanese vowels were recognized from voice in order to label the sensor values and automate training data collection.

Our experiments indicated that our mouth shape recognition method classified six mouth shapes (the mouth shapes of five Japanese vowels and closed mouth shape) with an average accuracy of approximately 99.9%; the classifiers trained with our automated labeling method achieved an average classification accuracy of approximately 96.3%. In addition the experiment on user-dependency indicated that the user-independent classification accuracy was 30.7%, which was lower by 68.7% than user-dependent classification accuracy. Therefore, our current system is difficult to construct a generalized classifier. Therefore, it is difficult to construct a generalized classifier using the current system.

In addition, an application that projected the user's mouth shapes to an avatar was developed. In transferring the mouth shapes, six mouth shapes were blended using class membership probabilities for each shape as weights. The application revealed that our system could transfer various mouth shapes on the avatar.

In spatial directional guidance using facial haptics, directional cues were provided on the cheeks using robotic arms attached to an HMD. Robotic arms were attached to the left and right sides of the bottom of an HMD to offer haptic cues on both sides of the cheek and placed photo reflective sensors on the tip of the arm as proximity sensors. Points on the cheek surface were collected using photo reflective sensors and fitted a quadratic surface to the points to estimate the cheek surface. To stimulate the cheek, the point on the estimated surface corresponding to the target azimuthal and elevational angles in the VR space was calculated and the point was mapped to the stimulation area.

The experiment on directional guidance accuracy by cheek haptics resulted in the absolute azimuthal pointing error of M=2.36 degrees, SD=1.55 degrees, and the absolute elevational absolute pointing error of M=4.51 degrees, SD=4.74 degrees. The experiment on spatial guidance also conducted to investigate task performance, usability, and workload using our guidance technique. Participants were imposed a target searching task. The experiment compared task completion time, SUS score, and NASA-TLX score in the three conditions. The first condition (visual condition) showed only visual information; the second condition (visual+audio condition) provided visual and audio cues; the third condition (visual+haptic condition) presented visual and haptic information. The result indicated that the averages of task completion time were M=6.68 s, SD=3.45 s in the visual condition; and M=6.20 s, SD=3.30 s in the visual+audio condition; and M=4.22 s, SD=2.05 s in the visual+haptic condition. The averages of the SUS score were M=44.17, SD=21.83 in the visual condition; and M=44.44, SD=23.24 in the visual+audio condition; and M=86.11, SD=8.76 in the visual+haptic condition. The averages of NASA-TLX scores were M=75.22, SD=22.93 in the visual condition; and M=71.03, SD=12.13 in the visual+audio condition; and M=35.00, SD=17.16 in the visual+haptic condition. Statistical tests revealed significant differences in task completion time, SUS score, and NASA-TLX score between the visual and visual+haptic conditions and visual+audio and visual+haptic conditions.

Our system presented a single direction by stimulating the cheek using robotic arms. However, the robot arms can provide feedback from the virtual environment and enable tangible interaction on a cheek for immersive HMD users. In the future, the potential application and interaction technique with virtual objects will be explored.

Through the above studies, the sensing and stimulation technologies were developed for facial surface interaction. The region around the mouth and cheek was focused on. Our studies revealed that embedded optical sensors detected facial surfaces and embedded robot arms provided stimulation successfully. Especially, this facial haptic stimulation study provided direct stimulation on the face precisely by using robotic arms. It is considered that this direct facial stimulation method expanded the possibilities of facial haptics in virtual and physical environments.

Our facial surface sensing and stimulation methods help to investigate the effectiveness of facial haptics on human augmentation and avatar embodiment. Especially, our framework is strongly beneficial in applications that provide vision, audio, and haptic cues, like VR applications.

Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor, Professor Maki Sugimoto, for supporting researches from Bachelor to Ph.D. course, for his patience and invaluable guidance. His incomparable supports helped me in all researches. I am grateful to him for letting me proceed with the research appropriately. I would like to thank the rest of my thesis committee too: Professor Hideo Saito, Associate Professor Yuta Sugiura, and Associate Professor Jean-Marie Normand.

In addition to my supervisor, I would like to thank Project Associate Professor Yuta Itoh at The University of Tokyo for giving great advices to my research. And I thank all members of Sugimoto Laboratory for the collaboration.

Reference

[1] Wilder Penfield and Theodore Rasmussen. *The cerebral cortex of man; a clinical study of localization of function.* Macmillan, 1950.

[2] Ivan E Sutherland. Sketchpad a man-machine graphical communication system. *Simulation*, 2(5):R–3, 1964.

[3] Hiroshi Ishii and Brygg Ullmer. Tangible bits: towards seamless interfaces between people, bits and atoms. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 234–241, 1997.

[4] Ivan E. Sutherland. A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, AFIPS '68 (Fall, part I), page 757–764, New York, NY, USA, 1968. Association for Computing Machinery.

[5] Oculus. https://www.oculus.com/.

[6] VRChat. https://hello.vrchat.com/.

[7] Jonghwa Kim, Stephan Mastnik, and Elisabeth André. Emg-based hand gesture recognition for realtime biosignal interfacing. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, IUI '08, page 30–39, New York, NY, USA, 2008. Association for Computing Machinery.

[8] Rui Fukui, Masahiko Watanabe, Masamichi Shimosaka, and Tomomasa Sato. Hand shape classification in various pronation angles using a wearable wrist contour sensor. *Advanced Robotics*, 29(1):3–11, 2015.

[9] Yasha Iravantchi, Mayank Goel, and Chris Harrison. *BeamBand: Hand Gesture Sensing with Ultrasonic Beamforming*, page 1–10. Association for Computing Machinery, New York, NY, USA, 2019.

[10] Granit Luzhnica, Jorg Simon, Elisabeth Lex, and Viktoria Pammer. A sliding window approach to natural hand gesture recognition using a custom data glove. In *2016 IEEE*

Symposium on 3D User Interfaces (3DUI), pages 81–90, 2016.

[11] Oliver Glauser, Shihao Wu, Daniele Panozzo, Otmar Hilliges, and Olga Sorkine-Hornung. Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Trans. Graph.*, 38(4), jul 2019.

[12] Daniel K.Y. Chen, Jean-Baptiste Chossat, and Peter B. Shull. Haptivec: Presenting haptic feedback vectors in handheld controllers using embedded tactile pin arrays. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–11, New York, NY, USA, 2019. Association for Computing Machinery.

[13] Sebastian Günther, Florian Müller, Markus Funk, Jan Kirchner, Niloofar Dezfuli, and Max Mühlhäuser. Tactileglove: Assistive spatial guidance in 3d space through vibrotactile navigation. In *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*, PETRA '18, page 273–280, New York, NY, USA, 2018. Association for Computing Machinery.

[14] Sidhant Gupta, Dan Morris, Shwetak N. Patel, and Desney Tan. Airwave: Non-contact haptic feedback using air vortex rings. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, page 419–428, New York, NY, USA, 2013. Association for Computing Machinery.

[15] Graham Wilson, Thomas Carter, Sriram Subramanian, and Stephen A. Brewster. Perception of ultrasonic haptic feedback on the hand: Localisation and apparent motion. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 1133–1142, New York, NY, USA, 2014. Association for Computing Machinery.

[16] Cristy Ho and Charles Spence. Head orientation biases tactile localization. *Brain research*, 1144:136–141, 2007.

[17] Masaaki Fukuoka, Adrien Verhulst, Fumihiko Nakamura, Ryo Takizawa, Katsutoshi Masai, and Maki Sugimoto. Facedrive: Facial expression driven operation to control virtual supernumerary robotic arms. In *SIGGRAPH Asia 2019 XR*, SA '19, page 9–10, New York, NY, USA, 2019. Association for Computing Machinery.

[18] Alan Transon, Adrien Verhulst, Jean-Marie Normand, Guillaume Moreau, and Maki Sugimoto. Evaluation of facial expressions as an interaction mechanism and their impact on affect, workload and usability in an ar game. In *2017 23rd International Conference on Virtual System Multimedia (VSMM)*, pages 1–8, 2017.

[19] Steven Hickson, Nick Dufour, Avneesh Sud, Vivek Kwatra, and Irfac Essa. Eyemotion:

Classifying facial expressions in vr using eye-tracking cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1626–1635, Jan 2019.

[20] Jun Rekimoto, Keishiro Uragaki, and Kenjiro Yamada. Behind-the-mask: A face-through head-mounted display. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, AVI '18, New York, NY, USA, 2018. Association for Computing Machinery.

[21] Shih-En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. Vr facial animation via multiview image translation. *ACM Trans. Graph.*, 38(4), July 2019.

[22] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. Facial performance sensing head-mounted display. *ACM Trans. Graph.*, 34(4):47:1–47:9, July 2015.

[23] Ho-Seung Cha, Seong-Jun Choi, and Chang-Hwan Im. Real-time recognition of facial expressions using facial electromyograms recorded around the eyes for social virtual reality applications. *IEEE Access*, 8:62065–62075, 2020.

[24] HTC VIVE PRO EYE. https://www.vive.com/jp/product/vive-pro-eye/overview/.

[25] HP Reverb G2 Omnicept Edition. https://www.hp.com/us-en/vr/reverb-g2-vr-headset-omnicept-edition.html.

[26] VIVE Facial Tracker. https://www.vive.com/us/accessory/facial-tracker/.

[27] Faceteq. https://www.emteqlabs.com/emteqpro-vive/.

[28] Victor Adriel de Jesus Oliveira, Luca Brayda, Luciana Nedel, and Anderson Maciel. Designing a vibrotactile head-mounted display for spatial awareness in 3d spaces. *IEEE Transactions on Visualization and Computer Graphics*, 23(4):1409–1417, April 2017.

[29] Roshan Lalintha Peiris, Wei Peng, Zikun Chen, and Kouta Minamizawa. Exploration of cuing methods for localization of spatial cues using thermal haptic feedback on the forehead. In *2017 IEEE World Haptics Conference (WHC)*, pages 400–405, 2017.

[30] Roshan Lalintha Peiris, Wei Peng, Zikun Chen, Liwei Chan, and Kouta Minamizawa. *ThermoVR: Exploring Integrated Thermal Haptic Feedback with Head Mounted Displays*, page 5452–5456. Association for Computing Machinery, New York, NY, USA, 2017.

[31] Hsin-Ruey Tsai and Bing-Yu Chen. Elastimpact: 2.5d multilevel instant impact using elasticity on head-mounted displays. In *Proceedings of the 32nd Annual ACM Sympo-*

sium on User Interface Software and Technology, UIST '19, page 429–437, New York, NY, USA, 2019. Association for Computing Machinery.

[32] Wen-Jie Tseng, Li-Yang Wang, and Liwei Chan. Facewidgets: Exploring tangible interaction on face with head-mounted displays. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, page 417–427, New York, NY, USA, 2019. Association for Computing Machinery.

[33] Fumihiko Nakamura, Katsuhiro Suzuki, Katsutoshi Masai, Yuta Itoh, Yuta Sugiura, and Maki Sugimoto. Automatic labeling of training data by vowel recognition for mouth shape recognition with optical sensors embedded in head-mounted display. In *ICAT-EGVE*, pages 9–16, 2019.

[34] Fumihiko Nakamura, Adrien Verhulst, Kuniharu Sakurada, and Maki Sugimoto. Virtual whiskers: Spatial directional guidance using cheek haptic stimulation in a virtual environment. In *Augmented Humans Conference 2021*, AHs'21, page 141–151, New York, NY, USA, 2021. Association for Computing Machinery.

[35] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021.

[36] Jose M. Chaquet, Enrique J. Carmona, and Antonio Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013.

[37] Jesus Suarez and Robin R. Murphy. Hand gesture recognition with depth images: A review. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 411–417, 2012.

[38] Kinect. https://azure.microsoft.com/ja-jp/services/kinect-dk/.

[39] RealSense. https://www.intel.co.jp/content/www/jp/ja/architecture-and-technology/realsense-overview.html.

[40] ZED 2i. https://www.stereolabs.com/zed-2i/.

[41] Ultraleap. https://www.ultraleap.com/product/stereo-ir-170/.

[42] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Transactions on Cybernetics*, 43(5):1318–1334, 2013.

[43] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, Arran Topalian,

Erroll Wood, Sameh Khamis, Pushmeet Kohli, Shahram Izadi, Richard Banks, Andrew Fitzgibbon, and Jamie Shotton. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Trans. Graph.*, 35(4), jul 2016.

[44] Liwei Chan, Chi-Hao Hsieh, Yi-Ling Chen, Shuo Yang, Da-Yuan Huang, Rong-Hao Liang, and Bing-Yu Chen. Cyclops: Wearable and single-piece full-body gesture input devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3001–3009, New York, NY, USA, 2015. Association for Computing Machinery.

[45] Dong-Hyun Hwang, Kohei Aso, Ye Yuan, Kris Kitani, and Hideki Koike. *Mono-Eye: Multimodal Human Motion Capture System Using A Single Ultra-Wide Fisheye Camera*, page 98–111. Association for Computing Machinery, New York, NY, USA, 2020.

[46] Jhe-Wei Lin, Chiuan Wang, Yi Yao Huang, Kuan-Ting Chou, Hsuan-Yu Chen, Wei-Luan Tseng, and Mike Y. Chen. Backhand: Sensing hand gestures via back of the hand. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology*, UIST '15, page 557–564, New York, NY, USA, 2015. Association for Computing Machinery.

[47] A. Ferrone, F. Maita, L. Maiolo, M. Arquilla, A. Castiello, A. Pecora, X. Jiang, C. Menon, A. Ferrone, and L. Colace. Wearable band for hand gesture recognition based on strain sensors. In *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 1319–1322, 2016.

[48] Jess McIntosh, Asier Marzo, and Mike Fraser. Sensir: Detecting hand gestures with a wearable bracelet using infrared transmission and reflection. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, UIST '17, page 593–597, New York, NY, USA, 2017. Association for Computing Machinery.

[49] W. K. Wong, Filbert H. Juwono, and Brendan Teng Thiam Khoo. Multi-features capacitive hand gesture recognition sensor: A machine learning approach. *IEEE Sensors Journal*, 21(6):8441–8450, 2021.

[50] Yasha Iravantchi, Yang Zhang, Evi Bernitsas, Mayank Goel, and Chris Harrison. *Interferi: Gesture Sensing Using On-Body Acoustic Interferometry*, page 1–13. Association for Computing Machinery, New York, NY, USA, 2019.

[51] Chen Liang, Chun Yu, Yue Qin, Yuntao Wang, and Yuanchun Shi. Dualring: Enabling

subtle and expressive hand interaction with dual imu rings. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(3), sep 2021.

[52] Nabeel Siddiqui and Rosa H. M. Chan. Multimodal hand gesture recognition using single imu and acoustic measurements at wrist. *PLOS ONE*, 15(1):1–12, 01 2020.

[53] Koumei Fukahori, Daisuke Sakamoto, and Takeo Igarashi. Exploring subtle foot plantar-based gestures with sock-placed pressure sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3019–3028, New York, NY, USA, 2015. Association for Computing Machinery.

[54] Harrison L. Bartlett and Michael Goldfarb. A phase variable approach for imu-based locomotion activity recognition. *IEEE Transactions on Biomedical Engineering*, 65(6):1330–1338, 2018.

[55] Chris Harrison, Desney Tan, and Dan Morris. Skinput: Appropriating the body as an input surface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 453–462, New York, NY, USA, 2010. Association for Computing Machinery.

[56] Martin Weigel, Vikram Mehta, and Jürgen Steimle. More than touch: Understanding how people use skin as an input surface for mobile computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 179–188, New York, NY, USA, 2014. Association for Computing Machinery.

[57] Masa Ogata, Yuta Sugiura, Hirotaka Osawa, and Michita Imai. *IRing: Intelligent Ring Using Infrared Reflection*, page 131–136. Association for Computing Machinery, New York, NY, USA, 2012.

[58] Masa Ogata, Yuta Sugiura, Yasutoshi Makino, Masahiko Inami, and Michita Imai. Senskin: Adapting skin as a soft interface. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, page 539–544, New York, NY, USA, 2013. Association for Computing Machinery.

[59] Martin Weigel, Tong Lu, Gilles Bailly, Antti Oulasvirta, Carmel Majidi, and Jürgen Steimle. Iskin: Flexible, stretchable and visually customizable on-body touch sensors for mobile computing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 2991–3000, New York, NY, USA, 2015. Association for Computing Machinery.

[60] Cheng Zhang, AbdelKareem Bedri, Gabriel Reyes, Bailey Bercik, Omer T. Inan, Thad E. Starner, and Gregory D. Abowd. Tapskin: Recognizing on-skin input for

smartwatches. In *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces*, ISS '16, page 13–22, New York, NY, USA, 2016. Association for Computing Machinery.

[61] Alexandra Ion, Edward Jay Wang, and Patrick Baudisch. *Skin Drag Displays: Dragging a Physical Tactor across the User's Skin Produces a Stronger Tactile Stimulus than Vibrotactile*, page 2501–2504. Association for Computing Machinery, New York, NY, USA, 2015.

[62] Alexandra Delazio, Ken Nakagaki, Roberta L. Klatzky, Scott E. Hudson, Jill Fain Lehman, and Alanson P. Sample. Force jacket: Pneumatically-actuated jacket for embodied haptic experiences. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery.

[63] bhaptics, 2015.

[64] Junji Watanabe and Hideyuki Ando. Pace-sync shoes: intuitive walking-pace guidance based on cyclic vibro-tactile stimulation for the foot. *Virtual Reality*, 14(3):213–219, 2010.

[65] Akira Matsuda, Kazunori Nozawa, Kazuki Takata, Atsushi Izumihara, and Jun Rekimoto. Hapticpointer: A neck-worn device that presents direction by vibrotactile feedback for remote collaboration tasks. In *Proceedings of the Augmented Humans International Conference*, AHs '20, New York, NY, USA, 2020. Association for Computing Machinery.

[66] Jilin Tu, T. Huang, and Hai Tao. Face as mouse through visual face tracking. In *The 2nd Canadian Conference on Computer and Robot Vision (CRV'05)*, pages 339–346, 2005.

[67] Yulia Gizatdinova, Oleg pakov, and Veikko 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, 2012.

[68] Yuki Hashimoto, Satsuki Nakata, and Hiroyuki Kajimoto. Novel tactile display for emotional tactile experience. In *Proceedings of the International Conference on Advances in Computer Enterntainment Technology*, ACE '09, page 124–131, New York, NY, USA, 2009. Association for Computing Machinery.

[69] Hiroyuki Kajimoto, Yonezo Kanno, and Susumu Tachi. Forehead electro-tactile display

for vision substitution. In *Proc. EuroHaptics*, page 11. Citeseer, 2006.

[70] Brais Martinez, Michel F. Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 10(3):325–347, 2019.

[71] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, pages 1–1, 2020.

[72] Yuta Matsunaga and Kenji Matsui. Mobile device-based speech enhancement system using lip-reading. In *2018 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pages 1–4, 2018.

[73] Chatsopon Deepateep and Pongsagon Vichitvejpaisal. Facial movement interface for mobile devices using depth-sensing camera. In *2020 12th International Conference on Knowledge and Smart Technology (KST)*, pages 115–120, 2020.

[74] Yukang Yan, Chun Yu, Wengrui Zheng, Ruining Tang, Xuhai Xu, and Yuanchun Shi. Frownonerror: Interrupting responses from smart speakers by facial expressions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI ’20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[75] Katsutoshi Masai, Yuta Sugiura, Masa Ogata, Kai Kunze, Masahiko Inami, and Maki 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.

[76] Katsutoshi Masai, Yuta Sugiura, and Maki 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, New York, NY, USA, 2018. Association for Computing Machinery.

[77] Takashi Kikuchi, Yuta Sugiura, Katsutoshi Masai, Maki Sugimoto, and Bruce 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, New York, NY, USA, 2017. Association for Computing Machinery.

[78] Takuma Hashimoto, Suzanne Low, Koji Fujita, Risa Usumi, Hiroshi Yanagihara, Chihiro Takahashi, Maki Sugimoto, and Yuta Sugiura. Tongueinput: Input method by tongue gestures using optical sensors embedded in mouthpiece. In *2018 57th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pages

1219–1224, 2018.

[79] Denys J.C. Matthies, Chamod Weerasinghe, Bodo Urban, and Suranga Nanayakkara. Capglasses: Untethered capacitive sensing with smart glasses. In *Augmented Humans Conference 2021*, AHs'21, page 121–130, New York, NY, USA, 2021. Association for Computing Machinery.

[80] Yutaro Suzuki, Kodai Sekimori, Yuki Yamato, Yusuke Yamasaki, Buntarou Shizuki, and Shin Takahashi. A mouth gesture interface featuring a mutual-capacitance sensor embedded in a surgical mask. In *International Conference on Human-Computer Interaction*, pages 154–165. Springer, 2020.

[81] Mayank Goel, Chen Zhao, Ruth Vinisha, and Shwetak N. Patel. Tongue-in-cheek: Using wireless signals to enable non-intrusive and flexible facial gestures detection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 255–258, New York, NY, USA, 2015. Association for Computing Machinery.

[82] Xuhai Xu, Haitian Shi, Xin Yi, WenJia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K. Dey. Earbuddy: Enabling on-face interaction via wireless earbuds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[83] Alvaro Cassinelli, Carson Reynolds, and Masatoshi Ishikawa. Augmenting spatial awareness with haptic radar. In *2006 10th IEEE International Symposium on Wearable Computers*, pages 61–64, Oct 2006.

[84] Oliver Beren Kaul and Michael Rohs. Haptichead: A spherical vibrotactile grid around the head for 3d guidance in virtual and augmented reality. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3729–3740, New York, NY, USA, 2017. Association for Computing Machinery.

[85] Matthias Berning, Florian Braun, Till Riedel, and Michael Beigl. Proximityhat: A head-worn system for subtle sensory augmentation with tactile stimulation. In *Proceedings of the 2015 ACM International Symposium on Wearable Computers*, ISWC '15, page 31–38, New York, NY, USA, 2015. Association for Computing Machinery.

[86] Yuka Sato and Ryoko Ueoka. Investigating haptic perception of and physiological responses to air vortex rings on a user's cheek. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3083–3094, New

York, NY, USA, 2017. Association for Computing Machinery.

[87] Ryoko Ueoka, Mami Yamaguchi, and Yuka Sato. Interactive cheek haptic display with air vortex rings for stress modification. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, page 1766–1771, New York, NY, USA, 2016. Association for Computing Machinery.

[88] Shigeo Yoshida, Takuji Narumi, Tomohiro Tanikawa, Hideaki Kuzuoka, and Michitaka Hirose. *Teardrop Glasses: Pseudo Tears Induce Sadness in You and Those Around You*. Association for Computing Machinery, New York, NY, USA, 2021.

[89] Hyunjae Gil, Hyungki Son, Jin Ryong Kim, and Ian Oakley. Whiskers: Exploring the use of ultrasonic haptic cues on the face. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery.

[90] Minkyeong Lee, Seungwoo Je, Woojin Lee, Daniel Ashbrook, and Andrea Bianchi. Activearring: Spatiotemporal haptic cues on the ears. *IEEE Transactions on Haptics*, 12(4):554–562, 2019.

[91] Arshad Nasser, Kexin Zheng, and Kening Zhu. *ThermEarhook: Investigating Spatial Thermal Haptic Feedback on the Auricular Skin Area*, page 662–672. Association for Computing Machinery, New York, NY, USA, 2021.

[92] Hui Tang and D. J. Beebe. An oral tactile interface for blind navigation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(1):116–123, 2006.

[93] Karan Ahuja, Chris Harrison, Mayank Goel, and Robert Xiao. Mecap: Whole-body digitization for low-cost vr/ar headsets. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, page 453–462, New York, NY, USA, 2019. Association for Computing Machinery.

[94] UmurAybars Ciftci, Xing Zhang, and Lijun Tin. Partially occluded facial action recognition and interaction in virtual reality applications. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 715–720, 2017.

[95] Xavier P. Burgos-Artizzu, Julien Fleureau, Olivier Dumas, Thierry Tapie, François LeClerc, and Nicolas Mollet. Real-time expression-sensitive hmd face reconstruction. In *SIGGRAPH Asia 2015 Technical Briefs*, SA '15, pages 9:1–9:4, New York, NY, USA, 2015. ACM.

[96] Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li. High-fidelity facial and speech animation for vr hmds. *ACM Trans. Graph.*, 35(6):221:1–221:14, November

2016.

[97] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Facevr: Real-time gaze-aware facial reenactment in virtual reality. *ACM Trans. Graph.*, 37(2), jun 2018.

[98] Yukang Yan, Chun Yu, Xin Yi, and Yuanchun Shi. Headgesture: Hands-free input approach leveraging head movements for hmd devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(4), dec 2018.

[99] Taizhou Chen, Lantian Xu, Xianshan Xu, and Kening Zhu. Gestonhmd: Enabling gesture-based interaction on low-cost vr head-mounted display. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2597–2607, 2021.

[100] Jianwen Lou, Yiming Wang, Charles Nduka, Mahyar Hamedi, Ifigeneia Mavridou, Fei-Yue Wang, and Hui Yu. Realistic facial expression reconstruction for vr hmd users. *IEEE Transactions on Multimedia*, 22(3):730–743, 2020.

[101] Chen Chen, Ke Sun, and Xinyu Zhang. Exgsense: Toward facial gesture sensing with a sparse near-eye sensor array. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (Co-Located with CPS-IoT Week 2021)*, IPSN '21, page 222–237, New York, NY, USA, 2021. Association for Computing Machinery.

[102] Takuro Nakao, Yun Suen Pai, Megumi Isogai, Hideaki Kimata, and Kai Kunze. Make-a-face: A hands-free, non-intrusive device for tongue/mouth/cheek input using emg. In *ACM SIGGRAPH 2018 Posters*, SIGGRAPH '18, New York, NY, USA, 2018. Association for Computing Machinery.

[103] Hiromi Nakamura and Homei Miyashita. Control of augmented reality information volume by glabellar fader. In *Proceedings of the 1st Augmented Human International Conference*, AH '10, New York, NY, USA, 2010. Association for Computing Machinery.

[104] Richard Li and Gabriel Reyes. Buccal: Low-cost cheek sensing for inferring continuous jaw motion in mobile virtual reality. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, ISWC '18, page 180–183, New York, NY, USA, 2018. Association for Computing Machinery.

[105] Mose Sakashita, Tatsuya Minagawa, Amy Koike, Ippei Suzuki, Keisuke Kawahara, and Yoichi Ochiai. You as a puppet: Evaluation of telepresence user interface for puppetry. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and*

Technology, UIST '17, pages 217–228, New York, NY, USA, 2017. ACM.

[106] Koki Yamashita, Takashi Kikuchi, Katsutoshi Masai, Maki Sugimoto, Bruce H. Thomas, and Yuta 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, New York, NY, USA, 2017. Association for Computing Machinery.

[107] Jinhyuk Kim, Jaekwang Cha, Hojun Lee, and Shiho Kim. Hand-free natural user interface for vr hmd with ir based facial gesture tracking sensor. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, VRST '17, New York, NY, USA, 2017. Association for Computing Machinery.

[108] Katsuhiro Suzuki, Fumihiko Nakamura, Jiu Otsuka, Katsutoshi Masai, Yuta Itoh, Yuta Sugiura, and Maki Sugimoto. Recognition and mapping of facial expressions to avatar by embedded photo reflective sensors in head mounted display. In *2017 IEEE Virtual Reality (VR)*, pages 177–185, March 2017.

[109] Cathy Fang, Yang Zhang, Matthew Dworman, and Chris Harrison. Wireality: Enabling complex tangible geometries in virtual reality with worn multi-string haptics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–10, New York, NY, USA, 2020. Association for Computing Machinery.

[110] Mohammed Al-Sada, Keren Jiang, Shubhankar Ranade, Mohammed Kalkattawi, and Tatsuo Nakajima. Hapticsnakes: multi-haptic feedback wearable robots for immersive virtual reality. *Virtual Reality*, 24(2):191–209, 2020.

[111] Mohammed Al-Sada, Keren Jiang, Shubhankar Ranade, Xinlei Piao, Thomas Höglund, and Tatsuo Nakajima. Hapticserpent: A wearable haptic feedback robot for vr. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, page 1–6, New York, NY, USA, 2018. Association for Computing Machinery.

[112] Matthias Hoppe, Pascal Knierim, Thomas Kosch, Markus Funk, Lauren Futami, Stefan Schneegass, Niels Henze, Albrecht Schmidt, and Tonja Machulla. Vrhapticdrones: Providing haptics in virtual reality through quadcopters. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*, MUM 2018, page 7–18, New York, NY, USA, 2018. Association for Computing Machinery.

[113] Chi Wang, Da-Yuan Huang, Shuo-wen Hsu, Chu-En Hou, Yeu-Luen Chiu, Ruei-Che Chang, Jo-Yu Lo, and Bing-Yu Chen. Masque: Exploring lateral skin stretch feedback

on the face with head-mounted displays. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, page 439–451, New York, NY, USA, 2019. Association for Computing Machinery.

[114] Takayuki Kameoka and Hiroyuki Kajimoto. Tactile transfer of finger information through suction tactile sensation in hmds. In *2021 IEEE World Haptics Conference (WHC)*, pages 949–954, 2021.

[115] Hong-Yu Chang, Wen-Jie Tseng, Chia-En Tsai, Hsin-Yu Chen, Roshan Lalintha Peiris, and Liwei Chan. Facepush: Introducing normal force on face with head-mounted displays. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, UIST '18, page 927–935, New York, NY, USA, 2018. Association for Computing Machinery.

[116] Wen-Jie Tseng, Yi-Chen Lee, Roshan Lalintha Peiris, and Liwei Chan. A skin-stroke display on the eye-ring through head-mounted displays. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[117] Yi-Hao Peng, Carolyn Yu, Shi-Hong Liu, Chung-Wei Wang, Paul Taele, Neng-Hao Yu, and Mike Y. Chen. Walkingvibe: Reducing virtual reality sickness and improving realism while walking in vr using unobtrusive head-mounted vibrotactile feedback. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[118] Nimesha Ranasinghe, Pravar Jain, Shienny Karwita, David Tolley, and Ellen Yi-Luen Do. *Ambiotherm: Enhancing Sense of Presence in Virtual Reality by Simulating Real-World Environmental Conditions*, page 1731–1742. Association for Computing Machinery, New York, NY, USA, 2017.

[119] Nimesha Ranasinghe, Pravar Jain, Nguyen Thi Ngoc Tram, Koon Chuan Raymond Koh, David Tolley, Shienny Karwita, Lin Lien-Ya, Yan Liangkun, Kala Shamaiah, Chow Eason Wai Tung, Ching Chiuan Yen, and Ellen Yi-Luen Do. Season traveller: Multisensory narration for enhancing the virtual reality experience. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery.

[120] Shi-Hong Liu, Neng-Hao Yu, Liwei Chan, Yi-Hao Peng, Wei-Zen Sun, and Mike Y. Chen. Phantomlegs: Reducing virtual reality sickness using head-worn haptic devices. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages

817–826, March 2019.

[121] Alexander Wilberz, Dominik Leschtschow, Christina Trepkowski, Jens Maiero, Ernst Kruijff, and Bernhard Riecke. Facehaptics: Robot arm based versatile facial haptics for immersive environments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[122] Mike Sinclair, Eyal Ofek, Mar Gonzalez-Franco, and Christian Holz. Capstancrunch: A haptic vr controller with user-supplied force feedback. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, page 815–829, New York, NY, USA, 2019. Association for Computing Machinery.

[123] Joanna Bergström and Kasper Hornbæk. Human–computer interaction on the skin. *ACM Comput. Surv.*, 52(4), aug 2019.

[124] Srinath Sridhar, Anders Markussen, Antti Oulasvirta, Christian Theobalt, and Sebastian Boring. Watchsense: On- and above-skin input sensing through a wearable depth sensor. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3891–3902, New York, NY, USA, 2017. Association for Computing Machinery.

[125] Manuel Prätorius, Dimitar Valkov, Ulrich Burgbacher, and Klaus Hinrichs. Digitap: An eyes-free vr/ar symbolic input device. In *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, VRST '14, page 9–18, New York, NY, USA, 2014. Association for Computing Machinery.

[126] Adiyan Mujibiya, Xiang Cao, Desney S. Tan, Dan Morris, Shwetak N. Patel, and Jun Rekimoto. The sound of touch: On-body touch and gesture sensing based on transdermal ultrasound propagation. In *Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces*, ITS '13, page 189–198, New York, NY, USA, 2013. Association for Computing Machinery.

[127] Gierad Laput, Robert Xiao, and Chris Harrison. Viband: High-fidelity bio-acoustic sensing using commodity smartwatch accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, page 321–333, New York, NY, USA, 2016. Association for Computing Machinery.

[128] Shu-Yang Lin, Chao-Huai Su, Kai-Yin Cheng, Rong-Hao Liang, Tzu-Hao Kuo, and Bing-Yu Chen. Pub - point upon body: Exploring eyes-free interaction and methods on an arm. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software*

and Technology, UIST '11, page 481–488, New York, NY, USA, 2011. Association for Computing Machinery.

[129] Masa Ogata and Michita Imai. Skinwatch: Skin gesture interaction for smart watch. In *Proceedings of the 6th Augmented Human International Conference*, AH '15, page 21–24, New York, NY, USA, 2015. Association for Computing Machinery.

[130] Robert Xiao, Teng Cao, Ning Guo, Jun Zhuo, Yang Zhang, and Chris Harrison. *LumiWatch: On-Arm Projected Graphics and Touch Input*, page 1–11. Association for Computing Machinery, New York, NY, USA, 2018.

[131] Yasutoshi Makino, Yoshikazu Furuyama, Seki Inoue, and Hiroyuki Shinoda. Haptoclone (haptic-optical clone) for mutual tele-environment by real-time 3d image transfer with midair force feedback. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1980–1990, New York, NY, USA, 2016. Association for Computing Machinery.

[132] Katsutoshi Masai, Kai Kunze, Daisuke Sakamoto, Yuta Sugiura, and Maki Sugimoto. Face commands - user-defined facial gestures for smart glasses. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 374–386, 2020.

[133] Emanuele Perini, Simone Soria, Andrea Prati, and Rita Cucchiara. Facemouse: A human-computer interface for tetraplegic people. In *European Conference on Computer Vision*, pages 99–108. Springer, 2006.

[134] Atieh Taheri, Ziv Weissman, and Misha Sra. Exploratory design of a hands-free video game controller for a quadriplegic individual. In *Augmented Humans Conference 2021*, AHs'21, page 131–140, New York, NY, USA, 2021. Association for Computing Machinery.

[135] Masaki Yuki, William W. Maddux, and Takahiko Masuda. Are the windows to the soul the same in the east and west? cultural differences in using the eyes and mouth as cues to recognize emotions in japan and the united states. *Journal of Experimental Social Psychology*, 43(2):303–311, March 2007.

[136] Zoric Goranka and Igor S. Pandzic. A real-time lip sync system using a genetic algorithm for automatic neural network configuration. In *2005 IEEE International Conference on Multimedia and Expo*, pages 1366–1369, July 2005.

[137] Oculus Lipsync Unity. https://developer.oculus.com/downloads/package/oculus-lipsync-unity/.

[138] Anna Gruebler and Kenji Suzuki. Design of a wearable device for reading positive expressions from facial emg signals. *IEEE Transactions on Affective Computing*, 5(3):227–237, July 2014.

[139] Itoi Kiyoaki, Misono Yasushi, and Kobayashi Yukio. Intelligent coding of facial expression using neural network and morphing. In *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No.01EX489)*, pages 352–355, May 2001.

[140] Katsutoshi Masai, Kai Kunze, Yuta Sugiura, Masa Ogata, Masahiko Inami, and Maki Sugimoto. Evaluation of facial expression recognition by a smart eyewear for facial direction changes, repeatability, and positional drift. *ACM Trans. Interact. Intell. Syst.*, 7(4), dec 2017.

[141] Alan F. Stokes and Christopher D. Wickens. Aviation Displays. In *Human Factors in Aviation*, pages 387–431. Elsevier, 1988.

[142] Joost X. Maier and Jennifer M. Groh. Multisensory guidance of orienting behavior. *Hearing Research*, 258(1-2):106–112, dec 2009.

[143] Bernhard Weber, Simon Schätzle, Thomas Hulin, Carsten Preusche, and Barbara Deml. Evaluation of a vibrotactile feedback device for spatial guidance. *2011 IEEE World Haptics Conference, WHC 2011*, pages 349–354, 2011.

[144] Jan BF Van Erp. Presenting directions with a vibrotactile torso display. *Ergonomics*, 48(3):302–313, 2005.

[145] Lisa M Pritchett, Michael J Carnevale, and Laurence R Harris. Reference frames for coding touch location depend on the task. *Experimental brain research*, 222(4):437–445, 2012.

[146] Maria Z Siemionow. *The Know-How of Face Transplantation.* Springer London, London, 2011.

[147] Lichao Shen, MHD Yamen Saraiji, Kai Kunze, Kouta Minamizawa, and Roshan Lalintha Peiris. Visuomotor influence of attached robotic neck augmentation. In *Symposium on Spatial User Interaction*, New York, NY, USA, 2020. Association for Computing Machinery.

[148] Jing Jin, Zongmei Chen, Ren Xu, Yangyang Miao, Xingyu Wang, and Tzyy-Ping Jung. Developing a novel tactile p300 brain-computer interface with a cheeks-stim paradigm. *IEEE Transactions on Biomedical Engineering*, 67(9):2585–2593, 2020.

[149] Theophilus Teo, Fumihiko Nakamura, Maki Sugimoto, Adrien Verhulst, Gun A. Lee,

Mark Billinghurst, and Matt Adcock. WeightSync: Proprioceptive and Haptic Stimulation for Virtual Physical Perception. In Ferran Argelaguet, Ryan McMahan, and Maki Sugimoto, editors, *ICAT-EGVE 2020 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*. The Eurographics Association, 2020.

[150] Alexander Marquardt, Christina Trepkowski, Tom David Eibich, Jens Maiero, and Ernst Kruijff. Non-visual cues for view management in narrow field of view augmented reality displays. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 190–201, 2019.

[151] Jaka Sodnik, Saso Tomazic, Raphael Grasset, Andreas Duenser, and Mark Billinghurst. Spatial sound localization in an augmented reality environment. In *Proceedings of the 18th Australia Conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, OZCHI '06, page 111–118, New York, NY, USA, 2006. Association for Computing Machinery.

[152] Tobias Rodemann, Gokhan Ince, Frank Joublin, and Christian Goerick. Using binaural and spectral cues for azimuth and elevation localization. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2185–2190, 2008.

[153] Young-Woo Park, Seok-Hyung Bae, and Tek-Jin Nam. How do couples use cheektouch over phone calls? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 763–766, New York, NY, USA, 2012. Association for Computing Machinery.