Hand-Over-Face Input Sensing for Interaction with Smartphones through the Built-in Camera

Mona Hosseinkhani Loorak
Huawei Noah's Ark Lab
mona.loorak@huawei.com

Wei Zhou
Huawei Noah's Ark Lab
wei.zhou1@huawei.com

Ha Trinh
Huawei Noah's Ark Lab
ha.trinh@huawei.com

Jian Zhao
University of Waterloo
jianzhao@uwaterloo.ca

Wei Li
Huawei Noah's Ark Lab
wei.li.crc@huawei.com

ABSTRACT

This paper proposes using face as a touch surface and employing hand-over-face (HOF) gestures as a novel input modality for interaction with smartphones, especially when touch input is limited. We contribute InterFace, a general system framework that enables the HOF input modality using advanced computer vision techniques. As an examplar of the usage of this framework, we demonstrate the feasibility and usefulness of HOF with an Android application for improving single-user and group selfie-taking experience through providing appearance customization in real-time. In a within-subjects study comparing HOF against touch input for single-user interaction, we found that HOF input led to significant improvements in accuracy and perceived workload, and was preferred by the participants. Qualitative results of an observational study also demonstrated the potential of HOF input modality to improve the user experience in multi-user interactions. Based on the lessons learned from our studies, we propose a set of potential applications of HOF to support smartphone interaction. We envision that the affordances provided by the this modality can expand the mobile interaction vocabulary and facilitate scenarios where touch input is limited or even not possible.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction(HCI)*; Interaction design; • **Computing methodologies** → *machine learning*;

MobileHCI '19, Oct. 01–04, 2019, Taipei, Taiwan

ACM ISBN 978-1-4503-9999-9/19/06…$15.00
DOI: 10.1145/3338286.3340143

KEYWORDS

Computer vision; input modality; smartphones

1 INTRODUCTION

With the success of mobile devices, touch-based interaction has become the dominant method of interacting with computing systems. Touch input, including tapping and flicking, is currently the leading interaction mechanism. However, there are many situations where touch is limited, for instance, when outside is too cold to remove the gloves for touch interactions, when driving a car where touch input is not recommended, and when the device is in a certain distance from the user. In these scenarios, users could benefit from alternative interaction mechanisms not involving touch input, although they might not be used all the time.

In addition to some input modalities such as mobile sensors [15], mid-air gestures [6, 41], and natural languages [28], *face-based* input, by utilizing the phone's front-camera, has been demonstrated as an effective means for interacting with smartphones. Face-based input has been used for auto-screen rotation [5], authentication [8], mobile interaction [45], and camera control [7]. One promising approach is to utilize face as a touch surface for mobile interaction, i.e., *hand-over-face* (HOF) gestures. This input modality has three main benefits: 1) the face offers a larger space for interaction without occluding the smartphone display, 2) the face is often touched and always available for interaction [32], and 3) the unique layout of the face allows for more semantic and intuitive interaction (e.g., face AR, and virtual makeup).

Recently, HOF gestures have been used for interaction (e.g., panning and zooming) with head-worn displays [39, 44], but not with smartphones. Thus, in this paper, we explore the use of the HOF input modality for interaction with smartphones in situations where touch input is limited. We introduce *InterFace* (§3), a general system framework, that employs computer vision techniques on the front-facing camera frames to identify HOF gestures. As a start, we focus on addressing the detection of one single hand gesture (i.e., index finger pointing) and triggering different functionalities based on the hand gesture location to the facial elements.

However, the technique is not limited to a specific hand gesture; and the framework could be extended to include more diverse hand gestures.

To demonstrate the feasibility and effectiveness of InterFace, we integrated it in the design of an Android selfie-taking application with the goal of improving both single-user and group selfie-taking experiences (§4). The application enables one or more individuals to select and apply Augmented Reality (AR) lenses (i.e., visual add-ons) onto their photos by pointing to different areas on their face while taking selfies. We chose this specific use case because it has received global attention within recent years [23]. Furthermore, photos often need to be taken when the smartphone is in a distance from the user (e.g., an extension rod is usually used), so that using touch input is challenging. However, we believe that the HOF input modality for smartphone interaction can be easily applied in other touch-challenging scenarios mentioned above.

In a within-subjects study with 18 participants, we compared the HOF input modality against the conventional touch input in the single-user selfie-taking scenario. Results of the study showed that the HOF input led to significant improvements in both accuracy and perceived workload. Participants also indicated a strong overall preference towards the HOF input, although there were concerns about the social acceptability of this modality. In an observational study focusing on multi-user photo taking scenarios, participants also reported positive experiences of the HOF input and commended its potential to support group-based interactions.

In summary, our main contributions are: 1) development of a general backend framework to enable the novel HOF input modality based on computer vision techniques, as a first exploration of HOF for smartphone interaction; 2) implementation of a selfie-taking application, enabled by our proposed system framework, as an examplar of using HOF input modality in practice, and 3) results from two user studies that demonstrate the potential of the HOF input modality for improving the user experience in both single- and multi-user interactions.

2 RELATED WORK

This research is inspired by recent developments of interaction methods for smartphones, in particular work on mid-air gestures, face-engaged interactions, and on-body input modalities.

Mid-air gestures: Gestural interactions are a natural way of human communication shown to be effective in diminishing some barriers between users and computer interfaces [11]. One of the promising types of gestural interactions are mid-air gestures, using one or two hands. The detection of mid-air input either requires additional sensor-based equipment (for instance on the shoes [3], or on the wrist [10, 24]) or needs a camera to detect the mid-air gestures performed by the hands [6, 34, 41]. In recent years, mid-air gestures have started appearing in some commercialized products, such as Gesture Control developed by BMW to control some car functionalities [1]. Despite the usefulness of mid-air hand gestures, they suffer from the lack of haptic feedback. Additionally, they do not have a spatial surface as a reference point, which results in dedicating each single gesture to only one interaction and thus, reducing the number possible interactions with the device. On the other hand, a single gesture such as tap can be assigned to multiple commands using HOF interactions depending on the face area that the user performs the interaction with.

Face-engaged interactions: Recognizing gestures based on computer vision techniques has shown to be an effective method in human computer interaction [34]. User's face can be easily detected using the smartphone's front-facing camera, while the user is holding the phone. This makes face an effective channel for interaction with the device. There have been some attempts at employing face tracking to provide an extra affordance for smartphone interfaces. User's gaze information can be used for natural scrolling [26, 33]. Eye movement and blinking can be used for mobile browsing and text entry [38]. Hansen et al. [13] describes the "mixed interaction space" between the user and the device and proposes using the face to perform image navigation similar to Image Zoom Viewer [9]. Along the same lines, a face tracking technique is applied in panorama viewing [20] and screen rotation with mobile phones [5]. Recently, facial gestures have emerged in some off-the-shelf systems, such as smiling in Huawei's smartphone camera for shutter release. Taking advantage of face-engaged input channel alone provides a number of new interaction possibilities. However, they still need effort from the user to make facial gestures (e.g., blinking, eye, and head movement), and some unintentional interactions may occur (e.g., blinking).

On-body interactions: A large body of work has been invested on using human body as an interaction surface. It has been explored for many different parts of the body such as palm [12, 42], nail [22], fingers [18], arms [27], and back of the hand [40]. Only a few studies have investigated the use of HOF as an input modality for interaction with head-worn displays [39, 44]. However, to the best of our knowledge, HOF modality as an input channel for interaction with smartphones has not been explored in the literature. This new input channel could enable a more ubiquitous and natural usage for the next generation of mobile platforms.

3 INTERFACE FRAMEWORK

We introduce InterFace framework and its architecture as a potential system capable of interpreting HOF input modality.

InterFace allows one or more people, present in the camera frame, to interact with the smartphone, while using their face as a touch surface.

InterFace Architecture Overivew

Figure 1 shows the architecture of our proposed InterFace framework. It consists of two major computer vision based components including: 1) *Face Landmark Detection & Localization*, and 2) *Hand Gesture Detection & Localization*.

Face Landmark Detection & Localization (FLDL): This component is responsible for detecting and localizing the position of facial landmarks on each of the frames received from the camera. The located landmarks will then be used in combination with the Hand Gesture Detection & Localization component to calculate the touched target area on the face.

Hand Gesture Detection & Localization (HGDL): Our proposed HOF input modality treats face as a touch surface. Thus, depending on the target application, various static or dynamic hand gesture types need to be detected and classified (e.g., index finger pointing, pinch to zoom, and facepalm). Moreover, the task to be performed on the smartphone depends not only on the gesture type but also on the location of the hand relative to the face elements. For instance, pointing index finger to lips could mean turning the microphone on during a phone conversation, while the same gesture of pointing the index finger on the nose could mean turning the microphone off. Identifying the hand gesture type and localizing the gesture in the frame is the responsibility of the HGDL component.

In InterFace system, each frame received from the smartphone's built-in camera (*Camera* component) goes through the two models embedded in the *FLDL*, and *HGDL* components to detect the coordinates of the facial landmarks, the gesture type, and the coordinates of the gesture in the frame. The outputs of these two components enter the *Interaction Control* unit. Using the gesture coordinates (output of HGDL) in relation to the facial landmarks coordinates (output of FLDL), the controller identifies the face area that the interaction has been performed on. This calculation is based on the greatest proximity (e.g., minimum distance) of facial landmarks with the gesture location. Utilizing this result and the gesture type (output of HGDL), the Interaction Control unit identifies the user interaction.

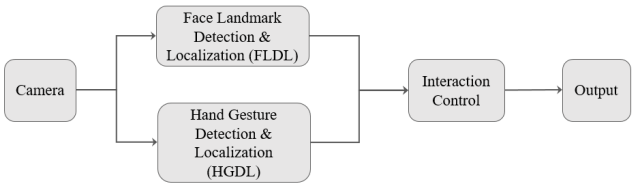

Figure 1: The InterFace system framework architecture.

The result will be then sent to the *Output* component, where depending on the target application an appropriate action will be performed and feedback will be provided to the user. For instance, if the user employs HOF gestures to change the music volume of a song while driving, the performed HOF interaction will be reflected in the audio level of the smartphone.

InterFace Implementation Details

Based on the overall architecture of InterFace described above, we implemented a potential prototype of the system in the form of an Android application to demonstrate its feasibility. In our implementation, we employed the Mobile Vision API [2] for FLDL component to detect 8 main facial landmarks (see Figure 3-b). For the HGDL component, in this research, we only considered one static and common hand gesture (i.e., pointing with index finger) to be detected on the face. However, depending on the use case, the HGDL component could be extended to include the detection and classification of other gestures. Focusing on this single gesture, the HGDL component further needs to localize the fingertip of the index finger within each frame to locate the gesture. However, there exist no publicly available model capable of detecting the index fingertip especially on the face. Thus, we collected a dataset of images with various backgrounds and lightings and annotated the index fingertip location within each image. We then trained an object detection model to localize the index fingertip within each frame. In the following, we describe in detail our data collection, model training process, and the implementation details of our Android application for detecting small index fingertips within the frames.

Data Collection

In our implementation, the goal of HGDL component is to localize the absolute location of the users' index fingertip within each of the frames received from the smartphone front-facing camera using computer vision techniques. Detecting fingertip from captured RGB frames of a smartphone's camera is challenging due to 1) background complexity, 2) hand orientation, 3) lighting varieties, and 4) image blurriness due to the camera movements. Furthermore, detecting fingertip over the face is even more challenging due to color similarity of hand with face skin. These challenges make traditional skin color-based fingertip detection techniques (e.g., [21, 36]) not suitable for recognizing HOF interactions with smartphones.

Deep learning (DL) has proved to be very effective in addressing computer vision problems. Specifically, DL-based object detection and localization (e.g., YOLO [37], and SSD [29]) appears to be a promising direction for locating the users' fingertip within each of the received frames from the camera. For training an object detection model

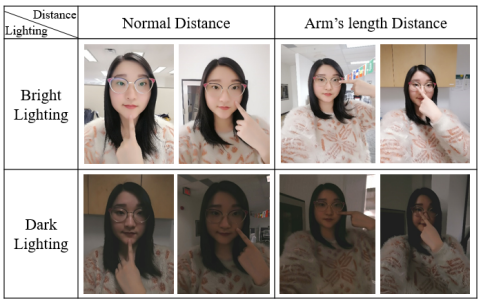

Figure 2: Examples of HOF collected image data under 2 sample lighting conditions, with 2 pre-defined distances, with varying backgrounds, and hand posture orientations.

capable of detecting fingertips, publicly available datasets do not contain appropriate HOF image data. For instance, Cam3D [31] and Hand2Face [35] datasets focus on employing hand over face occlusions to better recognize the human facial emotions. Thus, their hand gestures are limited to those relevant to facial emotions and very few instances of our target gesture (i.e., index finger pointing) can be found in these datasets. Thus, we collected a dataset of HOF images consisting of 8500 annotated images (the process of our data collection is described below). We combined our dataset with 22000 images collected by Huang et al. [19] for detecting finger key points from an egocentric vision with a mobile camera. Thus, in total, our collected dataset includes 30500 images, each annotated with the fingertip locations.

For collecting our HOF image dataset, we asked eight volunteers (four females and four males) to record 6—10 short videos-clips (20-40 seconds) from themselves individually with the front-facing camera of their smartphone, while using their index finger to point and touch on different areas of their face as well as the environment around them. We asked our volunteers to record videos with various indoor and outdoor backgrounds, with varying lighting conditions, distances (i.e., normal distance, and at arm's length distance), and different hand posture orientations to make the resulting trained model support fingertip detection in diverse situations. Figure 2 shows sample collected images considering the aforementioned criterias. Our collected images consist of the extracted frames from the recorded videos. For automatically annotating the location of index fingertip inside each extracted frame, we took advantage of Visual Object Tracking (VOT) which is the process of locating a moving object over time in a video [25]. Among various VOT techniques, we specifically chose CSRDCF technique [30] due to its high accuracy [25] and its integration into the OpenCV library.

Index Fingertip Detection Model Training

Using our collected dataset, we trained an object detection model for index fingertip detection and localization. For training, we used the single shot multibox detector (SSD) [29] with MobileNet [17] as the backbone feature extractor. SSD and MobileNet were chosen as they are considered as efficient network architectures (low computational burden) and implementations for the applications of mobile vision. For the training, we initialized the weights with a truncated normal distribution with a standard deviation of 0.03. The initial learning rate is 0.004, with a learning rate decay of 0.95 every 5,000 iterations. The input image size is 300 * 300 pixels as well. For our experiment, we randomly divided the dataset in a ratio of 8:2. The former part is used as the training data and the latter as evaluation data. Thus, 24,400 images were included in the training set and the evaluation set included 6,100 images. Another alternative for dividing the dataset is using Leave One Subject Out Cross Validation (LOSOCV). However, we do not expect a significant difference between the two approaches due to similarities in features of humans' fingertips. The performance of InterFace in our user studies (§5) confirms this as none of the participants in data collectioin phase were among the study participants and the system still performed with similar accuracy.

We trained the model on CentOS 7 (1708) OS with two NVIDIA Tesla P100 GPUs. In the evaluation, we set the Intersection Over Union (IOU) threshold to 0.6 and achieved 93% mAP (mean Average Precision) for fingertip detection on our test images. For the development of the Android application, a Sony Xperia XZ2 smartphone with an MSM8998 Snapdragon 845 CPU and 4GB RAM was used. The results on our testing smartphone revealed that the application can detect and localize the users' index fingertips 8 frames per second.

InterFace Implementation to Detect Small Fingertips

As we described above, for detecting index fingertip, we took advantage of computer vision based object detection models. However, localizing very small objects is a known challenge in object detection models. Positioning the smartphone in distant from the user's face for performing interactions results in obtaining frames with very small index fingertips, which might result in poor detection and localization. To alleviate this, in our developed Android application of InterFace, we increased the feature resolution of fingertips by magnifying each input frame with respect to the face boundaries of user(s) present in the frame.

4 USING INTERFACE FOR SELFIE-TAKING

To investigate the feasibility and usefulness of InterFace, we focus on investigating its application for improving the experience of selfie photo taking as an exemplar use case. we integrated our framework into an exemplar selfie-taking application. We chose selfie-taking as our representative use case as it is a commonly used application in which touch screen input is challenging, due to the frequent need of

holding the phone from an extended distance to capture mid- or wide-shot selfie photos.

Exploratory Study: Selfie-taking in Practice

To inform the design of our HOF-based selfie-taking application, we conducted a series of semi-structured interviews exploring current challenges in selfie-taking and common tasks supported in current selfie-taking applications. We recruited four students and professionals (3 females, 1 male, aged 17—34) who had experiences using existing selfie-taking applications. Each interview session lasted 45—60 minutes and were audiotaped.

Study findings: For taking selfies, two main tasks were highlighted by our participants: 1) adjusting camera settings (e.g., shutter release, and zooming), and 2) customizing facial appearance. In this paper, we focused on the second task which includes augmenting the face with AR lenses (e.g., virtual makeup or accessories), *"we always add stickers. Makes our photos look fun. [...] but it's hard to select the right sticker"*. Participants also mentioned that they prefer customizing their appearance before the photo is taken as they may not be the camera owner, *"sometimes its my friend's camera and I can't edit my face later"*.

Many face customization applications (e.g., Instagram, Faceu, and Snapchat) allow users to select AR lenses using touch input modality. However, using touch to navigate through the large collection of AR lenses and apply them to the face while holding the phone from a distance is a challenging task. HOF gestures could be a more natural and effective alternative for this task. Thus, we focused on addressing this challenge in our selfie-taking application.

Furthermore, our results show that selfie-taking can be categorized into two classes: 1) *single-user*, and 2) *group selfies*. Adjusting the facial appearance is not only a challenge in single-user scenarios but is also a big issue in group selfies as it requires simultaneous customization, *"most apps are designed for a single user. It's difficult to make everyone happy about the stickers,"* and *"the person whose photo is taken can't control anything, may not in a nice facial expression."* Some people in the group may prefer their face without any AR lenses, while others might prefer adding some, *"maybe we can choose for each person to add or not add the emojis."* Our HOF input modality could potentially provide effective support for multi-user facial customization, thanks to the ability to detect and track multiple fingertips and faces.

Facial Customization Tasks

Based on our study findings and literature review, we articulated the tasks of enhancing face with AR lenses into three main categories:

T1. Selection of the target AR lenses: In most selfie-taking applications, there is a relatively large database of AR lense that users can choose from. However, selecting the right one is challenging based on their small previews, *"looking through many small stickers isn't fun"*.

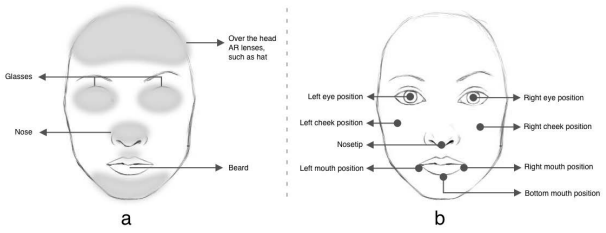

Figure 3: a) Facial areas where InterFace supports interaction and augmenting the face with AR lenses, b) 8 facial landmarks detected on each face.

T2. Modification of the selected AR lenses: The users needed ways to adjust the selected lens to their own face, *"some people's faces are wide. Some are narrow. Stickers are of the same size and can't be much customized"*. Examples of such adjustments include resize, and rotation.

T3. Creation of new AR lenses: Within currently available solutions for augmenting the facial features, there is not much room for users to be creative and generate their own AR lenses. Thus, HOF input modality can be effectively and intuitively employed for creating such ARs on the face.

HOF is a natural and intuitive input modality for performing all the aforementioned tasks. However, *Selection* is the only task that almost all AR-based selfie-taking applications allow users to perform and has been widely employed by users. Thus, as an initial step, we focus on addressing the selection task (**T1**) by integrating InterFace into an Android selfie-taking application that allows users to select and apply AR lenses on their face in real-time by pointing with index finger to their different facial areas.

Adapting InterFace for Selfie-taking Scenario

We extended the InterFace Android application to enable the selection of AR lenses on different areas of the face using HOF input modality. Our application allows one or more individuals, present in the camera frame, to choose and apply facial AR lenses using the simple and static gesture of touching their target facial area using their index fingertip (see Figure 4). Figure 3-a shows the target areas of the face that touching them results in augmenting the face with lenses using our application. For each of these facial areas, we embedded four alternative lenses to be selected from. However, this database of lenses could easily be extended.

As an example, when a facial area such as lip is touched for adding a beard, the available beard AR lenses appear on his chin sequentially every ten frames apart. When the user wants to make the selection, he simply removes his index fingertip from his chin area and the last shown beard AR lens will stay on his face. Covering the face with the hand

Figure 4: a) The man has first touched his forehead to add a hat and then touches his chin for a beard. The lady touches her eye to add glasses. b) The user selected elements corresponding to the touched areas (hat and beard for the man and glasses for the lady) are added on the users' faces.

results in removing all the added lenses. As each user's index fingertip and HOF gestures is tracked individually, multiple users can perform the above interactions at the same time without interfering each other in group selfie-taking.

5 EVALUATION OF HOF INPUT MODALITY

To evaluate the user performance and experience of the HOF input modality, we conducted two user studies. In the first study, we evaluated HOF against touch input in a single-user interaction scenario. We chose touch-based interaction as our baseline condition as touch has been considered the most common input modality used in smartphones to date. In the second study, we assessed the user experience of HOF in a multi-user interaction setting by analyzing qualitative feedback from pairs of participants who used the HOF input modality for group selfie-taking.

Single-user Study: HOF vs. Touch Input

We compared our HOF-based selfie-taking application to a functionally equivalent touch-based Android application. As in the HOF application, the touch-based application uses the Face Landmark Detection module to detect four target areas of the face, each of which has four alternative AR lenses to be selected from. To select an AR lens for a specific area, the user presses a finger at the target area on the touchscreen and holds the finger at that position on the screen to rotate through the alternative lenses, which appear every 10 frames apart. Once the user has found the desired AR lens, she removes the finger from the screen to make the selection. To remove all the added lenses, the user simply moves the phone away from the face. Both applications were run on a Sony Xperia phone with a 5-inch touchscreen display.

Participants: We recruited 18 participants (6 female, 12 male, ages 18-36, mean 24.6) with occupational backgrounds in science, technology, accounting, sales and administration. All participants were regular touchscreen-based smartphone users, and 55.6% of them used selfie-taking smartphone apps on a regular basis. 44% of participants indicated that they never used hand gestures to interact with smartphones in the past, while 55.6% of participants stated that they had tried using gestures on smartphones a few times. Participants were compensated for their participation.

Study design: The study was a 2x2 within-subjects, single-session factorial design with two factors: *Input Method* and *Distance*. The levels of *Input Method* were (HOF, Touch), and the levels for *Distance* were (Close, Far). For the "close" distance, we asked participants to hold the phone at a normal distance from which they normally interact with their phone, in order to take a close-up shot of their face. For the "far" distance, we asked participants to stretch their arms as far as they could and hold the phone at that distance while selecting AR lenses. We hypothesized that the HOF input method would lead to better user performance and experience, especially when users interact with their smartphone from a far distance. The ordering of the presented input methods was counterbalanced across participants.

Procedure: Following a sociodemographic questionnaire at the beginning of the session, we introduced participants to the task of applying leses to different areas of their face while taking selfies using two different input methods. For each application, participants progressed through 3 sets of tasks:

1. Guessability tasks: Before introducing our application, we first collected participants' feedback on what they considered as suitable methods for AR lens selection. We gave participants 4 tasks of applying a lens to their nose, eyes, lip, and forehead. For each task, we asked participants to suggest their own ways to perform the task, using the given input method. For the touch input, we asked them to suggest any methods that use the touchscreen. For the HOF input, we asked them to suggest and perform any methods that use hand gestures on the face. We encouraged the participants to suggest as many suitable methods as they would like.

2. Practice tasks: Following the guessability tasks, we introduced the tested application to the participants. We then asked them to perform 4 AR lens selection practice trials to familiarize themselves with the application.

3. Study tasks: Upon the completion of the practice trials, we asked participants to perform 10 close-distance trials and 10 far-distance trials, for a total of 20 trials per application. Each 10-trial set contains 4 trials that require selecting one AR lens per trial, and 6 trials that require selecting 2 lenses per trial. Participants were instructed to perform the trials as quickly and accurately as possible. The trials were shown on a 13-inch laptop display located in front of the participants. The ordering of the 20 trials were randomized across participants.

After each application, we asked participants to complete two questionnaires: (1) the NASA-TLX questionnaire measuring perceived task load [14]; (2) the Absolute Rating of Input Method questionnaire assessing the usability and social

acceptability of the tested input method. Upon completing both applications, participants filled out the *Relative Rating of Input Methods* questionnaire (Figure 5) to compare the two input methods. We concluded the session with a semi-structured interview, prompting for comparisons of the two input modalities as well as suggestions for improvement and further use cases. The study session took approximately 60 minutes and was videotaped. The phone screen was also video-captured for analysis of user performance. The applications were also instrumented to log all timestamped interactions with the smartphone, including detected landmarks and displayed AR lenses.

Results

We evaluated the user performance and experience of the two input methods using the following measures.

Guessability score: This metric was measured as the percentage of participants that suggested index finger-pointing or touching a specific facial area as one of suitable selection methods, as identified in the guessability tasks. This measure is an indicator of the intuitiveness of the input methods supported in our applications.

Table 1: Guessability scores of the two input methods

Task	Touch (%)	HOF (%)
Apply an AR lens to your nose	72.22	88.89
Apply an AR lens to your eyes	72.22	72.22
Apply an AR lens to your lips	72.22	72.22
Apply an AR lens to your forehead	72.22	66.67

Table 1 presents the guessability scores of the two input methods for each facial area. Overall, both finger-pointing and touching at a specific facial area were suggested by the majority (>50%) of the participants, demonstrating the intuitiveness of these methods. Besides finger-pointing, participants also suggested a variety of area-specific gestures (e.g., making circle gestures around the eye areas). However, index finger pointing remained the most commonly suggested gesture that can be uniformly applied for all facial areas.

User performance: We evaluated the user performance of the two input methods using two measures:

Completion Time: measured as the time taken (in milliseconds) to complete a trial, starting from the first touch/finger pointing until the participant successfully selects the target AR lenses.

Error Rate: measured as the number of errors made by the participant in each trial, manually counted from the video recordings of the phone screen during the trials. An error is defined as an instance at which the app incorrectly detects the facial area that the participant intends to touch on or point to (e.g., the participant points to the nose area but a lip lens is displayed). We excluded performance data of one participant due to technical difficulties. From the remaining participants, we collected 20 *trials* $*$ 2 *input methods* $*$ 17 *participants*, for a total of 680 trials. Time and error rate data were first aggregated by participant and the two factors being investigated. Results of Shapiro-Wilk tests showed that time and error rate data were not normally distributed (W>.69, p<.001). Thus, we performed the non-parametric Aligned Rank Transform procedure [43], which enables the use of ANOVA after aligning and ranking data. We then analyzed the transformed data using a generalized linear mixed-effects model of variance with *Input Method* and *Distance* as fixed effects and Participant as a random effect.

Table 2: Completion time and error rate for close and far distances of the two input methods (Mean (SD))

Distance	Touch		HOF	
	Time (ms)	Error rate	Time (ms)	Error rate
Close	9834.27 (432.06)	.17 (.05)	10855.57 (684.33)	.02 (.01)
Far	13033.81 (1111.95)	.47 (.09)	12209.81 (892.71)	.07 (.03)

Table 2 presents user performance results of the two input methods for close and far distances. The average completion time across all trials for the HOF input was 11532.69 ms (SD=566.23). For the touch input, it was 11434.04 ms (SD=650.04). There was a significant effect of Distance on completion time ($F_{1,64} = 13.76, p < .001$). However, there were no significant effect of *Input Method* and no significant interaction effect of *Input Method* $*$ *Distance*.

The average number of errors per trial for the HOF input was .05 (SD=.02). For the touch input, it was .32 (SD=.06). There was a significant interaction effect of *Input Method* $*$ *Distance* on error rate ($F_{1,64} = 15.74, p < .001$). To further examine the effect of each factor, we performed post-hoc pairwise comparisons using Wilcoxon signed-rank tests on the original, non-normal error rate data. Results showed that the HOF input method led to significantly lower error rates compared to the touch input for both close distance ($Z = -2.83, p = .005$) and far distance ($Z = -3.42, p = .001$). For the touch input, the error rate at the close distance was significantly lower than that at the far distance ($Z = -3.00, p = .003$). In contrast, for the HOF input, there was no significant difference in error rate between the close and far distances ($Z = -1.51, p = .13$). Overall, these results demonstrate the superior accuracy of the HOF input compared to the touch input, and the consistency in the performance of the HOF input across different distances.

Perceived Workload: Based on the results of NASA-TLX questionnaire, the average *overall workload score* for the HOF input was 42.61 (SD=3.09). For the touch input, it was 52.84 (SD=5.36). Results of ANOVA tests showed significant effects of the *input method* on the *overall workload score* ($F_{1,16} = 4.58, p = .048$), in favor of the HOF input.

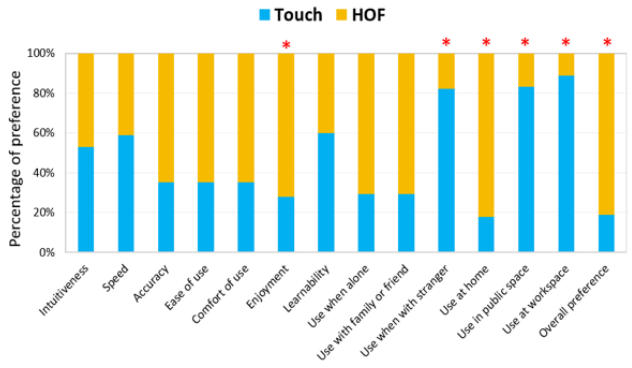

Figure 5: Relative rating of the two input methods (* indicates statistical significances).

Absolute Rating of Input Method: We asked participants to rate each input method using a 13-item, 5-point scale questionnaire. The questionnaire consists of two subscales:

Usability: This composite subscale consists of 7 usability measures, including *intuitiveness, speed, accuracy, ease of use, comfort of use, enjoyment*, and *learnability*. The composite usability rating (Cronbach's $\alpha = .84$) of the touch input was 3.47 (SD=0.70). For the HOF input, it was 3.86 (SD=0.72). Results of an ANOVA test showed that there was no significant effect of *Input Method* on the composite usability rating ($F_{1,16} = 3.69, p = .07$). Results of Wilcoxon signed-rank tests on individual scale measures showed that the HOF input was rated as significantly more *enjoyable* than the touch input ($Z = -2.52, p = .012$).

Social acceptability: This subscale consists of 6 measures assessing participants' willingness to use the input method in 6 social contexts (*alone, family or friend, stranger, home, public space* and *workspace*), as used by Serrano et al. [39]. The reliability (Cronbach's α) of our social acceptability questionnaire was 0.844. Participants expressed a strong willingness to use the HOF input method when being *alone* (median = 5.0, IQR=4.0-5.0), *at home* or *with family or friends* (median = 4.0, IQR=4.0-5.0). However, they expressed either neutral or low willingness to use this gesture input when being *with a stranger* or at a *public space* (median=3.0, IQR=1.5-4.0), or when being at a *workspace* (median=2.0, IQR=1.0-3.0).

Relative Rating of Input Methods: After participants completed both conditions, we asked them to indicate which of the two input methods performed better on the same 13 criteria included in the Absolute Rating questionnaire, in addition to their overall preference (figure 5).

Results of Chi-square tests showed the HOF input was rated significantly better in terms of *enjoyment* ($\chi^2(1) = 4.77, p = .029$), and *overall preference* ($\chi^2(1) = 4.77, p = .029$), compared to the touch input. Participants were significantly more willing to use the HOF input than the touch input if they were at *home* ($\chi^2(1) = 7.11, p = .008$). However, they expressed a significantly stronger willingness to use the touch input in *public* ($\chi^2(1) = 7.11, p = .008$), at *workspace* ($\chi^2(1) = 9.94, p = .002$), or when being *with a stranger* ($\chi^2(1) = 7.11, p = .008$).

In summary, quantitative results of the study showed that the HOF input method led to significantly higher accuracy, higher enjoyment level, and were overall preferred by the participants, compared to the touch input.

Qualitative Feedback: We performed thematic analysis on the transcribed interviews and categorized the participants' feedback into two themes.

1. User experience: Participants described their experiences of the HOF input method as "fun", "cool", "intuitive", and "enjoyable". Many of them commended the accuracy and consistency in the performance of the HOF input across different distances, *"There was not much of a difference when holding the phone close vs. far with the gesture. With the touch, if you are holding it further away, it's annoying. Also your face is smaller, so it's easier to get confused between eyes and forehead"*. The HOF method also eliminated the occlusion issues inherent in the touch method, *"I found it more comfortable because I was able to see on the screen what was happening with the gesture... [With the touch] I had to keep removing my finger to check what's behind"*. Although this study focused on single-user interactions, participants also envisioned the potential of HOF input to improve group-based interactions, *"I can imagine taking selfies with my kids. It would be funny to point at my daughter's eyes and change the sticker on her eyes"*. Despite the general positive experiences of the HOF method, four participants still expressed their overall preferences towards the touch input, mainly due to its familiarity and intuitiveness.

Participants offered several suggestions for improving the usability of the HOF method. They suggested to improve the robustness of the fingertip tracking. In addition, instead of time-based scrolling through alternative AR lenses, they proposed the use of swiping, tapping, or hand wave gestures to rotate through those lenses at their own pace.

2. Social acceptance: In consistent with the quantitative results, most participants expressed a strong willingness to use the HOF input method in private settings (e.g., at home, or with family or friends). However, they had concerns for using the gesture input in public settings, and preferred the touch input in those cases, *"if I'm in public places and more exposed, I'd prefer the touch one just because it's more hidden. It makes me fit more and less stand out"*.

Multi-user Observational Study

To gain a further understanding of the potential of the HOF input method to support group-based interactions, we conducted 3 observational study sessions, each of which

included two participants. Our aim was to observe users' interactions with the HOF input in the context of group selfie-taking, collect qualitative feedback on the value of HOF input for group interactions, and explore other use cases of this modality.

Participants: We recruited three pairs of participants (3 females, aged 17—44) from a variety of backgrounds including technology, accounting, and sales who had not participated in the previous studies. Five participants had previous experience using off-the-shelf products to augment their face. One mentioned that she always applies modification on her selfies such as AR lenses, skin smoothing, and adding text.

Procedure: Each observational study session began with a brief training session, in which we introduced our HOF-based selfie-taking application, and walked participants through a sample scenario covering the features of the application. Following the training session, participants filled in a background demographic questionnaire. We then asked them to perform a number of tasks that involved taking selfies with the other participant. We also asked them to think aloud while performing the tasks. Afterwards, the participants of each session took part in a semi-structured interview. Interviews focused on the participants' experience with our application, its benefits, challenges, future extensions, and other use case scenarios of HOF input channel. In each session, one researcher observed the user interactions with the camera and took notes. Each session lasted for about 30—60 minutes and was audio recorded.

Results

We analyzed the collected data and grouped comments by emerging themes.

Overall response: In general participants were very positive about HOF input channel , *"To me, it looks pretty straightforward. [...] I'm always looking at my phone. Why not use it to make changes"*. The usefulness of the HOF interaction was clear for the participants especially for taking group selfies, *"if you have a group, everyone wants to do it at the same time"*, and *"that's the dream [laughing]. My wife loves adding make-ups on our selfie photos but I don't need them, and there is no easy way to get rid of them"*.

HOF gestures for camera controls: Users found HOF interactions useful for controlling the camera settings such as the lighting, zoom level, and camera focus, *"I'd use it for group photos. Maybe not just for filters but if someone wants to make sure that their face is focused, they can point at themselves"*.

Support for devices with varying screen sizes: Participants hypothesized that HOF input channel can go beyond smartphones and can be employed in devices with varying screen sizes with embedded cameras. Smart watch is an example that using HOF interactions can eliminate the occlusion issues resulting from using touch as the input, using only the built-in camera of the device.

User experience: Users suggested future extensions of our prototype based on the user characteristics and context.

1) Prioritizing AR lenses according to user characteristics: Participants reported that they prefer the appearance of the AR lenses to be adjusted based on their own characteristics (e.g., gender, and age). *"It'd be cool if men and women can have different options to choose from"*, and *"if my age was 60, I'd have liked to see elements of 70's or 80's"*.

2) Filtering lenses by context: Participants suggested considering the contexts (e.g., gym, beach, and restaurants) and filter the lenses based on only the related contexts, *"when I'm working out at the gym, I can't easily find workout stuff to add to my photo before sharing with my friends"*.

6 DISCUSSION AND FUTURE WORK

Value of HOF Input Modality

The results of our evaluation studies clearly demonstrated the potential of the HOF input channel to improve the user experience of smartphone interaction in scenarios where touch input is limited. In the single-user selfie-taking study, HOF input led to significant improvements in both accuracy and perceived workload, and was rated as more enjoyable compared to the touch input. As a result, participants expressed strong preferences towards the HOF input, even though there were concerns about the social acceptability of this modality in public settings. Qualitative findings from our two user studies also indicated that the HOF input was not only effective for interacting with smartphones from a distance, but also especially useful for supporting simultaneous interactions in group settings. Furthermore, this new input channel shines not only for interaction with smartphones but also for interacting with other devices of different screen sizes such as smart watches, as suggested by our participants. Similarly, we hypothesize that large interactive displays commonly used for collaboration or art installations can also benefit from HOF input modality for allowing people to interact with them from a distance.

Applications of InterFace Framework

Based on the qualitative results of our studies and review of literature we propose a number of applications where HOF input modality, enabled by our InterFace framework, could potentially improve user experience.

Supporting smartphone interaction while driving: Smartphones are nowadays used for functionalities such as navigation, playing music, and making phone calls while driving. People often mount the device to their car front decks at a certain distance. This significantly reduces the ability to perform touch input. Serrano et al. [39] showed the

usefulness of panning and pinching gestures on the user's cheek and chin for navigation applications with head-worn displays. Thus, HOF gestures can potentially be employed to interact with smartphone applications such as navigation and music playing in the car.

Enabling visually impaired users to interact with smartphones: Bennett et al. [4] investigated how teens with visual impairment access smartphone photography. The results reveal that adding AR lenses on face is a challenge for this group of users, *"I can't tell which lens is coming up or where I need to click. [...] I'll have to get close to the camera, find the lens, and then back"*. Thus, the HOF input modality can potentially improve the experience of this group of users in taking photos.

Photo editing of faces with semantic interaction: People use photo editing softwares such as Adobe Photoshop for face editing. However, this task could be challenging with smartphones due to their limited screen size, and the need to navigate through menus for selecting functionalities. HOF gestures could assist in this scenario by mapping the editor's facial elements to the face of the target person selected for editing in the photo. The editor can then improve the target face by performing HOF gestures on her own face.

Enhancing shopping experience through Virtual Reality: Live virtual try-on of products such as beauty items has grown interest in the shopping domain recently. However, selection and filtering of products using touch and menus might be challenging for users. HOF gestures for purchasing facial related products (e.g., lipsticks, eye glasses) could be a potential way of trying products without the need to use public touch screens, and customers would be able to perform these interactions from a distance.

Video conference meetings: During the video conference meetings, it could be challenging to manage the interaction (e.g. muting/unmuting the microphone) when multiple people are in the same room and are participating using a single device. HOF gestures could potentially be useful to control the video conferencing device from a distance. For example, one of our participants suggested placing index finger on the nose for muting the device, *"instead of looking for a button on the device to mute it or asking someone else to do it, I'd prefer just shushing"*.

Future Directions

Performance improvement: As we described in §3, two DL models are constantly running to detect the face landmarks and fingertips. This might negatively affect the performance of our application. As future work, we plan to create a single model for both detecting the face landmarks and HOF gestures to maximize the performance of InterFace. Also, our study results indicate that the completion time of HOF is slightly higher (but not statistically different) than the touch condition for close distance. We hypothesize that the slow fingertip detection frame rate is the main factor affecting the completion time. In future, we will test this hypothesis with an improved fingertip detection model.

Expanding the gesture vocabulary: Using HOF input channel opens up many new interaction possibilities. In the application of selfie-taking, new gestures could be embedded to address the *Modification* (T2) and *Creation*(T3) tasks identified in §4. Examples include pinching to zoom in/out for resizing the lenses, tapping for skin smoothening, and sketching the shapes of desired lens on the target face area to filter lenses based on both shape and position.

HOF gesture activation: given the relatively high frequency at which people touch their face, Midas touch problem [16] (i.e., unintentionally issuing a command to the device) may occur. One solution is to use a delimiter at the start of the interaction (e.g., voice command, or pressing a button) to explicitly activate the HOF gestures. The frequency of occuring Midas touch problem depends on the target use case. For instance, for the selfie-taking scenario, this may not be a very frequent issue as people often do not unintentionally touch their faces before taking photos. Thus, we left out the examination of this factor in this first exploration.

Evaluating HOF input in new use cases: Depending on the target use case of HOF gestures, the accessibility and social acceptability of touching facial areas need to be further studied. For example, in driving, touching eyes might not be acceptable in comparison to cheeks or chin.

Touch vs. hover: Identifying whether the user has touched her face or has moved her hand over her face without touching is challenging with the current 2D photos taken by smartphones. In future, we aim to explore the use of depth sensors to differentiate touching vs. hovering over the face.

7 CONCLUSION

We presented HOF input modality as a novel input channel for interaction with smartphones, offering extra possibility of interacting with the phone under certain situations, e.g., when the phone is at a distance from the user. We introduced InterFace as a framework for realizing the HOF gestural interactions with smartphones. We demonstrated the usability of InterFace in practice by integrating it into an exemplar selfie-taking application. In a study comparing HOF with touch input for augmenting face with AR lenses, we found that HOF significantly improved accuracy and perceived workload, and was preferred by the participants. Qualitative evaluations of our application revealed the potential of HOF input to enhance user experience in both single and multi-user interactions, its suitability for new use cases, and future extensions of the system.

REFERENCES

[1] [n. d.]. BMW's Gesture Control. https://driving.ca/bmw/7-series/auto-news/news/how-it-works-bmw-gesture-control. Accessed: 2018-09-18.

[2] [n. d.]. Mobile Vision. https://developers.google.com/vision/. Accessed: 2018-09-17.

[3] Gilles Bailly, Jörg Müller, Michael Rohs, Daniel Wigdor, and Sven Kratz. 2012. ShoeSense: A New Perspective on Gestural Interaction and Wearable Applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1239–1248. https://doi.org/10.1145/2207676.2208576

[4] Cynthia L. Bennett, Jane E, Martez E. Mott, Edward Cutrell, and Meredith Ringel Morris. 2018. How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 76, 12 pages. https://doi.org/10.1145/3173574.3173650

[5] Lung-Pan Cheng, Fang-I Hsiao, Yen-Ting Liu, and Mike Y. Chen. 2012. iRotate: Automatic Screen Rotation Based on Face Orientation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 2203–2210. https://doi.org/10.1145/2207676.2208374

[6] Shaowei Chu and Jiro Tanaka. 2011. Hand Gesture for Taking Self Portrait. In *Human-Computer Interaction. Interaction Techniques and Environments*, Julie A. Jacko (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 238–247.

[7] S. Chu, F. Zhang, N. Ji, Z. Jin, and R. Pan. 2017. Pan-and-tilt self-portrait system using gesture interface. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*. 599–605. https://doi.org/10.1109/ICIS.2017.7960063

[8] Alexander De Luca, Alina Hang, Emanuel von Zezschwitz, and Heinrich Hussmann. 2015. I Feel Like I'm Taking Selfies All Day!: Towards Understanding Biometric Authentication on Smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1411–1414. https://doi.org/10.1145/2702123.2702141

[9] Eva Eriksson, Thomas Riisgaard Hansen, and Andreas Lykke-Olesen. 2007. Movement-based Interaction in Camera Spaces: A Conceptual Framework. *Personal Ubiquitous Comput.* 11, 8 (Dec. 2007), 621–632. https://doi.org/10.1007/s00779-006-0134-z

[10] Jun Gong, Zheer Xu, Qifan Guo, Teddy Seyed, Xiang 'Anthony' Chen, Xiaojun Bi, and Xing-Dong Yang. 2018. WristText: One-handed Text Entry on Smartwatch Using Wrist Gestures. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 181, 14 pages. https://doi.org/10.1145/3173574.3173755

[11] Sukeshini A. Grandhi, Gina Joue, and Irene Mittelberg. 2011. Understanding Naturalness and Intuitiveness in Gesture Production: Insights for Touchless Gestural Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 821–824. https://doi.org/10.1145/1978942.1979061

[12] Sean G. Gustafson, Bernhard Rabe, and Patrick M. Baudisch. 2013. Understanding Palm-based Imaginary Interfaces: The Role of Visual and Tactile Cues when Browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 889–898. https://doi.org/10.1145/2470654.2466114

[13] Thomas Riisgaard Hansen, Eva Eriksson, and Andreas Lykke-Olesen. 2006. Use Your Head: Exploring Face Tracking for Mobile Interaction. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*. ACM, New York, NY, USA, 845–850. https://doi.org/10.1145/1125451.1125617

[14] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139 – 183. https://doi.org/10.1016/S0166-4115(08)62386-9

[15] Ken Hinckley and Hyunyoung Song. 2011. Sensor Synaesthesia: Touch in Motion, and Motion in Touch. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 801–810. https://doi.org/10.1145/1978942.1979059

[16] Ken Hinckley and Daniel Wigdor. 2002. Input technologies and techniques. In *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*. Taylor and Francis, Chapter 9.

[17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[18] Da-Yuan Huang, Liwei Chan, Shuo Yang, Fan Wang, Rong-Hao Liang, De-Nian Yang, Yi-Ping Hung, and Bing-Yu Chen. 2016. DigitSpace: Designing Thumb-to-Fingers Touch Interfaces for One-Handed and Eyes-Free Interactions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1526–1537. https://doi.org/10.1145/2858036.2858483

[19] Y. Huang, X. Liu, L. Jin, and X. Zhang. 2015. DeepFinger: A Cascade Convolutional Neuron Network Approach to Finger Key Point Detection in Egocentric Vision with Mobile Camera. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*. 2944–2949. https://doi.org/10.1109/SMC.2015.512

[20] Neel Joshi, Abhishek Kar, and Michael Cohen. 2012. Looking at You: Fused Gyro and Face Tracking for Viewing Large Imagery on Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 2211–2220. https://doi.org/10.1145/2207676.2208375

[21] S. K. Kang, M. Y. Nam, and P. K. Rhee. 2008. Color Based Hand and Finger Detection Technology for User Interaction. In *2008 International Conference on Convergence and Hybrid Information Technology*. 229–236. https://doi.org/10.1109/ICHIT.2008.292

[22] Hsin-Liu (Cindy) Kao, Artem Dementyev, Joseph A. Paradiso, and Chris Schmandt. 2015. NailO: Fingernails As an Input Surface. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3015–3018. https://doi.org/10.1145/2702123.2702572

[23] James E Katz and Elizabeth Thomas Crocker. 2015. Selfies| selfies and photo messaging as visual conversation: Reports from the United States, United Kingdom and China. *International Journal of Communication* 9 (2015), 12.

[24] David Kim, Otmar Hilliges, Shahram Izadi, Alex D. Butler, Jiawen Chen, Iason Oikonomidis, and Patrick Olivier. 2012. Digits: Freehand 3D Interactions Anywhere Using a Wrist-worn Gloveless Sensor. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST '12)*. ACM, New York, NY, USA, 167–176. https://doi.org/10.1145/2380116.2380139

[25] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder. 2015. The Visual Object Tracking VOT2015 Challenge Results. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

[26] Manu Kumar and Terry Winograd. 2007. Gaze-enhanced Scrolling Techniques. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology (UIST '07)*. ACM, New York, NY, USA, 213–216. https://doi.org/10.1145/1294211.1294249

[27] Gierad Laput, Robert Xiao, Xiang 'Anthony' Chen, Scott E. Hudson, and Chris Harrison. 2014. Skin Buttons: Cheap, Small, Low-powered and Clickable Fixed-icon Laser Projectors. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. ACM, New York, NY, USA, 389–394. https://doi.org/10.1145/2642918.2647356

[28] Oliver Lemon. 2012. *Conversational Interfaces*. Springer New York, New York, NY, 1–4. https://doi.org/10.1007/978-1-4614-4803-7_1

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 21–37.

[30] Alan Lukezic, Tomas Vojir, Luka Cehovin Zajc, Jiri Matas, and Matej Kristan. 2017. Discriminative Correlation Filter With Channel and Spatial Reliability. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Marwa Mahmoud, Tadas Baltrušaitis, Peter Robinson, and Laurel D. Riek. 2011. 3D Corpus of Spontaneous Complex Mental States. In *Affective Computing and Intelligent Interaction*, Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 205–214.

[32] Marwa Mahmoud and Peter Robinson. 2011. Interpreting Hand-Over-Face Gestures. In *Affective Computing and Intelligent Interaction*. Springer Berlin Heidelberg, Berlin, Heidelberg, 248–255.

[33] Emiliano Miluzzo, Tianyu Wang, and Andrew T. Campbell. 2010. EyePhone: Activating Mobile Phones with Your Eyes. In *Proceedings of the Second ACM SIGCOMM Workshop on Networking, Systems, and Applications on Mobile Handhelds (MobiHeld '10)*. ACM, New York, NY, USA, 15–20. https://doi.org/10.1145/1851322.1851328

[34] Pranav Mistry and Pattie Maes. 2009. SixthSense: A Wearable Gestural Interface. In *ACM SIGGRAPH ASIA 2009 Sketches (SIGGRAPH ASIA '09)*. ACM, New York, NY, USA, Article 11, 1 pages. https://doi.org/10.1145/1667146.1667160

[35] Behnaz Nojavanasghari, Charles E. Hughes, Tadas Baltrusaitis, and Louis-Philippe Morency. 2017. Hand2Face: Automatic Synthesis and Recognition of Hand Over Face Occlusions. *CoRR* abs/1708.00370 (2017). arXiv:1708.00370 http://arxiv.org/abs/1708.00370

[36] Jagdish Lal Raheja, Karen Das, and Ankit Chaudhary. 2012. Fingertip Detection: A Fast Method with Natural Hand. *CoRR* abs/1212.0134 (2012). arXiv:1212.0134 http://arxiv.org/abs/1212.0134

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[38] I. Scott MacKenzie and Behrooz Ashtiani. 2011. BlinkWrite: efficient text entry using eye blinks. *Universal Access in the Information Society* 10, 1 (01 Mar 2011), 69–80. https://doi.org/10.1007/s10209-010-0188-6

[39] Marcos Serrano, Barrett M. Ens, and Pourang P. Irani. 2014. Exploring the Use of Hand-to-face Input for Interacting with Head-worn Displays. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3181–3190. https://doi.org/10.1145/2556288.2556984

[40] Srinath Sridhar, Anders Markussen, Antti Oulasvirta, Christian Theobalt, and Sebastian Boring. 2017. 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)*. ACM, New York, NY, USA, 3891–3902. https://doi.org/10.1145/3025453.3026005

[41] Robert Walter, Gilles Bailly, Nina Valkanova, and Jörg Müller. 2014. Cuenesics: Using Mid-air Gestures to Select Items on Interactive Public Displays. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '14)*. ACM, New York, NY, USA, 299–308. https://doi.org/10.1145/2628363.2628368

[42] Cheng-Yao Wang, Min-Chieh Hsiu, Po-Tsung Chiu, Chiao-Hui Chang, Liwei Chan, Bing-Yu Chen, and Mike Y. Chen. 2015. PalmGesture: Using Palms As Gesture Interfaces for Eyes-free Input. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15)*. ACM, New York, NY, USA, 217–226. https://doi.org/10.1145/2785830.2785885

[43] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[44] Koki Yamashita, Takashi Kikuchi, Katsutoshi Masai, Maki Sugimoto, Bruce H. Thomas, and Yuta Sugiura. 2017. 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)*. ACM, New York, NY, USA, Article 19, 8 pages. https://doi.org/10.1145/3139131.3139146

[45] Jian Zhao, Ricardo Jota, Daniel J. Wigdor, and Ravin Balakrishnan. 2016. Augmenting Mobile Phone Interaction with Face-Engaged Gestures. *CoRR* abs/1610.00214 (2016). arXiv:1610.00214 http://arxiv.org/abs/1610.00214