Face Commands – User-Defined Facial Gestures for Smart Glasses

Katsutoshi Masai* Keio University
Kai Kunze† Keio Media Design
Daisuke Sakamoto‡ Hokkaido University
Yuta Sugiura§ Keio University
Maki Sugimoto¶ Keio University

ABSTRACT

We propose the use of face-related gestures involving the movement of the face, eyes, and head for augmented reality (AR). This technique allows us to use computer systems via hands-free, discreet interactions. In this paper, we present an elicitation study to explore the proper use of facial gestures for daily tasks in the context of a smart home. We used Amazon Mechanical Turk to conduct this study (N = 37). Based on the proposed gestures, we report usage scenarios and complexity, proposed associations between gestures/tasks, a user-defined gesture set, and insights from the participants. We also conducted a technical feasibility study (N = 13) with participants using smart eyewear to consider their uses in daily life. The device has 16 optical sensors and an inertial measurement unit (IMU). We can potentially integrate the system into optical see-through displays or other smart glasses. The results demonstrate that the device can detect eight temporal face-related gestures with a mean F1 score of 0.911 using a convolutional neural network (CNN). We also report the results of user-independent training and a one-hour recording of the experimenter testing two of the gestures.

Index Terms: Human-centered computing—Interaction techniques; Human-centered computing—Ubiquitous and mobile computing design and evaluation methods

1 INTRODUCTION

When communicating with people, we often use facial expressions and non-verbal gestures to implicitly convey our intentions or influence their responses [46]. In interactive experiences, as envisioned in augmented reality, we require more discrete, implicit interaction modalities. In our research, we explore the use of non-verbal communication, especially facial gestures, in augmented reality and ubiquitous computing contexts.

As we see more and more augmented reality applications from research being applied in everyday life, we assume a larger adoption of optical see-through head-mounted displays (OST-HMDs). In scenarios, where OST-HMDs are used, facial gestures might be a natural interaction modality. The face is already augmented, adding sensors to the HMD seems simple (especially the optical sensors we use in our research), and other technology might seem cumbersome. One can imagine smartphones or remote controllers as common input devices, yet the user needs to carry them around, and they cannot be used "hands-free." A touch-bar or buttons on the HMD are other solutions, yet they have quite limited areas to deal with and might result in socially awkward interactions. Several virtual reality (VR)/AR headsets opt for mid-air input techniques (e.g., HoloLens). Yet, they require space and can lead to muscle fatigue. In addition, voice commands can be used hands-free, but they might also be socially awkward and cannot be used in noisy environments. Overall, we believe that future AR/VR headsets will use a combination of these approaches. We are contributing an exploration of facial gestures and a simple, low-cost, effective, and resource-constrained prototype system to this end.

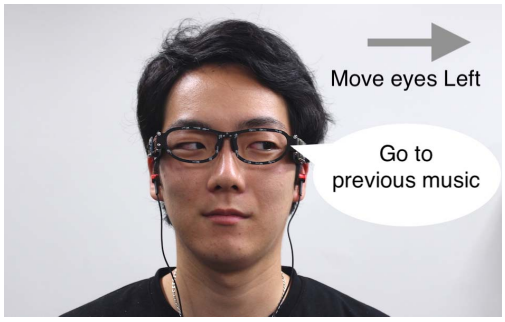

Figure 1: We defined face-related gestures used to perform daily tasks in an elicitation study. Then, we verified the gestures' technical feasibility using smart eyewear.

A lot of research efforts have focused on designing interactions to be more explicit and obvious for bystanders to understand. We are exploring the opposite direction and investigating how to make interactions more subtle and discreet. This paper focuses on the face as an input modality. Facial gestures can be subtle, and we can use them to operate devices intuitively, similar to communicating non-verbally with other people. Moreover, facial gestures do not require the use of the hands and are appropriate for noisy environments.

To explore the design space of face-related gestures, we conducted a gesture elicitation study.This method aims at improving the usability of performing gestures by incorporating feedback from users. We defined face-related gestures (Figure 1) as any gestures that involve the movement of the facial muscles, eyes, or head or a combination of these movements. We excluded hand-to-face manipulation since it was already explored in other studies [31, 55].

For the study, we first created a design reference by conducting interviews and obtaining feedback regarding the difficulty of designing face-related gestures from scratch in a preliminary study. Then, we chose 15 daily mobile tasks based on previous work [52, 55] and the context of smart homes. An elicitation study was conducted online using Amazon Mechanical Turk (MTurk) (N = 37). We coded the proposals into facial action units (AUs) using the Facial Action Coding System (FACS) [14] and made a user-defined gesture set that summarizes frequently proposed gestures for each task. Then, we examined the gesture set's technical feasibility using smart eyewear with optical sensors following the principle used by Masai et al. [40, 41]. Considering possible integration with OST-HMDs, we adopted a glass-shaped optical sensing system [39, 64]. This work demonstrates the potential of face-related gestures for interacting with smart devices using an optical sensing system.

The contribution of this work is as follows:

- We explored the use of face-related gestures for daily mobile tasks. We conducted a gesture elicitation study via MTurk (N= 37). We summarized the findings and developed a user-defined set of face-related gestures based on these insights.

*e-mail:masai@imlab.ics.keio.ac.jp
†e-mail:kai@kmd.keio.ac.jp
‡e-mail:sakamoto@ist.hokudai.ac.jp
§e-mail:sugiura@keio.jp
¶e-mail:sugimoto@imlab.ics.keio.ac.jp

978-1-7281-8508-8/20/$31.00 ©2020 IEEE
DOI 10.1109/ISMAR50242.2020.00064

- We adopted a CNN approach to wearable sensor data for temporal face-related gesture detection and classification. Then, we evaluated the technical feasibility of detecting and classifying the eight kinds of temporal facial gestures included in the user-defined set. We achieved an F1 score of 0.911, with 13 participants.
- We conducted a one-hour false-positive study for two types of gestures to show the possibilities for using face-related gestures in real life.

2 Related Work

Our work builds on the research areas of user-defined gestures, hand-to-face input, and facial gesture input methods.

Before the emergence of the concept of user-defined gestures, gesture sets were developed primarily by system designers. They focused on technical feasibility more than how users behave. Thus, these gesture sets have tended to be complex and not very intuitive. Wobbrock et al. conducted a guessability study and employed users in the design of gestures [63]. They showed participants a referent (i.e., the computational effect to be caused by using a gesture, such as playing music) and asked them to create a sign (i.e., a symbol to activate the referent). According to Morris et al., a gesture set proposed by users is preferred to a gesture set designed by human-computer interaction researchers [44]. User-defined design is now a standard method and applied in a wide range of domains, such as as mobile tasks [52], smartwatches [5], augmented reality [47], and television control [58]. Furthermore, studies were conducted on gestures utilizing various body parts, such as the head [65], the upper limbs [61], a single hand [8], and the feet [15]. The method works well even when users are unfamiliar with the gestures.

In terms of HMDs, many recent studies proposed hand-to-face input methods using elicitation techniques [13, 31, 55]. Serrano et al. [55] conducted a guessability study of hand-to-face inputs, finding that the cheek is a promising area regarding possible inputs. Lee et al. [31] conducted a similar study and focused on social acceptability by running the experiment in a cafe. They summarized five design strategies for the inputs based on the results. Both researchers validated their techniques by making prototypes that enable inputs using the cheek, ear, and jaw. Dierk et al. conducted an elicitation study focusing on the use of hat technology as an interaction modality [13]. Their consensus gesture set included touching a hat in addition to head gestures.

Some studies have utilized the surface of the device [20, 62], and researchers have also proposed hand-to-face input techniques from the perspective of technical feasibility [9, 25, 32, 42, 64]. Yamashita et al. attached an optical sensor array to an HMD to allow gesture inputs via the cheek [64]. Similarly, Kikuchi et al. used an optical sensor array integrated into an earphone to enable gesture inputs via the ear [25]. Hand-to-face gesture recognition with eyewear devices has also been done using electrooculography (EOG) [32] or an optical sensor array [42]. Furthermore, Loorak et al. proposed a hand-to-face input method using a smartphone camera to improve the smartphone user experience [34]. Vega et al. proposed an interface for sending an input command to a computer utilizing capacitive touch sensors and by touching hair extensions [59]. Researchers have designed a user-defined gesture set for hand-to-face inputs that have undergone considerable technical verification. While hand-to-face input is effective, especially in an eyes-free context, hygiene concerns about touching the face remain. We focused on face-related gestures, which are suitable for hands-free settings.

Researchers have developed various techniques to detect facial movements [10, 30, 45]. The typical way to detect facial gestures is a computer vision approach [37], which mostly focuses on emotion recognition. Jota and Wigdor explored the design space of eyelid gestures using a commodity camera and proposed various application cases [23]. Špakov and Majaranta introduced a hands-free interaction system combining gaze pointing and head gestures [60], while Gizatdinova et al. utilized face and visual gesture detection to manipulate a scrollable virtual keyboard [17]. However, a vision-based approach is limited to specific spaces as it is influenced by lighting and obstacles and has bulky processing systems.

Considering the context of mobile tasks in daily life, we focus on a wearable solution. For head gestures, the use of HMDs for the mobile context has been proposed [11, 65]. The use of head tilting gestures for mobile interaction is explored in detail in terms of the head tilt angle and velocity using three-axis accelerometers and magnetometers attached to a hat [11]. Rantanen et al. presented a prototype to detect frowning and eyebrow lifting using capacitive sensors [49]. Wearable EOG glasses allow the wearer to use eye movements as an input modality [7, 12, 22]. Kanoh et al. presented eyewear to detect eye movements using three EOG electrodes in an unrestricted way [24]. The same device was used to recognize kiss gestures to obtain passive context sensing and awareness [33]. Manabe et al. developed an earphone-based input device utilizing bio-potential electrodes that can detect eye gestures [36]. Many studies have focused on the use of various facial gestures as input for wearable devices. Expression Glasses use piezoelectric sensors to detect interest or confusion from a facial expression [54]. Masai et al. classified the primary emotion of facial expressions using eyewear with embedded optical sensor arrays [40]. They also developed an eye-gesture classification system using the eyewear [38]. Iravantchi et al. created hand and face gesture recognition prototypes using acoustic interferometry [21]. Their face mask classified eight facial gestures with high accuracy and was implemented to fit HMDs. Rostamina et al. detected upper facial action units using EOG-based eyewear [51]. Goel et al. proposed a tongue-in-cheek method to detect lower facial gestures using a non-contact X-band Doppler integrated into a headphone [18]. Regarding mouth movement gestures, a silent speech interface has been proposed [16, 27]. Some researchers have considered gesture sets that combine head, facial, and eye gestures. Matthies et al. developed earphone-like devices to detect 25 face-related gestures using electric sensing technologies [43]. Their Earfield sensing technique uses a contact-based four-electrode earplug. Our approach works for cases where earphones/earplugs might not be acceptable, and our sensing system does not require contact. The gesture set for the Earfield sensing technique includes facial, eye, and head gestures, focusing on one participant only. Based on its technical performance (90% accuracy in stable conditions), they developed five gestures that can be used in real-world scenarios. Amesaka et al. utilized an internal sound in the ear canal to classify 21 facial expression states [3]. Their methods showed high accuracy in classifying lower facial expression changes and head tilting. The CanalSense system can recognize jaw, face, and head movement using barometers embedded in earphones [4].

Although many techniques have been developed to recognize facial gestures utilizing various sensors, we do not yet know which facial gestures are suitable for specific tasks. Also, only a few devices can handle multiple face-related gestures simultaneously. Therefore, we propose a user-defined design of facial-related gestures for daily mobile tasks and develop a wearable solution that recognizes face-related gestures using a minimally invasive eyewear device with a sensing system that can be integrated into OST-HMDs.

3 Approach for Defining The Gesture Set

In this section, we describe the protocol of the guessability study and its outcome. Our primary goal is to define a usable face-related gesture set for daily mobile tasks and create guidelines for designing such gestures. Unlike typical guessability studies [52, 61, 63], we used two-stage studies, which are inspired by Yan et al. [65]. First, we ran a study to understand users' conceptual models and explore taxonomies to help design face-related gestures. This study included an interview with local students and an online survey. Next, we

conducted an online elicitation study. We analyzed the data and summarized the implications of designing face-related gestures. Finally, we decided a suitable gesture set for daily mobile tasks.

Face-Related Gestures We defined face-related gestures as gestures that can be made using movements of the facial muscles, eyes, and head. Most of them can be described using the FACS [14]. We excluded hand-to-face gestures because they have already been explored in previous studies [31, 55]. Face-related gestures meet the gesture criteria, as they are natural, easy to perform, and memorable [63]. First, facial gestures are natural since we use the face in daily communication. Second, they are easy to perform, as we can explicitly control the facial muscles. Third, they are intuitive and memorable because we can use various association techniques, such as the feelings produced by particular tasks and product metaphors. Users can associate the emotion of a facial expression with the one related to a specific task. Besides, this approach allows subtle, hands-free inputs regardless of environmental noise. The gestures can be performed by users who cannot speak or move their hands. We explore the use of face-related gestures as few studies have explored how people use such gestures as interaction techniques.

Tasks and Technology Table 1 shows the task list. We chose 15 daily mobile tasks relevant to Internet of Things (IoT) and smart home contexts from a previous study on mobile interaction to expand the usability of the gestures in daily life [52, 55].

Table 1: The task list shown to the participants.

Category	Number	Tasks
Call	C1	Make a call
	C2	Answer a call
	C3	Ignore a call
	C4	Hang up a call
TV	T5	Turn on a TV
	T6	Turn off a TV
Music	M7	Play music
	M8	Stop music
	M9	Raise the volume
	M10	Lower the volume
	M11	Go to next track
	M12	Back to previous track
Picture	P13	Take a picture
Light	L14	Turn on a light
	L15	Turn off a light

Considering that our target operations are daily mobile tasks that fit smart home contexts, we assume the following requirements concerning the technology:

- It can detect various face-related gestures and can send the signals to smart devices.
- Its form factor can be implemented as a wearable HMD.
- It does not interfere with daily life.

3.1 Exploration Study to Develop a Design Reference

We conducted a preliminary elicitation study in which some users described the difficulty of designing face-related gestures. Therefore, we made the study exploratory, similar to that of Yan et al. [65], to help users develop gestures they deemed difficult to design. In this study, we first interviewed twelve university students in Japan and asked them to create the gestures for nine mobile tasks proposed by Serrano et al. [55]. They were asked not to consider technical feasibility and conflicts in the gesture set. As we wanted to learn about their design process regarding the gestures, we asked why/how they chose their proposed gestures. In addition to the student interviews, we recruited participants online using MTurk. The aim was to include diverse ideas and perspectives of people with different backgrounds from students enrolled in a local university. Twenty workers described how they would use face-related gestures to control devices in smart home contexts. We asked them to identify which six gestures are a good fit (two each for eye-gestures, facial gestures, and head gestures) and which three gestures (one for each gesture) are not a good fit. The latter question helped to gain insight into how to design the face-related gestures. The participants provided text input describing the gestures.

Result We summarized the results to create a reference, which is similar to the taxonomy developed by Lee et al. [32]. Our reference includes four concepts to consider: accuracy, ease of use, social acceptance, and intuitiveness. For accuracy, two main strategies were used to avoid false positives. One involved using infrequent behaviors that do not often happen in daily life (e.g., a wink instead of a blink gesture). The other strategy was to combine gestures or repeat simple gestures. Examples are tilting the head left, then looking down and left with the eyes, or blinking twice. For ease of use, the participants proposed subtle and simple gestures. For social acceptance, socially inappropriate behaviors, such as yawning, were not recommended. For intuitiveness, associating a gesture with a metaphor for the product (i.e., habits or senses) was used. For example, to make a call, the associated head tilt motion was proposed. Associations with feelings caused by gestures and ones caused by tasks were also used. The reference describes all the aspects the users proposed for designing gestures, although the factors within the reference have trade-offs, such as accuracy and intuitiveness. The reference includes a taxonomy as an anchor point for developing gestures, such as the categories of area and flow (see Table 2).

Table 2: Taxonomy for designing face-related gestures.

Area		head, eye, mouth, eyebrow, lips, cheek, nose, tongue
Flow	Combination	one gesture, two gestures at the same time, one gesture then another gesture
	Frequency	an instant, held, twice, three times, repeatedly

3.2 Elicitation Study

We conducted an elicitation study of face-related gestures for daily mobile tasks. We showed the effect of a gesture to users and asked them to propose a gesture that would cause such an effect [52, 63].

3.2.1 MTurk Approach

For gesture elicitation, we designed an online study using MTurk. We adopted MTurk to include a broad user base, as most studies for gesture elicitation are biased in terms of specific cultural backgrounds. The advantages of using an online tool are that it is cost-effective and less time-consuming [1, 2]. We controlled the workers' qualifications for those with an approval rate of greater than 95% on human intelligence tasks (HITs) and the number of HITs approved as greater than 50. The study took 20 to 30 minutes for each participant, each of whom was rewarded with 3 US dollars for successfully completing the study. We used two yes/no validation quizzes to check for bots. However, as this process is not perfect, we manually checked the answers. If the answers were not related to emotion/ face-related movement, we removed them from our dataset. 37 participants completed the study correctly by text. The gestures

were not recorded via video as MTurk policy did not allow us to collect personal identification information, which includes the face. Table 3 provides a summary of the method.

Table 3: Summary of the approach.

	MTurk
Participants	37
Proposals	486
Gestures	612
Briefing	By text
Elicited	1 for each task
Group	Personal
Recording	Text input only
Media	Google form

3.2.2 Procedure

The participants provided text responses on a Google form that they were redirected to from MTurk. The format consisted of a briefing section, followed by a gesture elicitation section, and finally, a question asking participants to describe the pros/cons of the face-related gestures.

The briefing section first gave an overview of a smart home's context, wearable devices, what a face-related gesture is, and the tasks. Then, participants were shown a developed reference on how to design the gestures. The final part included two validation quizzes to confirm that they had read and understood the briefing. The gesture elicitation section showed a video or image of the referent and asked participants to try out the gestures they developed. For example, we showed a video of a ceiling light turning off for L15. Then, the participants entered face-related gestures that they felt were a good fit for the task. This section also asked them to evaluate their proposals in terms of whether the gesture is a good match for the task, how easy the gesture is to perform, and how obvious it is. These questions were asked to prompt them to design the gestures considering these criteria and try the gestures. The process was repeated for each task. The participants completed the proposals from C1 (make a call) to L15 (turn off a light) in numerical order. The final section asked about the pros/cons of the face-related gestures developed by the participants. This question was asked to gain insight into how the user felt about the face-related gestures.

3.2.3 Analysis and Results

With regard to the 15 tasks, the MTurk participants (N = 37) made 486 valid proposals. The number of proposals differed depending on the tasks since we eliminated invalid gestures that did not meet our definition from 61 participants.

First, we factored the proposals into the gestures. We counted repetitive gestures as one gesture. All told, the number of factored gestures was 612. Then, we manually coded them into AU combinations that refereed to the FACS [14]. Using the FACS, we coded the proposals based on the text, and on the image results of searching for the text on Google. We created categories if there were no appropriate labels for the gestures in the FACS. We coded 589 out of 612 gestures using the index. The coding was performed to categorize the gestures as either head (AU51, AU52, AU53, AU54, AU55, AU56, M59, M60), eye (AU61, AU62, AU63), upper face (AU1, AU2, AU4, AU5, AU43, AU45, AU46), or lower face (AU12, AU15, AU24, AU27) gestures. We categorized similar movements (e.g., "opening eyes" and "lifting eyebrows" movements can be associated with the same emotion—surprise) into one AU group. The results of the coding index appear in Table 4. They only show the gestures that appeared more than or equal to ten times in the coding.

Table 4: Action units index from the text proposals (N = number of proposals).

AU Grouping	**Examples**	**N**
AU12	smile, raise lips, happy	83
AU43AU45	close (shut) eyes, sleeping, blink	76
AU1AU2AU5	lift (raise) eyebrows, raise one eyebrow, open eyes, excited	62
AU15AU24	sad, lips down, unhappy face, bad face	36
AU46	wink , close one eye	33
AU27	open mouth	28
M59	head shake, head nod side to side, head tilt to double side	28
M60	head nod, head up and down	28
AU4	frown, narrow (lower, furrow) eyebrows, angry, shrink face	25
AU55AU56	head tilt (to one side, right, left)	22
AU54	head (tilt) down	20
AU51	head left	18
AU53	raise head, head tilt up, head up	18
AU52	head right	12
AU62	eyes right	12
AU61	eyes left	10
AU63	eyes up	10

Area for Interaction We categorized the proposed AUs into four groups (i.e., head, lower face, upper face, and eyes). Figure 2 shows the detailed area distribution for each task. Overall, the ratio of head-related, lower-face, upper-face, and eye-related proposals was 0.283, 0.330, 0.309, and 0.077, respectively. The participants proposed face gestures (upper and lower face) most frequently (0.638). The results suggest that face gestures have potential as interaction methods.

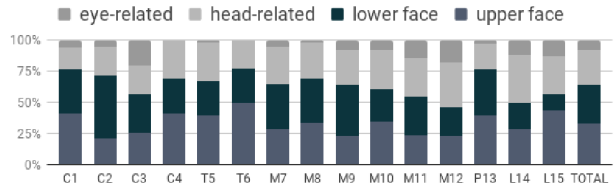

Figure 2: The proposed interaction area.

Complexity Next, we analyzed the complexity of the proposed gestures. We divided the gestures into two categories: single-site motion gestures and other complex gestures, such as a combination of gestures, sequential gestures, and repetitive gestures. Figure 3 shows the ratios of the results obtained in each experiment. The proportion of single-site motion gestures was high (average: 0.645).

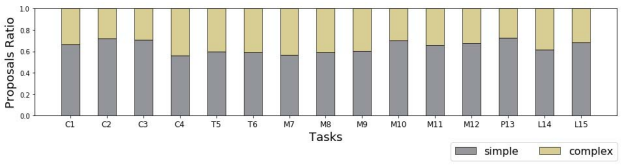

Figure 3: Complexity of the proposed gestures.

Association of Tasks with Non-Verbal Meanings We examined the relationship between tasks and gesture characteristics. Figure 4 shows the result. We considered positive or high arousal

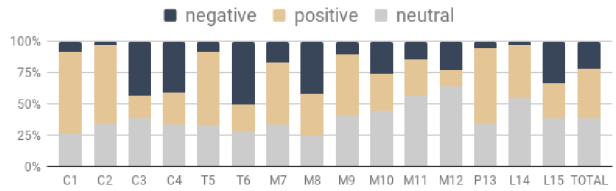

Figure 4: The positive/negative gestures included in the proposals.

gestures to be the following four types of gestures: 1) communication gestures (AU46), 2) agreement gestures (M60), 3) high arousal gestures (the movement included in the action units of surprise, such as AU1AU2AU5, and AU27), and 4) positive gestures (AU12). Negative or low arousal gestures included disapproval gestures (M59), and AUs containing in negative emotions, such as AU4 and AU15.

We conducted a Student's t-test on the starting action groups (C1, C2, T5, M7, M9, P13, L14) and ending action groups (C3, C4, T6, M8, M10, L14). The results showed a significant difference in the number of positive or high arousal gestures between the starting and ending action groups ($p < 0.01$). Similarly, there was a significant difference in the number of the negative or low arousal gestures between the two action groups ($p < 0.01$). Thus, the tactics of association differed depending on the action groups. The starting action groups tended to include the positive or high arousal gestures as they had a ratio of 2.09 (the number of positive or high arousal gestures divided by the number of the negative or low arousal gestures), while the ending action groups had a ratio of 0.198.

3.2.4 Candidates for the Gesture Set

To examine the trends in terms of the gesture proposals for each group, we picked out candidates for the gesture set. As one proposal (e.g., "wink + smile") can be coded into two or more gestures ("wink" and "smile") in our study, and a general agreement rate is difficult to use. We also think that our dataset included more noisy answers than general elicitation studies. Our validation process can not observe each MTurk participant (they may have just tried to finish quickly). Out of the ten or more gestures included in all the proposals (i.e., those shown in Table 4), scores (S) were assigned using the following formula for those that occurred three or more times for any one task. The scores were assigned after the normalization to decide on a suitable gesture for each task.

$$S_{ik} = \frac{N_{ik}}{N_i} * \frac{N_{ik}}{N_k} \quad (1)$$

where N_i is the total number of gestures proposed for task i, N_k is the total number of gesture k proposed, and N_{ik} is the number of gesture k proposed for task i. After calculating S_{ik}, which is the score of gesture k for task i. For example, if we consider a smile gesture for making a call (N_{ik}), the number of proposed smiles is 10. The number of proposed smiles for all referents (N_k) is 83, and the number of proposed gestures for making a call (N_i) is 34. Therefore, score S_{ik} = 10/34*10/83 = 0.0354. Our agreement score considers a bias of suggesting gestures that are easy to come up regardless of the referent. Table 5 shows the gestures with the three highest scores for each task. In this table, we used the first example gesture of each AU in Table 4 (e.g., smile instead of AU12) for ease of understanding. Table 5 also shows a raw percentage agreement (the number of a gesture divided by the number of gestures proposed for each referent.)

According to the table, participants tended to design gestures with reference to the existing interfaces and the non-verbal information that they associated with a task (e.g., opening the mouth to make a call). In particular, they suggested gestures that required moving a part of the face in the same direction as the interface of a task operation (e.g., turning right for the "next song" or raising an eyebrow to "turn up the volume"). Additionally, there was a tendency for gestures to be suggested as pairs in the same area categories for paired tasks. For example, gestures proposed for turning on/off a light included "raising the head" and "lowering the head." Both of them involved moving the head in opposite directions.

Table 5: Candidates for the gesture set [gesture name (our score * 10^2, the raw agreement rate)].

Task	Gesture Proposals
C1 Make a call	smile (3.54, 0.29), lift eyebrows (1.71, 0.18), wink (1.43, 0.12)
C2 Answer a call	smile (6.21, 0.67), open mouth (0.85, 0.08), lift eyebrows (0.68, 0.11)
C3 Ignore a call	sad (3.59, 0.18), shake head (2.29, 0.13), frown (0.92, 0.08)
C4 Hang up a call	close eyes (3.37, 0.26), frown(2.56, 0.13), nod head (1.47,0.10)
T5 Turn on a TV	lift eyebrows (3.04, 0.21), smile (2.27,0.21), raise head (2.07, 0.09)
T6 Turn off a TV	close eyes (4.74, 0.3), shake head (2.23, 0.13), sad (1.79, 0.13)
M7 Play music	smile (3.64, 0.28), shake head (1.43, 0.1), close eyes(1.18, 0.15)
M8 Stop music	sad (2.34, 0.14), frown (2.27, 0.11), open mouth(1.3, 0.09)
M9 Raise the volume	lift eyebrows (3.44, 0.24), smile (1.55, 0.18), open mouth(0.85, 0.08)
M10 Lower the volume	frown (2.7, 0.14), head left (2.4, 0.11), head down(2.16, 0.11)
M11 Go to next track	head right (5.08, 0.12), eyes right (5.08, 0.12), sad(1.11, 0.10)
M12 Back to previous track	head left (6.98, 0.13), eyes left (6.41, 0.13), sad (1.17, 0.10)
P13 Take a picture	wink (2.87, 0.16), smile (1.55, 0.18), open mouth (1.5, 0.11)
L14 Turn on a light	raise head (5.00, 0.15), smile (1.48, 0.18), head nod (1.43, 0.10)
L15 Turn off a light	close eyes (3.37, 0.26), head down (1.15, 0.08), wink (0.7, 0.08)

3.2.5 Discussion on Face-Related Gestures

In the final section of the elicitation study, We asked MTurk users the following question: "What do you think are the advantages/disadvantages of face-related gestures over other input modalities (e.g., voice input)?" We also asked the local university students (N = 11) this question. We summarized the pros and cons of face-related gestures from the participants' comments and proposals. Many participants commented on the ability to perform the gestures. Since all of us have heads, disabled people who cannot speak can also make face-related gestures. One participant stated that face-related gestures are not limited to a specific language. Some argued that the gestures could be silent, which means that they do not bother people near the user and provide security, thus allowing them to be used in a public space, such as a library. Many mentioned the ease of performing the gestures. They said that they take less effort and save time as they are fast and easy to perform. One participant noted that this helps multitask because the actions can be performed mechanically if the user gets used to the gestures. Regarding mouth gestures, silent voice gestures indicating "yes" and "no" were included in the elicitation study. This silent voice approach is a useful face-related gesture as it can compensate for the shortcomings of voice inputs. However, many participants expressed concern that face-related gestures may trigger unwanted tasks since some of the

gestures are often performed in daily life. One participant stated that input gestures could bother or be misunderstood by people facing the user. These concerns suggest that a trigger command is necessary to avoid such problems. Another concern is memorability, as users will need to remember which facial gestures are used for which tasks. One person preferred voice commands because they can be performed without any prior programming or adjustment. One suggested method was to use voice inputs for the face-related gestures until users get used to them so as they can take time to adjust to the new technique.

3.2.6 Elicitation Study Takeaway

Based on this analysis, gesture designers should consider the following recommendations with regard to face-related gestures. From the MTurk feedback, it seems that facial gestures are acceptable to the users, as is the use of head and eye movements. Users tend to propose rather simple gestures used in everyday life (e.g., a broad smile), but this leaves room for misinterpretation and potential high false-positive recognition. According to the observation in the exploration study, the proposed gestures are slightly more exaggerated than naturally occurring facial expressions. Yet, it remains to be seen if we can detect them in unconstrained everyday recordings. Possible ways to prevent misclassification are the use of segmentation gestures or limiting the gestures to a specific location (people working on a computer or reading usually do not show strong facial expressions, etc.). Although we showed two different strategy references ("Use infrequent behaviors to avoid false positives" vs. "Use simple and common gestures to increase the ease of use"), as mentioned, most of the proposed gestures are simple. We also advise avoiding socially unnatural behavior, as seen in the exploration study. Second, we suggest using the metaphors associated with tasks. A possible metaphor for a task is related to the situation in which it arises, emotions, non-verbal gestures, or an existing interface for executing the task. Finally, for paired tasks, such as turning a television on/off or music volume up/down, paired gestures should be used. Gestures can be paired using contrasting or same notions using the same facial area category. We think that this helps make gestures more intuitive and memorable.

3.3 Finalized Gesture Set

We created the finalized gesture set using the gesture candidates in Table 5 and following the above takeaway. Table 6 shows the result. We mainly selected from the first candidate in each case in Table 5. Then, we added the second option for C1 and C2 to avoid gesture conflicts in the same category, and for L15 to form a pair with L14. For M11, we had two candidates with the highest score. We added the paired gesture of M11 for M12 as a second option. We think that second options can be combined with first candidates to trigger tasks to avoid false-positives.

4 Proof of Concept

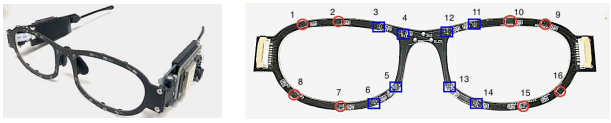

Figure 5: The appearance (left) and the sensor layout (right) of the device. The sensors circled in red have a longer focal length than the ones in the blue rectangles.

Given that the gesture set included all facial, eye, and head gestures, we searched for a method that could detect these gestures with a wearable device. A head gesture can be detected by adding an IMU, as it has already been evaluated in the previous studies [13,65]. Therefore, a method that could detect both face and eye movements using another sensing modality was desirable for the evaluation. To this end, we focused on Masai et al.'s method [38, 40, 41], which uses the skin deformation around the eyes through photo-reflective sensors on smart eyewear. Such deformation occurs during both facial movements and eye movements, and we think that this method can detect both types of gestures. Since the sensors they used have small form factors (e.g., NJL5901AR-1-TE1 is 1.3 mm x 1.6 mm x 0.6mm), the device is wearable, allowing users to perform daily mobile tasks without it restricting their pose or position. Moreover, the form factor of eyewear is socially acceptable. This is an essential factor for practical use, and recent eyewear computing research has considered such a form factor [40, 51, 57]. Because of the small form factor, the system has the potential to be used in a wide range of applications since the sensors can be integrated into commercial HMDs or everyday eyeglasses because of their small form factor [56,64]. Other advantages of photo-reflective sensors are their low cost, contactless characteristic, and fast processing speed.

Table 6: The final result for each task

Task	Final Results
C1	smile (lift eyebrows)
C2	smile (open mouth)
C3	sad
C4	close eyes
T5	lift eyebrows
T6	close eyes
M7	smile
M8	sad
M9	lift eyebrows
M10	frown
M11	head right (eyes right)
M12	head left (eyes left)
P13	wink
L14	raise head
L15	close eyes (head down)

In addition to determining user-friendly gestures that are user-defined, we improved the hardware and software in their facial expression recognition method [40, 41] and eye gestures classification method [38]. For the hardware, we changed the sensor configuration to incorporate facial movements and eye movements simultaneously. We added an IMU to make the classification robust in terms of head movement noise. For the software, we used a CNN to consider the detection of time-series gestures, which is robustly distinguishable from unintended sensor changes caused by natural movements and gestures [53]. Our method has advantages concerning deliberate gesture inputs to interact with smart devices in daily life.

4.1 Hardware

We used an eyewear prototype for our technical evaluation (see Figure 5). The device follows the same sensing principle used in a previous study [40]. The main parts of the device are photo-reflective sensors, an IMU sensor, and micro-controllers. The IMU sensor (Adafruit BNO055 Absolute Orientation Sensor) is located on the side of the glasses. The IMU transmits a four-dimensional quaternion. Two microcontrollers, one located on each side of the glasses, are connected with I2C communication to handle the analog inputs of 16 photo-reflective sensors. One is an ESP32-based microcontroller (Adafruit HUZZAH32) that functions via Bluetooth, and another is a Pro Micro, which acquires data from the IMU. The IMU data includes the effects of head movements and the effects of changes in the device caused by changing facial expressions. This device is driven by a 3.3-volt battery.

A photo-reflective sensor consists of an infrared LED and a phototransistor. The sensors measure proximity through reflection intensity. Figure 5 shows the sensor layout. We placed 16 photo-reflective

sensors (i.e., eight NJL5901AR-1-TE1 sensors [a short focal length] and eight NJL5909RL-4 sensors, both produced by the New Japan Radio Co. Ltd.) on the front frame of the eyewear prototype. The sensors measure reflection intensity, which is changed by skin deformation around the eyes caused by eye movements and upper and lower facial muscle movements, and mouth movements. We adopted two kinds of sensors with different focal lengths for stability. We put the sensors with the longer focal length on the far side of the frame because the curvature of a face changes the distance range measured by the sensors. Additionally, to improve light sensitivity, we used higher register values (200k ohm) for the phototransistors of the sensors that measure longer distances at the end of the front frame than for the ones close to the center (47k ohm).

The device acquires a 20-dimensional data sample per reading. The sampling frequency is 100 Hz. The data are sent via Bluetooth Serial (at a baud rate of 115,200 bps) to a laptop, where the data are stored and processed. We applied a CNN to the 120 sequential data samples to classify temporal gestures.

4.2 Software

4.2.1 Data Processing

After smoothing out noise by applying a moving average of 10 frames, we set the window size to 120 and shifted the window by 10 frames. For each data, we applied "change detection" with different threshold values to extract each gesture's range. We used three thresholds to augment the data if the estimated range was different. After obtaining the respective training data features, we resized each sensor's time series data with bilinear interpolation from the OpenCV library. It unifies the sizes of all the gesture data to 52 (features) x 120 (frames) dimensions. In addition to gesture data, we used the noise data for the non-gesture class. We applied the change detection algorithm with different thresholds and feature extractions, and then we resized to a window of 120 frames every ten frames.

4.2.2 Change Detection

Change detection finds the beginning and end of changes in certain windows. First, the algorithm divided the data samples in the window every five frames. Then, it calculated the *Changed Value* in the frames for each sensor dimension as follows:

$$ChangedValue = sum(sqrt((d2)^2)) \quad (2)$$

where *d2* is the difference between the sensor values in two frames. If the average of the *Changed Value* in the five frames/total *Changed Value* in the window was larger than the threshold, the algorithm regarded it as a changed frame. It regarded the range between the first and last of the changed frames as the range of change in the window. If the algorithm could not find the changed frame, it returned the whole data sample to the window. Other factors, such as head motions and blinking, cause changes, but this algorithm aims at avoiding false negatives rather than false positives. This is because false positives can be filtered out for being non-gestures by the gesture classifier later on.

4.2.3 Feature Extraction

After the range was specified, we extracted the features of 52 (16 + 16 + 16 + 4) dimensions for each time axis. The first 16 dimensions are values obtained from optical sensors within the range, which were standardized to a zero mean and unit standard deviation. The next 16 dimensions were differential values obtained from the optical sensors within the range, which were standardized to a zero mean and unit standard deviation. The other 16 dimensions were optical sensor values, which were subtracted by each initial value and then divided by the standard deviation of the data samples from the whole window. The last four dimensions' values were standardized to a zero mean and unit standard deviation within the range obtained from the IMU. This feature extraction boosts the training of a CNN.

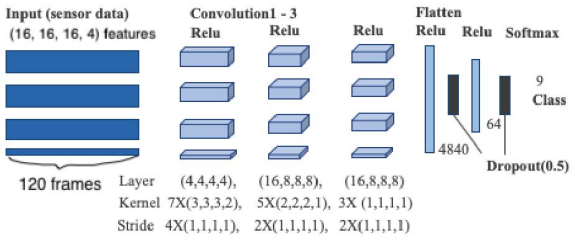

Figure 6: The CNN architecture for gesture recognition.

4.2.4 Network Architecture

We used a CNN for gesture detection and classification. A CNN is used for activity recognition in temporal sensor data. The approach outperforms handcrafted features and shallow feature learning algorithms, such as support vector machines [66]. For multimodal deep learning algorithms, the modality-specific architecture showed higher accuracy [48]. Based on this, our CNN architecture is the modality-specific as shown in Figure 6. For each sensor modality and processing feature, the channels were divided and convoluted. We used three convolutional layers for each channel with Rectified Linear Unit (ReLU) activation. Convolution was applied to both the temporal and sensor dimensions since close sensor data correspond with each other. For strides, we used the temporal direction of each sensor dimension. Each layer was normalized with L2 regularization. Then, the algorithm concatenated the layers and constructed a fully connected layer connected to another fully connected layer. Finally, it was connected to the last layer associated with the gesture classes (eight gesture classes and one non-gesture class). For the last layer, we applied softmax activation. The optimizer of the learning rate was Adam [28], and the CNN was trained with 50 epochs and a batch size of 256. We implemented the algorithm using Python and Keras. The weight W of the gesture class was calculated using the following formula when the amount of non-gesture class was large. In the formula, the *noise* is the amount of non-gesture data, and *all* is the total volume of data. The weights of all the gesture classes were the same as we assumed that there were no large differences in the amount of data among the gesture classes. G corresponds to the number of gesture classes, which is eight in the evaluation.

$$W = log(((noise/all) * G)/(1 - (noise/all))) \quad (3)$$

For detection, we made three inputs from data samples in each window using three different thresholds (0.01, 0.02, and 0.03) for "change detection." Then, we applied the CNN to these inputs. Each output contains a nine-dimensional vector (eight gesture classes and one non-gesture class) in the range of 0 to 1. Each class of the outputs was smoothed using the three time series sequences of moving averages. Finally, the algorithm multiplied the outputs for each class in the same time frame. If one of the gesture class values was greater than 0.5, then it outputs the class. Otherwise, it outputs the non-gesture class. As long as the output class was the same in the time series, the algorithm regarded them as one gesture. This process can reduce false positives due to unexpected data behavior while performing gestures.

5 TECHNICAL EVALUATION

To verify that the user-defined face-related gestures are technically feasible in a minimally invasive way, i.e., using the technology that can be potentially integrated into a wide range of wearable devices (OST-HMDs, HMDs, eyewear, etc.), we evaluated accuracy and precision of intentional temporal face-related gestures using an optical sensing system integrated into the eyewear device. This sensing technique is more affordable than comparable methods and can be

Figure 7: Eight face-related gestures for evaluation.

applied to other wearable devices as the hardware configuration, including sensors and microcontrollers, is fairly simple. We have summarized this section's results below:

- The average F1 score when detecting and classifying the 8 kinds of gestures with 13 participants was 0.911.
- The average accuracy of the classification with user-independent training was 61.4%.
- The one-hour recording of making two gestures every four minutes showed F1 scores of 0.921 and 0.867 for wink and smile gestures, respectively.

5.1 The gesture set used for the evaluation

We selected the gestures from Table 5 for the evaluation as follows: AU12 (smile), AU27 (open mouth), LP (lip movements related to negative emotions), AU1+2 (lift eyebrows), AU4 (frown), AU43 (close eyes firmly), AU46 (right eye wink), and AU61 (eyes left). We excluded head gestures that have been evaluated using an IMU in the previous research [13, 65]. This gesture set is novel compared to existing validated gesture sets for a smart eyewear [38, 40] in terms of considering both facial and eye gestures, and including mouth only movement. We assumed that all the gestures start from and end with a neutral facial expression. By creating this deliberate flow, we aimed to distinguish gestures for manipulation from spontaneous ones. We included only one of each of the paired gestures as the classification of another paired gesture will be possible if one gesture can be classified. For example, "move eyes right (AU62)" can be classified if "move eyes left (AU61)" is classified correctly. For AU43, we asked the participants to close their eyes firmly to avoid natural blinks being classified as AU43. We did not consider the quality of gesture execution as some participants may not be able to make certain gestures, such as a wink. Instead, we evaluated whether the detection of their intended gestures was feasible.

5.2 Participants and Data Acquisition Procedure

We recruited 13 participants (9 males and 4 females) in their twenties via word-of-mouth sampling at a university in Japan. All the participants are Asian, and they were each compensated with approximately ten US dollars. We ran all the recordings sequentially in the laboratory after receiving approval from the bioethics committee of Faculty of Science and Technology / Graduate School of Science and Engineering, Keio University.

Procedure Each participant sat on a chair in front of a laptop, which was placed on a desk. Participants wore the prototype with an eyewear band strap for stability. The experimenter introduced the prototype and confirmed it worked adequately for each participant. If sensor values were saturated, the experimenter adjusted the size of one or both of the nose pads. The experimenter then explained that they would make 8 different face-related gestures 20 times each in a limited time (160 gestures in total). The experimenter told them that each gesture should start and end with a neutral facial expression. Next, the experimenter gave instructions for using the recording software. Each gesture recording consisted of two phases. In the preparation phase (2,000 ms), the software indicated the next gesture with a word and image. In the action phase (1,500 ms), the software asked them to perform the gesture. The software recorded sensor values, gesture labels, and timestamps during the action phase. The experimenter asked the participants to make gestures quickly, as the recording time was limited. The software recorded the eight kinds of gestures in a periodic order to prevent participants from making the wrong gestures. The software was implemented using the Processing language. These recordings were used later for training. Then, the experimenter recorded noise data for 30 seconds while each participant 1) was seated and in a relaxed state, 2) moved his or her face, head, and body, and 3) walked around the room. These noise data were assigned to the non-gesture class for training. Later, the experimenter asked the participants to make all 8 types of gestures 4 times within 90 seconds. The order of the gestures was randomized and given by written instructions. In this session, participants made the gestures at their own pace. The process was repeated three times while participants were sitting on a chair. Overall, 96 gestures (eight types x four times x three sessions) were made by each participant. We recorded sensor data and videos of the participants with the laptop's built-in camera and placed timestamps to synchronize the images and sensor data. These recordings were later used as test data. The test dataset is closer to a realistic scenario than training data. It contains not only complete gesture data but also gesture transition data and gesture data that are missing the information at the start or the end.

5.3 Results

For gesture recordings in the first session, each gesture data had approximately 150 frames. Two or three sets of training data were generated from each set of gesture data. Overall, around 2,500 sets of training data were collected from each participant, including the augmented data, and about two-thirds were from the non-gesture class. The ratios of the sample numbers are almost the same for the eight gesture classes. The test data consist of three recordings of 90 seconds with a 100 Hz sampling rate. As the window is shifted for every ten frames, the size of the test data for each participant leads to approximately 2,640. We checked the outputs by superimposing the classification result on the captured images in the recording of the test data (the training dataset and the test dataset were recorded with different procedures). The timing was matched using the timestamps. Figure 8 shows the F1 score of classifying the gestures in the test data while the users sat on a chair. After the CNN was trained with all the participants' training datasets, the whole architecture was retrained with each participant's training dataset and tested independently. Since the number of non-gesture classes was larger than the others, we do not show them in Figure 8. Each participant performed 8 kinds of gestures 12 times. The average F1 score was 0.911, ranging from 0.827 to 0.994. Table 7 summarizes the results by gesture. AU46 (right wink) had the best score, and AU1+2 (raising eyebrow action) had a high F1 score of 0.966.

We observed specific patterns of false positives for the gestures. For example, the end of P9's eye-closing gesture was misrecognized as a gesture that narrows the eyebrows. The end of P10's smile gesture was misrecognized as a gesture of opening the mouth. The end of P12's opening the mouth gesture was misrecognized as a lip

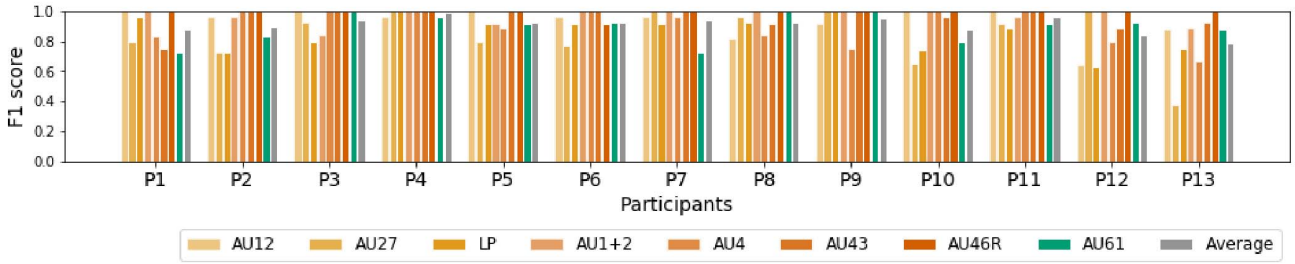

Figure 8: F1 scores of detecting and classifying eight kinds of temporal face-related gestures for each participant.

Table 7: Average scores of all participants (R: recall, P: precision, F1: F1 score).

	AU12	AU27	LP	AU1+2	AU4	AU43	AU46R	AU61	Average
R	0.955	0.904	0.910	0.987	0.968	0.974	0.994	0.917	0.951
P	0.903	0.762	0.798	0.945	0.825	0.938	0.994	0.867	0.875
F1	0.928	0.827	0.850	0.966	0.891	0.956	0.994	0.891	0.911

gesture. The data near the end of the gestures tended to be classified incorrectly as the ends of some gestures were lacking in the training dataset due to the short recording time for each gesture's data.

We also tested performance with user-independent training. We trained the CNN (10 epochs) using 15,000 randomly selected training data (gesture and non-gesture classes) from 12 participants. The CNN predicted the gesture class data (8 kinds x 20 times) of the users whose data was not used for the training. The gesture class data were processed using "change detection" with a threshold of 0.02, using 120 frames from the beginning of each trial's data. We used the highest value of the CNN output as the prediction result. We repeated this 10 times for all 13 participants. Figure 9 shows the mean accuracies and standard deviations. The overall mean accuracy was 61.4% (SD = 11.6%). The detection was a challenge as 44.5% of the data from P6 (40.0% accuracy) and 49.0% from P13 (40.5% accuracy) were regarded as being part of the non-gesture class.

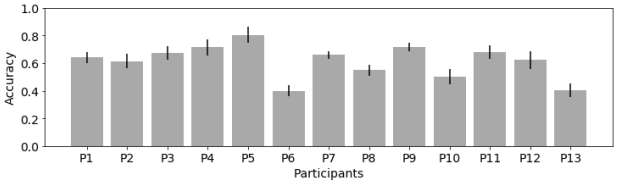

Figure 9: Performance with user-independent training.

5.4 One-Hour False Positive Test

We further explored the false positives of two gestures (AU46 and AU12). For this exploration, the experimenter performed the recording. After recording the gestures 20 times and 90 seconds of the non-gestures, the CNN was trained with 3 classes. Then, the experimenter performed a one-hour recording with the device. The alarm was set every four minutes, and each time the alarm went off, the experimenter winked twice and smiled twice. Including the gestures at the beginning of the recording and one additional wink in the middle, the experimenter performed a total of 31 winks and 30 smiles. During the recording, the experimenter either sat on a chair, angled his head, stood and stretched, walked around, watched a movie that induced smiles, or surfed the internet using a computer. Despite the experimenter correcting the glasses' position occasionally and performing daily activities, 29 out of the 31 winks and 26 out of the 30 smiles were detected correctly. There was one false positive concerning a wink during the recording. It was caused by scratching the right side of the nose, which was not included in the training data. The F1 scores were 0.921 and 0.867 for AU46 and AU12, respectively. This supports the possibility of using such gestures in real-life settings.

6 Discussion

The technical evaluation showed that the average recall of face-related gestures was 0.951, but the average precision was 0.875. The main issue was false positives, which caused unintentional commands to be detected.

First, the false detection of gestures can be influenced by the data included in the training dataset. If we incorporate data that are likely to be mistaken as target gestures, the network can learn the difference. For example, the blink action included in the test dataset was not mistakenly recognized as closing the eyes firmly. These gestures are visually similar, but the network learned the difference based on the speed and strength of spontaneous blinks and intentional eye closing. Therefore, if the spontaneous movement is included in the training dataset as noise data, the algorithm can differentiate it from intentional gestures. This data collection can reduce false positives, but it might be costly. Also, some users' gestures with regard to closing their eyes and lowering their eyebrows had a visual similarity. While these gestures were classified correctly most of the time, they did cause some false positives. The algorithm can consider such patterns and improve the F1 scores by improving the network architecture and data acquisition protocol.

Second, the independence of gestures needs to be considered to avoid false positives. For example, when users were instructed to open their mouths, they sometimes raised their eyebrows at the same time. However, they did not open their mouths when instructed to raise their eyebrows. Another user lowered his or her eyebrows in conjunction with closing his or her mouth. Both movements are associated with emotions (surprise and anger), and the accompanying actions are also parts of the associated emotions. This synchronicity may happen, so it is helpful to use synchronized movements for the same command. This prevents unintentional commands but may make user-defined gesture sets non-intuitive.

Third, we could use an infrequent gesture, such as a wink, as a trigger command, and this was shown to give a high F1 score in the one-hour study. A trigger command could make detection more robust but could also make the gesture flow more complex.

Finally, in addition to a trigger command, it is crucial to understand which object the user intends to operate and which object the user is currently operating to avoid mis-interactions. OST-HMDs allow users to confirm their intentions on displays. Also, research

efforts have been made to sense users' context for such AR systems [19]. We expect that combining these technologies will help avoid false positives and improve the user experience.

7 Limitations and Social Implications

As outlined above, facial expressions have a lot of positive attributes (hands-free, subtle, etc.). They can be used by people who cannot move their bodies and provide them with the tools to interact with smart devices. We considered the device design's social acceptability, but our study did not address the sociological aspect of performing gestures. Social acceptability is essential to improving the user experience. As there are various aspects of social acceptability, we need to develop appropriate measures for each aspect in naturalistic settings [29]. This is outside of the scope of this paper; however, we need to validate gestures in terms of social acceptability. According to Rico and Brewster [50], a user's willingness to perform gestures depends on the environment and the audience. We think social acceptability will be different for each face-related gesture. After defining the gesture set, we need to explore when and with whom users would be motivated to use a specific gesture. Also, encouraging the use of certain facial gestures with our technology may alter the role and usage of facial expressions in human-to-human communication. We are already observing similar effects and considerations to change the technology regarding speech assistants [35]. The effects of facial gestures on human communication should therefore be carefully considered. Another concern regarding sociological aspects is how other people interpret intentional face-related gestures. How will the design of the wearable device affect usability and social acceptance? Obstructing or covering part of the face might make the user accept facial gestures more while harming non-verbal communication. The effects need to be studied in situ and will be dependent on application cases, the environment, and the company the person is in. Even though these considerations are outside of this paper's scope, we provide an exploration of such use cases and a working prototype implementation.

We developed a gesture set of face-related gestures for daily mobile tasks. This set may require adjustment depending on the culture and each user's preferences because such factors influence the muscular control of facial muscles. As a follow-up to the technical evaluation, the experimenter asked each participant about his or her preferred gestures. They tended to prefer the gestures AU12 or AU27 and stated that AU4 and LP are challenging to perform. We tested the gestures with Asian participants, so determining whether this tendency is present in other cultures requires further research. Also, the preference can be influenced by how easily the technology recognizes the gestures.

For the elicitation study, we could not record the MTurk participants visually. Instead of visually validating the data, we collected a lot of text data to confirm the trend. Therefore, there may have been a bias in linking texts with AUs. Also, the gesture proposals were coded by one of the authors only. We made the grouping clear by creating an index (Table 4), but the grouping of AUs was subjective.

Moreover, due to the high noise ratio, our use of MTurk had room for improvement. We had to collect data from approximately three times the target number of participants. The noise answers included nonsensical responses, such as just "good" or "yes," even though the participants received explicit instructions to propose gestures. Some responses included the description of the image/video of the referent. We assume that many assignments on MTurk involve the labeling of pictures for machine learning and that they responded to the images and videos without reading the briefing. Therefore, the briefing could be more direct and succinct. Our validation quizzes could be improved since they were just two yes/no questions, which could have been answered without reading the briefing. We could reduce noise answers if we divided the tasks into explicitly simple ones, similar to those used in machine translation [6].

Head motion, positional drift, walking, and re-worn conditions can influence data collected from optical sensors on an eyewear device [40]. Our algorithm considers the relative change in a time series to lessen the influence of initial sensor states in a sampling window. We undertook a case study of a one-hour recording where the experimenter made indoor activities. The study showed a promising result in terms of adjusting for a shift. Still, we need to evaluate such influences with our algorithm. Additionally, the sensors employed for the system are vulnerable to high-intensity light, limiting the system's use to indoor settings. To use it outdoors, we need to develop covers to protect the sensors from ambient light or filters, utilizing the synchronous detection technique.

With the current algorithm, the sensor data are classified into one of nine classes. The system does not consider multilabel gestures simultaneously (i.e., two or more). If a user makes two gestures simultaneously, the resulting sensor data could be categorized as either of them or in the non-gesture class. This limits diversity in terms of the gestures. Also, adding new gesture classes will require re-training the architecture. We are considering making detectors for each class (eye movement, mouth movement, etc.) by incorporating multiple CNN architectures. Another way to address the customizability issue is to use a similarity-based approach, such as one-shot learning. We explore the algorithms in future work.

The current dataset only contains 13 users, which impacts its generalizability. Our main result of the technical evaluation used the test data and training data from the same users. Our results of user-independent training suggest the potential for generalizability, but the mean accuracy was only 61.4%. Our system requires individual user training. Given that we are extending the use of glasses, an already individualized accessory, we think this limitation is not so severe. Our system requires a calibration phase, like for other wearable devices such as eye-tracker. If we assume people use personal VR/AR glasses (and do not share them), a personalized classifier might be acceptable as a one-time adjustment when the user sets up the system. The prototype system could be used as it is for this setup. We are considering transfer learning (e.g., a domain adaptation framework [26]) to reduce the cost of user-dependent training processes and make the system generalizable. We want to investigate how to shorten the calibration by analyzing the influence of the training dataset's size on accuracy. In the future, we can imagine similar auto-calibration approaches, such as those in place for eye-tracking systems.

8 Conclusion

In this paper, we developed and tested face-related gestures as interaction methods. We conducted an elicitation study using MTurk. We coded the text proposals into AUs manually. The gestures were considered to be hands-free, subtle, and able to be used in various settings by the participants. Based on the results, we summarized the findings and developed a user-defined gesture set. Then, we evaluated the feasibility of face-related gestures in the developed set using an eyewear device. We used a CNN for detection and classification. The average F1 score when detecting and classifying the eight kinds of gestures with 13 participants was 0.911. The mean accuracy of the classification with user-independent training was 61.4%. The one-hour recording of making two gestures every four minutes showed F1 scores of 0.921 and 0.867 for wink and smile gestures. This finding suggests the possibility of the real-life use of the gestures. In future work, we plan to integrate the technique with an HMD prototype to explore the user experience of the gestures, such as their social acceptability.

Acknowledgments

The authors wish to thank the reviewers. This work was supported by JST AIP-PRISM JST AIP-PRISM Grant Number JPMJCR18Y2 and JSPS KAKENHI Grant Numbers JP18H03278, and JP16H05870.

REFERENCES

[1] A. X. Ali, M. R. Morris, and J. O. Wobbrock. Crowdsourcing similarity judgments for agreement analysis in end-user elicitation studies. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, UIST '18, pp. 177–188. ACM, New York, NY, USA, 2018. doi: 10.1145/3242587.3242621

[2] A. X. Ali, M. R. Morris, and J. O. Wobbrock. Crowdlicit: A system for conducting distributed end-user elicitation and identification studies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 255:1–255:12. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300485

[3] T. Amesaka, H. Watanabe, and M. Sugimoto. Facial expression recognition using ear canal transfer function. In *Proceedings of the 23rd International Symposium on Wearable Computers*, ISWC '19, p. 1–9. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3341163.3347747

[4] T. Ando, Y. Kubo, B. Shizuki, and S. Takahashi. Canalsense: Face-related movement recognition system based on sensing air pressure in ear canals. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, UIST '17, p. 679–689. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3126594.3126649

[5] S. S. Arefin Shimon, C. Lutton, Z. Xu, S. Morrison-Smith, C. Boucher, and J. Ruiz. Exploring non-touchscreen gestures for smartwatches. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 3822–3833. ACM, New York, NY, USA, 2016. doi: 10.1145/2858036.2858385

[6] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich. Soylent: A word processor with a crowd inside. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pp. 313–322. ACM, New York, NY, USA, 2010. doi: 10.1145/1866029.1866078

[7] A. Bulling, D. Roggen, and G. Tröster. It's in your eyes: Towards context-awareness and mobile hci using wearable eog goggles. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, UbiComp '08, pp. 84–93. ACM, New York, NY, USA, 2008. doi: 10.1145/1409635.1409647

[8] E. Chan, T. Seyed, W. Stuerzlinger, X.-D. Yang, and F. Maurer. User elicitation on single-hand microgestures. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 3403–3414. ACM, New York, NY, USA, 2016. doi: 10.1145/2858036.2858589

[9] J. Cheng, A. Okoso, K. Kunze, N. Henze, A. Schmidt, P. Lukowicz, and K. Kise. On the tip of my tongue: a non-invasive pressure-based tongue interface. In *Proceedings of the 5th Augmented Human International Conference*, pp. 1–4, 2014.

[10] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1548–1568, Aug 2016. doi: 10.1109/TPAMI.2016.2515606

[11] A. Crossan, M. McGill, S. Brewster, and R. Murray-Smith. Head tilting for interaction in mobile contexts. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '09. Association for Computing Machinery, New York, NY, USA, 2009. doi: 10.1145/1613858.1613866

[12] M. Dhuliawala, J. Lee, J. Shimizu, A. Bulling, K. Kunze, T. Starner, and W. Woo. Smooth eye movement interaction using eog glasses. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 307–311, 2016.

[13] C. Dierk, S. Carter, P. Chiu, T. Dunnigan, and D. Kimber. Use your head! exploring interaction modalities for hat technologies. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, pp. 1033–1045. ACM, New York, NY, USA, 2019. doi: 10.1145/3322276.3322356

[14] P. Ekman and W. Friesen. *Facial action coding system*. Consulting Psychologists Press, Stanford University, Palo Alto, 1977.

[15] K. Fukahori, D. Sakamoto, and T. 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, pp. 3019–3028. ACM, New York, NY, USA, 2015. doi: 10.1145/2702123.2702308

[16] M. Fukumoto. Silentvoice: Unnoticeable voice input by ingressive speech. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, UIST '18, pp. 237–246. ACM, New York, NY, USA, 2018. doi: 10.1145/3242587.3242603

[17] Y. Gizatdinova, O. ï¿œpakov, and V. Surakka. Face typing: Vision-based perceptual interface for hands-free text entry with a scrollable virtual keyboard. In *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*, pp. 81–87, Jan 2012. doi: 10.1109/WACV.2012.6162997

[18] M. Goel, C. Zhao, R. Vinisha, and S. 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, pp. 255–258. ACM, New York, NY, USA, 2015. doi: 10.1145/2702123.2702591

[19] J. Grubert, T. Langlotz, S. Zollmann, and H. Regenbrecht. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 23(6):1706–1724, 2017.

[20] J. Gugenheimer, D. Dobbelstein, C. Winkler, G. Haas, and E. Rukzio. Facetouch: Enabling touch interaction in display fixed uis for mobile virtual reality. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, pp. 49–60. ACM, New York, NY, USA, 2016. doi: 10.1145/2984511.2984576

[21] Y. Iravantchi, Y. Zhang, E. Bernitsas, M. Goel, and C. Harrison. Interferi: Gesture sensing using on-body acoustic interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 276:1–276:13. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300506

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

[23] R. Jota and D. Wigdor. Palpebrae superioris: Exploring the design space of eyelid gestures. In *Proceedings of the 41st Graphics Interface Conference*, GI '15, pp. 273–280. Canadian Information Processing Society, Toronto, Ont., Canada, Canada, 2015.

[24] S. Kanoh, S. Ichi-nohe, S. Shioya, K. Inoue, and R. Kawashima. Development of an eyewear to measure eye and body movements. 08 2015. doi: 10.1109/EMBC.2015.7318844

[25] T. Kikuchi, Y. Sugiura, K. Masai, M. Sugimoto, and B. H. Thomas. Eartouch: Turning the ear into an input surface. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '17, pp. 27:1–27:6. ACM, New York, NY, USA, 2017. doi: 10.1145/3098279.3098538

[26] K. Kikui, Y. Itoh, M. Yamada, Y. Sugiura, and M. Sugimoto. Intra-/inter-user adaptation framework for wearable gesture sensing device. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, ISWC '18, p. 21–24. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3267242.3267256

[27] N. Kimura, M. Kono, and J. Rekimoto. Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 146:1–146:11. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300376

[28] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[29] M. Koelle, S. Ananthanarayan, and S. Boll. Social acceptability in hci: A survey of methods, measures, and design strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–19. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376162

[30] Y. Lai, B. Tag, K. Kunze, and R. Malaka. Understanding face gestures with a user-centered approach using personal computer applications as an example. In *Proceedings of the Augmented Humans International Conference*, pp. 1–3, 2020.

[31] D. Lee, Y. Lee, Y. Shin, and I. Oakley. Designing socially acceptable

hand-to-face input. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, UIST '18, pp. 711–723. ACM, New York, NY, USA, 2018. doi: 10.1145/3242587.3242642

[32] J. Lee, H.-S. Yeo, M. Dhuliawala, J. Akano, J. Shimizu, T. Starner, A. Quigley, W. Woo, and K. Kunze. Itchy nose: Discreet gesture interaction using eog sensors in smart eyewear. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ISWC '17, pp. 94–97. ACM, New York, NY, USA, 2017. doi: 10.1145/3123021.3123060

[33] R. Li, J. Lee, W. Woo, and T. Starner. Kissglass: Greeting gesture recognition using smart glasses. In *Proceedings of the 11th Augmented Human International Conference*, AH '20. Association for Computing Machinery, New York, NY, USA, 2020.

[34] M. H. Loorak, W. Zhou, H. Trinh, J. Zhao, and W. Li. Hand-over-face input sensing for interaction with smartphones through the built-in camera. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19, pp. 32:1–32:12. ACM, New York, NY, USA, 2019. doi: 10.1145/3338286.3340143

[35] I. Lopatovska and H. Williams. Personification of the amazon alexa: Bff or a mindless companion. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pp. 265–268, 2018.

[36] H. Manabe, M. Fukumoto, and T. Yagi. Conductive rubber electrodes for earphone-based eye gesture input interface. In *Proceedings of the 2013 International Symposium on Wearable Computers*, ISWC '13, pp. 33–40. ACM, New York, NY, USA, 2013. doi: 10.1145/2493988.2494329

[37] B. Martinez and M. F. Valstar. *Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition*, pp. 63–100. Springer International Publishing, Cham, 2016. doi: 10.1007/978-3-319-25958-1_4

[38] K. Masai, K. Kunze, and M. Sugimoto. Eye-based interaction using embedded optical sensors on an eyewear device for facial expression recognition. In *Proceedings of the 11th Augmented Human International Conference*, AH '20. Association for Computing Machinery, New York, NY, USA, 2020.

[39] K. Masai, K. Kunze, M. sugimoto, and M. Billinghurst. Empathy glasses. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, p. 1257–1263. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2851581.2892370

[40] K. Masai, K. Kunze, Y. Sugiura, M. Ogata, M. Inami, and M. Sugimoto. Evaluation of facial expression recognition by a smart eyewear for facial direction changes, repeatability, and positional drift. *ACM Trans. Interact. Intell. Syst.*, 7(4):15:1–15:23, Dec. 2017. doi: 10.1145/3012941

[41] K. Masai, Y. Sugiura, M. Ogata, K. Kunze, M. Inami, and M. Sugimoto. Facial expression recognition in daily life by embedded photo reflective sensors on smart eyewear. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, IUI '16, p. 317–326. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2856767.2856770

[42] K. Masai, Y. Sugiura, and M. Sugimoto. Facerubbing: Input technique by rubbing face using optical sensors on smart eyewear for facial expression recognition. In *Proceedings of the 9th Augmented Human International Conference*, AH '18. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3174910.3174924

[43] D. J. C. Matthies, B. A. Strecker, and B. Urban. Earfieldsensing: A novel in-ear electric field sensing to enrich wearable gesture input through facial expressions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pp. 1911–1922. ACM, New York, NY, USA, 2017. doi: 10.1145/3025453.3025692

[44] M. R. Morris, J. O. Wobbrock, and A. D. Wilson. Understanding users' preferences for surface gestures. In *Proceedings of Graphics Interface 2010*, GI '10, pp. 261–268. Canadian Information Processing Society, Toronto, Ont., Canada, Canada, 2010.

[45] T. Nakao, Y. S. Pai, M. Isogai, H. Kimata, and K. Kunze. Make-a-face: a hands-free, non-intrusive device for tongue/mouth/cheek input using emg. In *ACM SIGGRAPH 2018 Posters*, pp. 1–2. 2018.

[46] B. Parkinson. Do facial movements express emotions or communicate motives? *Personality and Social Psychology Review*, 9(4):278–311, 2005. PMID: 16223353. doi: 10.1207/s15327957pspr0904_1

[47] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn. User-defined gestures for augmented reality. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pp. 955–960. ACM, New York, NY, USA, 2013. doi: 10.1145/2468356.2468527

[48] V. Radu, C. Tong, S. Bhattacharya, N. D. Lane, C. Mascolo, M. K. Marina, and F. Kawsar. Multimodal deep learning for activity and context recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4):157:1–157:27, Jan. 2018. doi: 10.1145/3161174

[49] V. Rantanen, P.-H. Niemenlehto, J. Verho, and J. Lekkala. Capacitive facial movement detection for human–computer interaction to click by frowning and lifting eyebrows. *Medical & Biological Engineering & Computing*, 48(1):39–47, Jan 2010. doi: 10.1007/s11517-009-0565-6

[50] J. Rico and S. Brewster. Usable gestures for mobile interfaces: Evaluating social acceptability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, p. 887–896. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1753326.1753458

[51] S. Rostaminia, A. Lamson, S. Maji, T. Rahman, and D. Ganesan. W!nce: Unobtrusive sensing of upper facial action units with eog-based eyewear. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(1):23:1–23:26, Mar. 2019. doi: 10.1145/3314410

[52] J. Ruiz, Y. Li, and E. Lank. User-defined motion gestures for mobile interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pp. 197–206. ACM, New York, NY, USA, 2011. doi: 10.1145/1978942.1978971

[53] C. Saito, K. Masai, and M. Sugimoto. Classification of spontaneous and posed smiles by photo-reflective sensors embedded with smart eyewear. In *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '20, p. 45–52. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3374920.3374936

[54] J. Scheirer, R. Fernandez, and R. W. Picard. Expression glasses: A wearable device for facial expression recognition. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '99, p. 262–263. Association for Computing Machinery, New York, NY, USA, 1999. doi: 10.1145/632716.632878

[55] M. Serrano, B. M. Ens, and P. P. Irani. Exploring the use of hand-to-face input for interacting with head-worn displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pp. 3181–3190. ACM, New York, NY, USA, 2014. doi: 10.1145/2556288.2556984

[56] K. Suzuki, F. Nakamura, J. Otsuka, K. Masai, Y. Itoh, Y. Sugiura, and M. Sugimoto. Recognition and mapping of facial expressions to avatar by embedded photo reflective sensors in head mounted display. In *2017 IEEE Virtual Reality (VR)*, pp. 177–185, March 2017. doi: 10.1109/VR.2017.7892245

[57] M. Tonsen, J. Steil, Y. Sugano, and A. Bulling. Invisibleeye: Mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):106:1–106:21, Sept. 2017. doi: 10.1145/3130971

[58] R.-D. Vatavu. User-defined gestures for free-hand tv control. In *Proceedings of the 10th European Conference on Interactive TV and Video*, EuroITV '12, pp. 45–48. ACM, New York, NY, USA, 2012. doi: 10.1145/2325616.2325626

[59] K. Vega, M. Cunha, and H. Fuks. Hairware: The conscious use of unconscious auto-contact behaviors. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pp. 78–86. ACM, New York, NY, USA, 2015. doi: 10.1145/2678025.2701404

[60] O. Špakov and P. Majaranta. Enhanced gaze interaction using simple head gestures. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pp. 705–710. ACM, New York, NY, USA, 2012. doi: 10.1145/2370216.2370369

[61] M. Weigel, V. Mehta, and J. 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, pp. 179–188. ACM, New York, NY, USA, 2014. doi: 10.1145/2556288.2557239

[62] J. Weppner, A. Poxrucker, P. Lukowicz, S. Ishimaru, K. Kunze, and

K. Kise. Shiny: an activity logging platform for google glass. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pp. 283–286, 2014.

[63] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1083–1092. ACM, New York, NY, USA, 2009. doi: 10.1145/1518701.1518866

[64] K. Yamashita, T. Kikuchi, K. Masai, M. Sugimoto, B. H. Thomas, and Y. Sugiura. Cheekinput: Turning your cheek into an input surface by embedded optical sensors on a head-mounted display. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, VRST '17, pp. 19:1–19:8. ACM, New York, NY, USA, 2017. doi: 10.1145/3139131.3139146

[65] Y. Yan, C. Yu, X. Yi, and Y. Shi. Headgesture: Hands-free input approach leveraging head movements for hmd devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(4):198:1–198:23, Dec. 2018. doi: 10.1145/3287076

[66] J. B. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pp. 3995–4001. AAAI Press, 2015.