D-Touch: Recognizing and Predicting Fine-grained Hand-face Touching Activities Using a Neck-mounted Wearable

Hyunchul Lim
Cornell University
Ithaca, USA
hl2365@cornell.edu

Ruidong Zhang
Cornell University
Ithaca, USA
rz379@cornell.edu

Samhita Pendyal
Cornell University
Ithaca, USA
sp2377@cornell.edu

Jeyeon Jo
Cornell University
Ithaca, USA
jj693@cornell.edu

Cheng Zhang
Cornell University
Ithaca, USA
chengzhang@cornell.edu

ABSTRACT

This paper presents D-Touch, a neck-mounted wearable sensing system that can recognize and predict how a hand touches the face. It uses a neck-mounted infrared camera (IR), which takes pictures of the head from the neck. These IR camera images are processed and used to train a deep-learning model to recognize and predict touch time and positions. The study showed D-Touch distinguished 17 Facial related Activity (FrA), including 11 face touch positions and 6 other activities, with over 92.1% accuracy and predict the hand-touching T-zone from other FrA activities with an accuracy of 82.12% within 150 ms after the hand appeared in the camera. A study with 10 participants conducted in their homes without any constraints on participants showed that D-Touch can predict the hand-touching T-zone from other FrA activities with an accuracy of 72.3% within 150 ms after the camera saw the hand. Based on the study results, we further discuss the opportunities and challenges of deploying D-Touch in real-world scenarios.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; **Human computer interaction (HCI)**.

KEYWORDS

Hand-face touching recognition and prediction, Deep learning, Computer vision

ACM Reference Format:
Hyunchul Lim, Ruidong Zhang, Samhita Pendyal, Jeyeon Jo, and Cheng Zhang. 2023. D-Touch: Recognizing and Predicting Fine-grained Hand-face Touching Activities Using a Neck-mounted Wearable. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3581641.3584063

IUI '23, March 27–31, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0106-1/23/03…$15.00
https://doi.org/10.1145/3581641.3584063

1 INTRODUCTION

As recommended by CDC[1] and WHO[2], one critical step to reduce the risk of infection with COVID-19 is to avoid touching the face with bare hands since the virus enters through the mucous membranes of the eyes, nose, and mouth (i.e., facial T-zone [36, 49]), mostly by self-inoculation. However, it is challenging for people to stop touching their faces since touching the face is a natural and habitual behavior that many people often practice in a variety of daily activities. People can touch their eyes, nose, and mouth more than 23 times an hour [29] based on the result shown in an observational study. Furthermore, the frequency of touching the face is an informative indicator of the level of stress [22]. Therefore, understanding how people touch their faces can potentially help alleviate multiple important health challenges.

To detect how people touch their faces, the dominant research approach is to deploy a camera in front of the user to observe the user's behaviors. These recorded videos are either manually decoded by researchers in observational studies[5], or analyzed by computer vision algorithms [3] [4] [5, 34, 39]. Unfortunately, these systems will not work if the user is not facing the camera or in the context where it is hard to place a frontal camera(e.g., in motion). To overcome this limitation, researchers in the wearable community have developed a variety of wearable-based hand-face touching (HFT) recognition systems using a smartwatch[11, 21], a wrist band[5][71], rings[6], necklace[21], and eyeglasses[40]. These wearable-based systems enable tracking hand-face touching behaviors in the wild. However, all of these projects can only identify whether the hand touches the face only when the distance between the hand and face is small enough. Such an operation can raise many false-positive errors in the wild because many daily activities require the user to raise hands closer to the face, such as putting on earbuds or eyeglasses, eating, and drinking. People even must hold their hands

[1] Centers for Disease Control and Prevention, https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html
[2] World Health Organization, who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public
[3] How YOU Can Use Computer Vision to Avoid Touching Your Face!, https://medium.com/microsoftazure/how-you-can-use-computer-vision-to-avoid-touching-your-face-34a426ffddfd
[4] How to stop touching your face? Hands and face detection with Python, https://medium.com/analytics-vidhya/how-to-stop-touching-your-face-hands-and-face-detection-with-python-60ecb0d28d69s
[5] Immutouch, https://immutouch.com/
[6] Saving Face, https://www.media.mit.edu/projects/saving-face/overview/

close to their faces when they wear masks. Thus, such devices may falsely record the touching face behaviors when people conduct other activities after washing their hands.

In order to accurately recognize hand-face touching behaviors in the wild, the recognition system needs to move beyond binary classification about whether the hand touches the face. It also needs to distinguish which areas on the head that are or will be touched by hands. Recognizing different touch areas can help distinguish different hand activities around the head. Furthermore, touching different areas carries different risks to health. For instance, contacting a mucous membrane such as the facial T-zone would introduce higher risks of transmitting the virus than touching non-mucous areas such as the chin and cheek. Therefore, it is critical to recognize which areas are touched by hand.

To address this challenge, we develop D-Touch, a neck-mounted wearable sensing system that detects when and where the hand touches the face. The sensing principle of D-Touch is that although the images captured from the neck do not include the complete picture of the head, nor directly show which areas are touched by hands, these incomplete pictures of the head are highly informative to infer the hand-face touching activities, including whether and where the hand touches the face. D-Touch uses a neck-mounted infrared camera (IR), which captures images of activities (e.g., hand touching face) around the head from the neck. These images are learned by a customized deep learning model to recognize 17 Face-related Activities, such as 11 facial areas(e.g., mouth, eye, forehead, eyebrow, cheek, and chin) and 6 daily activities(e.g., eating, drinking, and calling), with an average accuracy of 92.1%.

Accurately recognizing where the hand touches the face is the first step towards alleviating the health risks introduced by hand-face touching behaviors. In order to reduce these behaviors, many researchers are developing various behavior intervention technologies [11, 13, 21, 71]. However, to provide just-in-time intervention, the system needs to predict the behavior in advance rather than simply detect the touch behavior. Therefore, D-Touch made the first effort to predict whether and where the hands touch the face. In our user study, D-Touch was able to predict 2 hand-face touch-related activities with an average accuracy of 82.12%, 150 ms after the hand first appeared in the captured image. To further understand how early intervention needs to be provided to stop hand-face touching behavior, we conducted another user study. It showed that if a user can be notified 150 ms before the hand touches the face, participants were able to stop the touching behavior with a success rate of 72.1%. This first-of-its-kind study result is preliminary and requires much more data to draw any conclusion. However, it indicates the potential of using D-Touch for hand-face touching behavior intervention.

To the best of our knowledge, D-Touch is the first wearable-based sensing system that can both recognize and predict where the hand touches the face, and it is also the first study that explores how early the system should predict hand-face touching behavior in advance for proper intervention. The contribution of this paper is:

- We implemented the first wearable system that can distinguish a rich set of hand activities around and on the face, including 1) 11 touched areas on the face; 2) 6 other activities in which hands are close to the face, and 3) other daily activities.
- We made the first attempt to predict the areas of touching on the face in advance before the touching actions happen, which offers potential future opportunities for behavior intervention.
- We conducted user studies to evaluate the performance of the system which showed promising accuracy in both recognizing and predicting a variety of hand-face touching activities in both lab settings and in the wild (10 participant's homes without limiting participant's behavior).
  for recognizing and predicting the hand-face touching behaviors with 10 participants in the wild (their homes without any constraints).
- We discussed the limitations of the current system and the challenges and opportunities of applying D-Touch at scale in real-world applications.

2 RELATED WORK

This section reviews the literature related to hand-face touching behavior and discusses existing sensing technology for detecting hand-face interaction with non-wearable or wearable devices.

2.1 Hand-face touching behavior

Touching the face with hands is a common and sometimes unconscious behavior. There are various reasons why people touch faces with their hands, such as people's habitual characteristics (e.g., rubbing eyes, scratching nose, curling fingers against mouth, twirling mustaches) [29, 42] or behavioral disorders like Onychophagia [63]). Furthermore, people tend to stroke the chin while listening to others, pondering questions, or expressing embarrassment or fatigue [22, 37, 38, 44]. Additionally, a few observational studies have conducted experiments and showed that people touch their face between 9 to 54 times on average per hour [1, 2, 5, 8, 11]. For example, in [29], an average count of 23 face-touches per hour is observed on 26 students. Similarly, in another study [2], ten subjects were recruited to perform office-type work in controlled setting for three hours. The video was recorded and manually analyzed, which showed on average, each participant touched their face for 15.7 times. However, these studies review the hand-face touching behaviors by manually analyzing the video data, which is not efficient and can not be largely deployed to collect the hand-face touching behavior data. To comprehensively understand human hand-face touching behavior, researchers seek to build computing technologies to track when, where, and how the hand touches the face in daily activities.

2.2 Hand face touching detection with non-wearable-based approach

Frontal cameras have been used extensively in the computer vision (CV) field to study people's facial movements [58, 66, 72] and hand gestures [41, 65] as two separate topics. Researchers have developed both classical methods based on hand-crafted features [1, 8, 23, 27,

28, 35, 46, 47, 60, 67, 68], and deep-neural-network-based algorithms [33, 45, 61, 62] to detect human faces in images captured by a frontal camera. Moreover, researchers apply various CV algorithms on images or videos to detect, track, and recognize hand gestures [3, 14–18, 30, 51, 70] that are static [19, 50, 52] or dynamic [12, 57, 64].

Hand-over-face occlusion is a challenging problem for face detection researchers in the CV field, as facial features may be erroneously detected or lost because of occlusion. To address this issue, researchers in [39] tried to learn gesture descriptors such as hand shape, hand action, and occluded facial region. Similarly, [5] provides a large-scale annotated data set for hand-face behaviors in social interaction and a CNN model for detecting touching actions with an accuracy of 83.76%. Additionally, in [34], researchers built a novel input modality for interaction with smartphones with hand-over-face gestures. Due to the recent COVID-19 pandemic, researchers have started paying more attention to detecting hand-face touching behavior with frontal-camera-based vision methods. For example, the website "Do not Touch Your Face"[7] use images from a webcam to determine whether the user is touching his face.

However, because hand-face touching (HFT) is a deeply ingrained and highly unintentional behavior that happens irregularly every day, it is impossible to deploy a frontal-camera to detect HFT constantly in people's daily life. Moreover, the frontal-camera-based methods usually requires a suitable view-angle and lighting condition, which is limited in many human activities.

2.3 Hand face touching detection with wearable-based approach

Along with trying to understand human behavior, researchers also build wearable devices to track activities and then support user self-awareness and intervention [55]. For example, Fitbit+ is used to reduce sedentary behavior [48], and various devices have been used to detect eating behavior for avoiding unhealthy weight gain [75]. Furthermore, because of the COVID-19 pandemic, people have started paying more attention to hand-face interaction behaviors. Because of this increased awareness, we must develop wearable devices for users to track their touching behavior.

Researchers have developed various systems using sensors in wearable devices to address the limitations of detecting hand-face contact with a frontal camera. For example, Face Guardian[8] uses a magnet on the wrist and phone that has compass ability placed on the neck to detect the distance of hand to face by measuring the magnetic field. It will provide vibration and audio alerts every time the app detects there is one touching action. Similarly, researchers in the Saving Face project[9] and [21] built devices by analyzing electromagnetic or capacitive fields between a transmitter ring or a bracelet on hand and the receiving sensor around the face or on the neck. Another solution proposed by researchers in Saving Face project is developing an HFT notification app by monitoring ultrasonic frequency changes from earphones. Likewise, NoFaceContact [71] utilizes near-field communication (NFC) with the reader on the wrist and a tag on the ear. However, these devices need two sensors wearing on the neck and wrist, respectively. Researchers also provide solutions with only one smartwatch to detect HFT behaviors by investigating the data from accelerometer [11, 56], inertial and magnetic sensors [21] or gravimeter [10]. TouchAlert[56] utilizes the sensors found in common wearable devices to train a deep learning model for predicting variable length face-touching at an early stage of its occurrence. Recently, FaceSense [26] uses multiple sensors, i.e., thermal and physiological sensors, to detect touches and recognize the facial zone of the touch.

Nevertheless, these devices all work on identifying when a hand comes closer to the face rather [11, 21, 26, 56] than detecting fine-grained hand-face touching behaviors. People sometimes hold their hands up not to touch the face but to do other activities like drinking, eating, and wearing glasses. Previous devices will mistakenly keep recording touching actions for these activities, despite them not being dangerous. Furthermore, identifying and predicting areas that people are touching is essential because touching facial areas like the eyes, nose, and mouth will raise the probability of infecting with the virus compared to other parts of the face. D-Touch is the first wearable technology that can recognize and even predict areas that the hand touches.

3 D-TOUCH

In this section, we discuss the goals and principles of designing a wearable system that can address our research question.

3.1 Design principle

Our goal is to design a wearable technology that can accurately recognize and predict which areas the hand touches the face. In order to capture the hand activities around the head, we decide to use IR cameras.

Most of the existing camera-based solutions would place the camera in front of the user to capture the face, which has two limitations. First, if the camera is set up in the environment, it may not work well when the user is in motion or does not face the camera. Second, it can only work when users wear/hold the camera in front of the face, which is inconvenient and less socially appropriate. Moreover, it is unable to track throughout the day continuously.

We consider several form factors and positions for the camera inspired by previous research (i.e., cap [54], headphones [10], necklace [9], chest-mounted camera [7, 25], shoes [4], and wristband [24, 31]). We evaluate different positions and form factors for our device with the design goals in mind, including a camera on a hat, on the chest, on a watch, and headphones. We decide to place the camera on a neck-mounted form factor for several reasons: 1). Placing the camera on the neck would offer the best view to observe the hand activities around the head without much occlusion. 2). People have already used to wearing decorations or devices around the neck, such as necklace, and neck-mounted Bluetooth speakers. People are more acceptable to wear smart neck-mounted wearables. 3). A neck-mounted device would offer larger space for design and can carry relatively larger weight, compared to other locations (e.g., ear), the space in front of the neck offers a large flat surface, which can steadily hold a relatively larger device (e.g., with larger battery).

[7] https://lazerwalker.com/dont-touch-your-face/
[8] https://matter.childmind.org/face-guardian.html
[9] https://www.media.mit.edu/projects/saving-face/overview/
[10] https://immutouch.com/guard

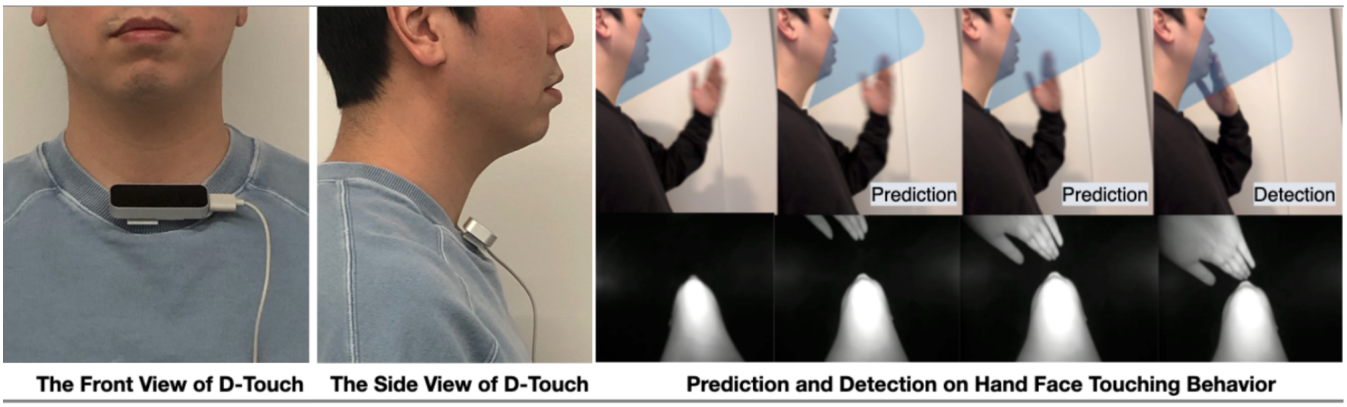

Figure 1: D-Touch: A tie-clip wearable device using an IR camera (i.e. LeapMotion). By pointing a camera toward the face from the tie clip, D-Touch is able to capture hand-face touching(HFT) behaviors for recognition and prediction.

After choosing the neck as the position of a wearable device, we further examine different neck-mounted form factors. We decide to attach the camera to a clip (e.g., tie-clip). This allows our device to be used in a similar manner as a brooch, as people could easily remove and attach the device throughout the day. Furthermore, by pointing the camera toward the face from the tie clip, the camera would only capture the neck and part of the face (e.g., chin, jaw), which contains less privacy about the user and surrounding environment.

In the end, we evaluate different options for cameras, including RGB camera, depth camera, and IR camera. RGB cameras would offer high resolution on images with minimal weight and size. However, segmenting the human body from different backgrounds in RGB images is still a challenging task in computer vision. Depth cameras would help separate the human body from the background in the depth images. Nevertheless, the current depth cameras' size and weights are too large, making them inappropriate as a neck-mounted wearable. As a result, we chose Infra-red (IR) camera, which emits the IR lights to the human body and receives the reflected IR lights from the body in the images. It perfectly satisfies our needs. On the one hand, it is relatively easy to segment the human body from the background. On the other hand, the dimension and weight of existing IR cameras are small enough to fit nicely into a neck-mounted wearable (e.g., tie-clip).

As a result, we developed D-Touch that uses an IR camera as the sensing device attached to a tie-clip. To assess the likelihood of using D-Touch, we conducted a small online survey with 22 participants. Participants ranged from students to working professionals and had an age range of 19 to 45 years old. Of the 22 participants, 15 (68%) said that they would consider using D-Touch given certain conditions such as reduced privacy concerns, comfort, and affordability. However, if D-Touch offered additional health features, such as eating tracking or hand-face touching intervention, 20 participants (90%) expressed their willingness to use it. The survey results showed that a neck-mounted form factor for D-Touch has the potential as a health-tracking device.

In the following sections, we will present how we use this neck-mounted sensing device for 1) recognizing and 2) predicting whether and when the hand touches the face.

3.2 Research Idea and Questions

With our form factor, i.e., a neck-mounted tie-clip device, we explore the possibility of detecting which parts of the face are being touched. We conducted a preliminary experiment by attaching an RGB camera to the neck. As shown in Figure 2, we find that the images captured by the neck-mounted camera are highly informative on the hand positions (e.g., nose, mouth, chin, cheek, eye, eyebrow, and forehead) and other daily activities related to the face (e.g., eating, drinking, and calling). Therefore, we suspect that the images of the face captured from a neck-mounted camera can be used to distinguish whether and where the hand touch the face. Also, we found that our hands in the captured images have both a different hand shape and trajectory depending on the hand-face behavior, indicating that we can potentially predict which face area is about to be touched before the hand touches the face. In order to provide the necessary intervention to stop users from touching their faces with hands, we need to predict their behaviors as early as possible before hand-face touch occurs. Based on the findings, we proposed key research questions on hand face touching classification and prediction behind D-Touch:

- `Research Question 1`: Is it possible to classify and predict hand face touch behavior with sequence images of the face and a hand captured by a camera-mounted tie-clip device on the neck before the hand touches the face?
- `Research Question 2`: How early D-Touch should predict the HFT activities to provide proper intervention to stop touching the face with the hand?

4 SYSTEM DESIGN

To answer our research questions, we design the D-Touch with a prototype, image preprocessing, and deep learning algorithms. Fig 4 shows the overview of D-Touch.

4.1 Prototype

Wearing an IR camera on the neck could provide obvious differences in image appearance when hands touch the face in various

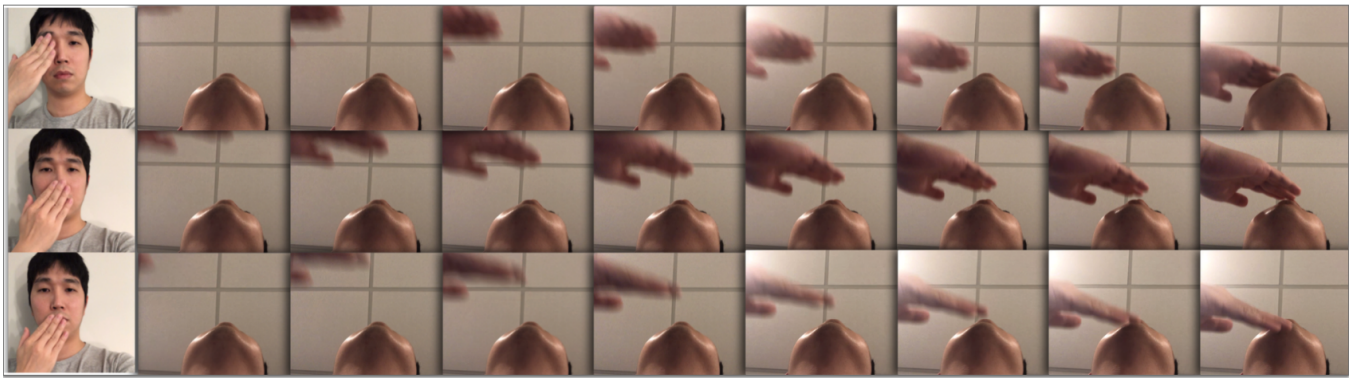

Figure 2: Hands in the captured images have both different hand shape and trajectory depending on the hand ace behavior. Thus, it would be possible to classify and predict when and where the user is about to touch their faces using images captured before the initial contact.

ways. Here, we use LeapMotion[11] for classifying and predicting the HFT behaviors. The rationale for using LeapMotion is to capture the hand as early as possible so that we can acquire maximal information of the hand as it moves. LeapMotion's wider field of view and longer tracking range (i.e., FoV : 140 x 120 degrees and Tracking Range: 60 to 80 cm [12]) helps to capture the hand earlier than a camera with a narrower view would. Also, with the help of an advanced IR light source and an IR filter, LeapMotion is able to adjust the direction of the lens toward the space in front of the face with the least amount of background noise.

The sensing method of LeapMotion employs pattern-less IR light on the object and IR camera to capture IR images. LeapMotion utilizes more advanced hardware (two monochromatic IR cameras and three brighter IR LEDs) and an advanced exposure strategy and algorithm. Thus, it has a wider field of view and longer tracking range (i.e., 140×120° typical field of view and depth of up to 60cm (24") preferred, up to 80cm (31") maximum), allowing for earlier capturing of the hands, along with less lighting issues. LeapMotion is connected to a laptop via CSI interface using a 3 meter long USB 2/3 hybrid cable and transmits data with a resolution of 150x150 at 40 fps. The camera module is also encased in a 3D printed shirt-clip. The shirt-clip case, along with the camera module, can be secured at the top of the participant's shirt, with the camera pointing upward towards the participant's chin, as shown in Figure 3.

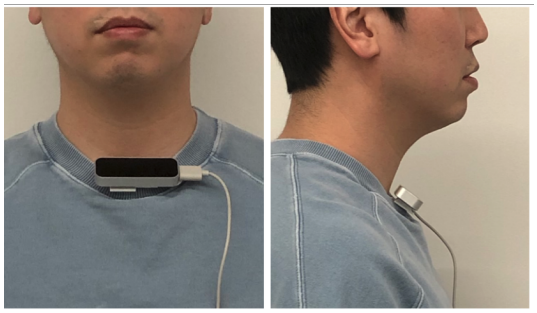

Figure 3: Experiment setting with LeapMotion. The clip is placed on the participant and the camera is pointed toward the chin.

4.2 IR Image Processing

The primary purpose of image processing is to reduce noise and other interfering factors by using image preprocessing algorithms to help subsequent neural networks improve inferring ability and generalization by data augmentation techniques.

Image Preprocessing: As shown in Figure 1, LeapMotion provides infrared-format data as a grey-scale image. We utilize the resized gray-scaled images (150x150) from LeapMotion without particular image preprocessing as input data for our deep learning pipeline.

Data Augmentation: Camera shifting is an important issue for any wearable camera. Shifting leads to a significant variance in images, which may ultimately fail our image analysis algorithms. For example, walking poses a particular challenge, as the camera position and angle shift with every stride. The camera also shifts whenever the user remounts the device. Although we try to improve the camera's stability with a suitable form factor design, movement is inevitable throughout daily activities. Therefore, we deploy data augmentation techniques to simulate such changes by manipulating the data. More specifically, we drop a probability with Gaussian distribution to decide whether to conduct translation, rotation, scaling or any combination of these three on the images before feeding them into the model in each epoch. The threshold of the Gaussian probability model we decide to augment is 0.5. The range of each manipulation is translation range from -30 to 30 pixels, rotation range from $-20°$ to $20°$, and scaling range from 0.8 to 1.2, respectively.

4.3 Deep Learning pipeline

In this section, we introduce the structure and training configurations of our deep learning model TouchNet, the framework being a combination of convolutional neural network (CNN) and Long short-term memory (LSTM). We define two sub-models: the CNN

[11] https://www.ultraleap.com/product/leap-motion-controller/
[12] https://www.ultraleap.com/company/news/blog/how-hand-tracking-works/

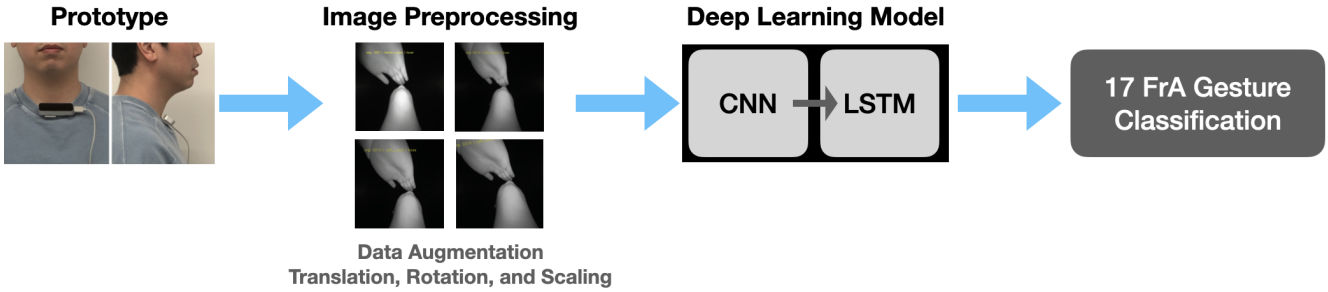

Figure 4: System Overview: D-Touching consists of three parts: Prototype, Image Preprocessing, and Deep Learning model

model for feature extraction and the LSTM model for interpreting the features across time steps.

Image-related tasks using CNN have achieved significant advances in recent years [6, 32, 53, 74] like hand gesture recognition [41, 65] and facial expression recognition [58, 66, 72]. As VGG 16 [59] has demonstrated its performance to be highly effective for many image tasks, which is also easy to achieve convergence, we chose VGG 16 as the backbone of TouchNet for our image classification task. Next we deploy an LSTM model to handle the time series data followed by a fully connected layer with a softmax function to classify FrA gestures.

4.3.1 TouchNet Architecture. : TouchNet is composed of two main parts: the backbone based on VGG 16 and LSTM, and a classification block. The VGG 16 backbone of TouchNet consists of five blocks, with the first two blocks having two convolutional layers and a max-pooling layer, and the last three blocks having three convolutional layers and a max-pooling layer. All the convolutional layers have a kernel size of 3×3 and use "same" padding, ensuring that the spatial dimensions of the feature maps remain the same after convolution. The convolutional layers are followed by rectified linear unit (ReLU) activations, which introduce non-linearity into the network. Max-pooling layers are used to reduce the spatial dimensions of the feature maps while retaining the most important information. The max-pooling layers have a kernel size of 3×3 and use "valid" padding, meaning that the output size is smaller than the input size. The TimeDistributed layer is used to wrap the entire sequence of CNN layers, integrating the CNN and LSTM models for improved performance in recognizing hand-face touching gestures. The fully connected layers are then used to learn high-level representations of the input data, with a dropout layer (probability=0.8) placed between each pair of fully connected layers to prevent overfitting. Finally, a Softmax layer is used to produce the final prediction, with 17 units corresponding to 17 facial-related gestures shown in Fig 5.

4.3.2 TouchNet Training. : We built the deep learning network under the TensorFlow framework. During the training process, the hyper-parameters we chose are listed as follows: gradient descent optimizer with a learning rate of 0.01 and batch size of 30. When training the network, we first use the model pre-trained on Imagenet dataset [20][13] to accelerate the training process. We then resize the images to 150×150 to meet the input requirement of the VGG16 network. Then we wrapped the VGG 16 network for LSTM by using the Time Distributed layer. We randomly shuffle the data, divide the data into several batches and train the model with 500 epochs for all participants. For training the model, we use our server with the GPU (AMD Thread- ripper 3960X CPU and RTX2080Ti GPUs with 256GB memory). Lastly, we evaluate the performance of the length of the input image sequence. Since our data sample is limited, we use a 5-fold cross-validation evaluation where 80% of data were used as the training and the last 20% as the testing data. Five models are trained and evaluated with each fold.

[13]https://github.com/tensorflow/models/tree/master/research/slim

5 EVALUATION

In order to evaluate how D-Touch can distinguish and predict Face-related Activities (FrA), we conducted a user study with 10 participants with three experiments: 1) 17 FrA classification and prediction in a controlled study setting, 2) an intervention study where we evaluated how early intervention is needed to stop face-touching behavior, and 3) the in-the-wild study at participants' home where we study how D-Touch would recognize and predict FrA in the wild. Each experiment was conducted under safety procedures approved by the Institutional Review Board (IRB) of the authors' institution and completed within 100 minutes.

5.1 Participants

We recruited 10 participants for three experiments, with 3 of them being female. Their ages ranged from 18 to 35, and the average age was 26.1 (SD: 4.1). All participants completed all three experiments without any issues. There were 3 participants with long hair and 1 participant wearing glasses, while there were no left-handed participants.

5.2 17 Face-related Activity

We consider 17 Face-related Activity for classification and prediction to validate D-Touch. To define fine-grained hand-face touching activities, we first consider where people touch the face with hands according to the previous and CDC's guidelines, and then we divide the face into 11 areas including facial T-zone (i.e., eyes, nose, and mouth) and other facial areas (i.e., forehead, cheek, and chin) as shown in Figure 5. Moreover, we aim to differentiate hand-face touching from other face-related activities which may bring hands close to the face but do not involve direct touching. We choose 6 common daily activities: 1) eating, 2) drinking, 3) calling, and

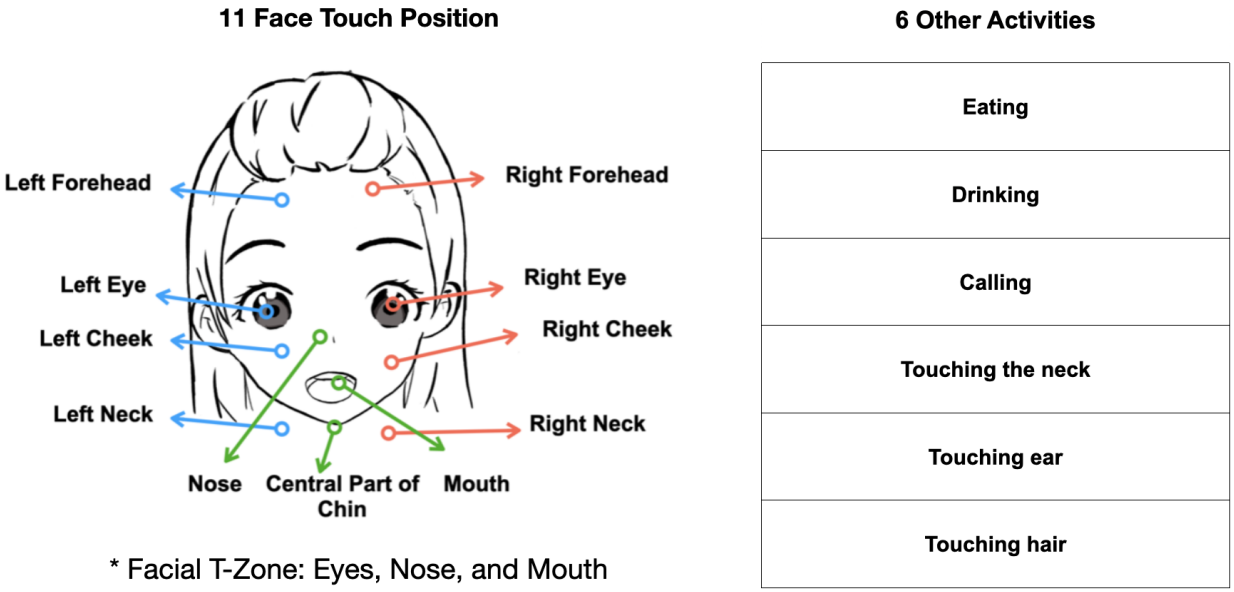

Figure 5: The 11 Facial Areas and 6 Common Daily Activities related to the Face.

touching 4) the neck, 5) ear, and 6) hair. These six activities are included in our study to avoid false positive errors in detecting hand-face touching.

5.3 Device Setup and Study Location

For data collection on D-Touch, participants were asked to wear a clip-shaped device with Leap Motion camera on the collar of their shirts as shown in Fig 3. The researcher then helped the participants adjust the device to a fixed position and fine-tuned the device to a suitable angle to capture images of the face. Since all experiments require participants to touch their faces with their hands, the experiments were conducted in participants' homes to avoid the risk of exposure to COVID-19. Also, participants were asked to wash their hands carefully before conducting the experiments. Each experiment was conducted under safety procedures approved by the Institutional Review Board (IRB) of the authors' institution and completed within 100 minutes.

For the in-the-wild data collection, the device was connected to a laptop using a 3-meter-long wire, allowing the participants to move freely in their homes within a radius of 3 meters. Furthermore, the researcher was not present in the experiment, which allowed the participants to freely conduct activities at home. All participants reported that although they had limited space to move around due to the wire, they did not find it difficult to move around their home while wearing the device with the long wire.

5.4 Experiment 1. Classifying and predicting 17 FrA gestures

To evaluate how D-Touch recognizes and predicts these FrA gestures, we collected data on 17 FrA gestures from ten participants in a controlled setting at their homes.

5.4.1 Procedure. The study was conducted at the participant's home. Participants were asked to sit in front of the table. We placed a screen in front of the participants to display the instruction for the actions that the participants need to perform. To simulate FrA gestures in the real world, participants were asked to perform gestures with two hand types (i.e., the left and right hand) respectively in random order by following the video instruction. In addition, to collect various images of hand shapes when performing FrA gestures, we asked participants to perform each gesture twice consecutively with different hand poses.

5.4.2 Data Collection. Participants repeated five times for each gesture, which lead to 10 samples per each gesture (2 different hand poses x 5 times). The order of each gesture was randomized. The 340 gestures were collected per participant (17 FrA x 2 hand types x 10 samples) with about 56,000 image frames (40 fps x 60s x 20 m). As a result, we collected 3,400 gestures with about 560,000 images from ten participants. This experiment took about 25 minutes including a practice session.

5.4.3 Labeling for Ground Truth. The labeling for ground truth is done manually for all participants in this study. We used another front-facing camera to record which parts of the face they touched and when the touch occurred. For a hand touch gesture, we label all frames in that segment (from the first frame shown in the camera to the last frame when the hand disappears). We pay more attention to the frames for the prediction task before the touching action happens.

5.4.4 Result. We first present the statistical analysis of how long it takes for participants to perform the 17 FrA gestures after the hands enter the camera frame. The hand appearing at the top of the frame was determined by using a threshold. Then, we report the

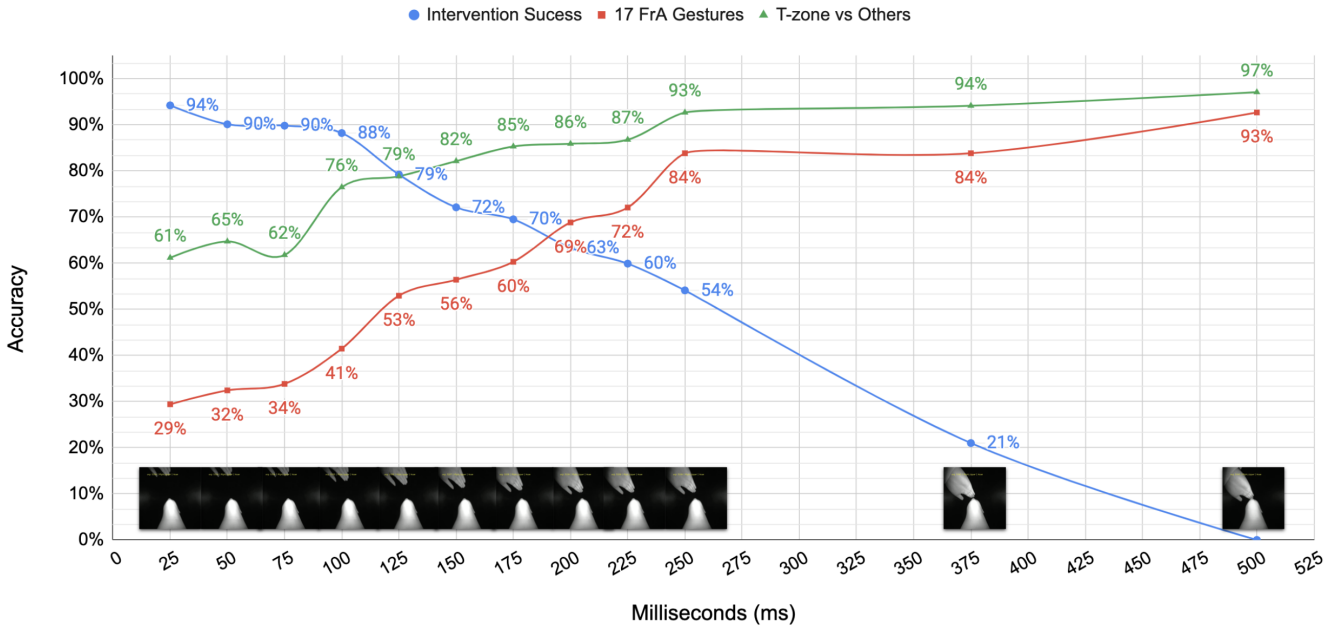

Figure 6: A trade-off between accuracy and intervention success

performance of our models in classification and prediction using different lengths of image input (5 to 20 images), evaluated with 5-fold cross-validation. The best input length for the classification task was found to be 20 sequential images (500ms), as no improvement in performance was seen with shorter or longer input lengths.

Table 1 summarized the statistical values of performing FrA gesture. On average, it took 425.6 ms to perform FrA gestures (SD = 121.4 ms, ranges = 250 - 950 ms). We found that it took a shorter time to detect the moving hand that was touching the parts of the face that is further away from the camera such as the forehead. In addition, it also showed that the moving speed of the hand was much slower in other non-face touching behavior such as eating, drinking, and calling than the movement speed of the hand when touching the face.

The result showed that D-Touch was able to distinguish the 17 FrA gestures with an average accuracy of 92.1% within about 500 ms after the hand appeared in the camera. As shown in Fig 6, the initial accuracy seems when only using the first couple of frames after the hand shown in the camera, as it contains less information about the moving trajectory of the hand. But the accuracy increased when the model see more frames of hand in the camera. The accuracy increased to 92% on 17 FrA activities within about 500 ms after the hand appeared in the camera, indicating that our technology can be used in tracking people's hand-face touching behavior in detail for their health.

In real-world scenarios, it may not be necessary to distinguish all 17 activities. For example, the most important category for COVID-19 prevention is to distinguish T-Zone v.s. other activities to prevent their hand-face touching activities. In this case, D-Touch can distinguish hand-touching T-Zone vs other activities with an accuracy of 85.9% within 200 ms as shown in Table 2. This indicates that D-Touch has the potential to provide prediction in advance for intervening in the hand-face touching behavior. However, it should be noted that there is a trade-off between accuracy and the time to predict touching T-zone as such prediction should be used to give users intervention to stop touching the face with their hands appropriately. We will explore the appropriate time for the intervention study in Experiment 2. Figure 7 summarizes the result on D-Touch using a confusion matrix.

5.5 Experiment 2. Intervention Study

The ultimate goal of our prediction on FrA is to provide intervention for stopping touching the face in advance, especially the facial T-zone for safety. Here, we investigated how early the intervention needs to be provided in order to allow the user to stop the hand-face touching behavior. This experiment was conducted in a controlled setting after Experiment 1 at the participant's home.

5.5.1 Procedure. In this experiment, we asked participants to sit in front of the table and then touch the mouth or nose with their dominant hand wearing D-Touch using different hand poses. Compared to touching other facial areas in T-Zone, touching the nose or mouth takes the shortest time (an average of about 450 ms). We developed a system using D-Touch which can recognize when the hand appeared in the camera. We gave an alarm sound to the participants at 12 different time points (25, to 500 ms) after the hand to appeared in the camera. The participants were instructed to stop the hand movement as soon as they hear the sound. If the hand touched the face, we counted it as a failed intervention. Otherwise, we counted it as a successful intervention.

5.5.2 Data Collection. Each participant repeated the touch behavior 10 times at each time point (12 in total). The order of the time

Table 1: Basic statistics on Face-related Activity (FrA) gesture set from Experiment 1 (milliseconds) and accuracy (percentage - %)

Label #1	Sample	Mean	SD	Min	Max	Median	Label #2	Accuracy
Total	**3400**	**425.6**	**121.4**	**250.0**	**950.0**	**425.0**	-	**96.1**
Left Eye	200	502.5	65.0	400.0	600.0	512.5	Facial T-zone	100.0
Mouth	200	477.5	54.6	400.0	575.0	487.5	Facial T-zone	81.0
Nose	200	375.0	63.5	250.0	450.0	362.5	Facial T-zone	85.0
Right Eye	200	392.5	37.4	350.0	450.0	387.5	Facial T-zone	85.0
Calling	200	487.5	71.9	375.0	625.0	500.0	Other Activities	100.0
Chin	200	422.5	53.3	350.0	500.0	425.0	Other Activities	88.0
Drinking	200	760.0	63.7	675.0	850.0	762.5	Other Activities	100.0
Eating	200	617.5	98.6	500.0	750.0	587.5	Other Activities	100.0
Hair	200	395.0	49.7	300.0	475.0	400.0	Other Activities	92.0
Left Cheek	200	462.5	44.5	400.0	550.0	450.0	Other Activities	100.0
Left Ear	200	392.5	47.2	325.0	475.0	400.0	Other Activities	92.0
Left Neck	200	362.5	82.7	300.0	575.0	325.0	Other Activities	90.0
Left Forehead	200	435.0	146.8	350.0	850.0	400.0	Other Activities	92.0
Right Cheek	200	402.5	39.9	350.0	475.0	387.5	Other Activities	92.0
Right Ear	200	392.5	40.9	325.0	475.0	400.0	Other Activities	85.0
Right Neck	200	407.5	200.0	275.0	950.0	337.5	Other Activities	96.0
Right Forehead	200	462.5	145.9	350.0	850.0	425.0	Other Activities	96.0

Table 2: Evaluation on D-Touch: Accuracy, Recall, Precision, and F-score from Experiment 1 and Intervention Accuracy (percentage - %). Time: time was recorded after the hand shown in the camera. ms: milliseconds.

	17 FrA Gestures				**T-zone vs Others**				**Intervention**
Time	Accuracy	Recall	Precision	F-score	Accuracy	Recall	Precision	F-score	Success Rate
25 ms	29.4	29.4	0.0	0.0	61.2	60.3	59.2	59.8	94.2
50 ms	32.4	32.4	0.0	0.0	64.7	53.9	77.9	63.7	90.1
75 ms	33.8	33.8	0.0	0.0	61.8	51.2	77.3	61.6	89.8
100 ms	41.5	42.2	46.9	44.4	76.5	72.3	75.6	73.9	88.2
125 ms	52.9	52.9	55.2	54.0	78.8	76.4	77.7	77.1	79.2
150 ms	56.4	58.8	57.9	58.3	82.1	75.5	79.8	77.6	72.1
175 ms	60.3	60.3	65.7	62.9	85.3	75.2	81.3	78.1	69.5
200 ms	68.8	65.3	69.9	67.5	85.9	78.5	80.9	79.7	63.3
225 ms	72.1	72.1	75.6	73.8	86.8	80.5	82.0	81.2	59.9
250 ms	83.8	83.8	86.3	85.0	92.7	90.9	89.2	90.0	54.1
375 ms	83.8	83.8	83.8	83.8	94.1	94.0	90.7	92.3	21.0
500 ms	92.7	93.0	93.3	93.1	97.1	96.2	96.7	96.4	0.0

points was randomized. In total, we collected 120 samples per participant (12-time points x 10 repetitions) with about 27,000 image frames (40 fps x 60s x 10 m). We collected 1,200 samples in total with about 270,000 images from ten participants. The experiment took about 15 minutes including a practice session.

5.5.3 Result. Overall, the intervention study showed that for 150 ms notice in advance (input length = 6 image frames), the participants were able to stop their face-touching activity with an average success rate of 72.1%(SD = 4.1 %). As shown in Fig 6, the initial accuracy seems high (e.g., over 88 %) until about 100 ms but the accuracy of stopping hand-face touching gestures decreased over time, reaching up 20% within about 375 ms after the hand was detected. It is evident that earlier interventions result in a higher success rate in stopping hand-to-face touching. However, there is a trade-off between accuracy and time. Despite the results being from a controlled experiment, this is a promising first step toward understanding the appropriate time to intervene in hand-face touching behaviors. This will be further discussed in section 6.1 and 6.8.

5.6 Experiment 3. Deploying D-Touch in the wild

In this section, we explored how participants performed FrA gestures in the wild and whether D-Touch is able to recognize and predict the gestures. We collected face-related Activity (FrA) data at the participant's home for 1 hour without the presence of the researcher. Here, we did not give participants alarms to intervene in their hand-face touching behavior. Our goal for this experiment

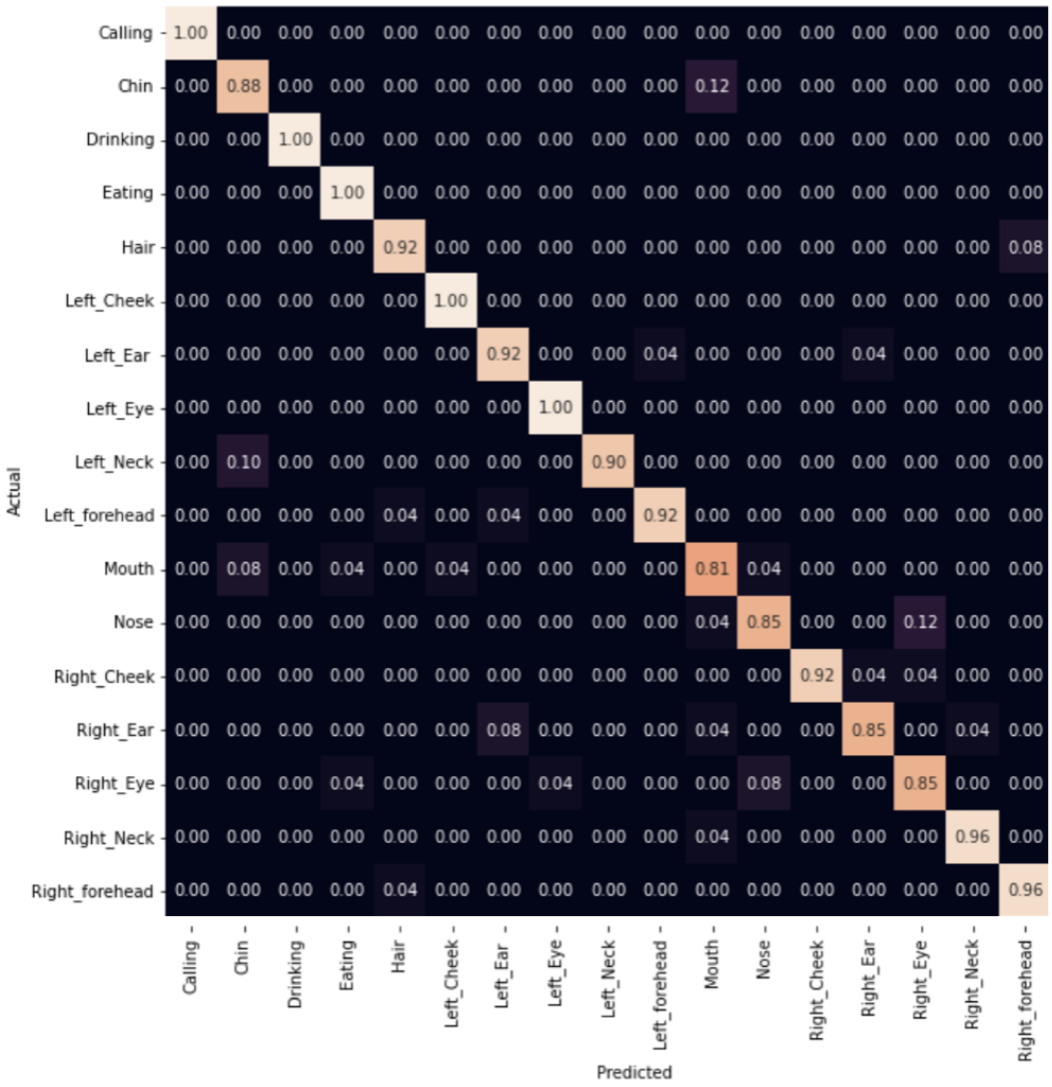

Figure 7: Confusion Matrix of 17 FrA Gesture Detection within 500 ms after the hand shows up in the frame

is to collect hand-face touching behaviors in the wild for further analysis of D-Touch.

5.6.1 Procedure. The researcher helped set up D-Touch with the participants and left them at their homes. The participants were asked to freely do any activities at their homes without any constraints. The purpose of this setup is to collect their natural and unconscious hand-touching activities at their homes where they are most comfortable. Even though the device was connected to a laptop with a 3-meter-long wire, participants are able to move freely in their homes within a radius of 3 meters. To emphasize once again, all participants reported no discomfort in moving their hands and moving around their homes. After the experiment, the researcher returned to the participants' homes to collect the experiment device.

5.6.2 Data Collection. The experiment took 60 minutes for data collection. We collected about 162,000 image frames (40 fps x 60s x 60 m) per participant, and 1.62 M frames in total from ten participants.

5.6.3 Labeling for Ground Truth, Removal of Noise, and Training Models. We manually labeled all frames for all participants by watching the collected images from Leap Motion since it was hard to get ground truth using an additional front-facing camera. For some images, it was not clear what kind of activity the participants performed, so three researchers labeled the data and then chose the most out of the three. In the case where the three researchers disagreed on the images, we asked the participants to confirm their activities. For a hand touch gesture, we labeled all frames in that segment (from the first frame shown in the camera to the last frame when the hand touched the face). We removed some noises such as lighting while participants are moving if the noises appear less than 150 milliseconds (i.e., the five input images). Here, we trained

Table 3: Basic statistics on Face-related Activity (FrA) on the in-the-wild data set (millisecond) and Accuracy (percentage- %)

Label #1	Sample	Mean	SD	Min	Max	Median	Label #2	Accuracy
Total	**503**	**399.0**	**134.4**	**164.7**	**1223.5**	**364.7**	-	**78.2**
Left Eye	12	321.6	69.2	235.3	411.8	323.5	Facial T-zone	75.0
Mouth	92	410.4	196.8	176.5	1011.8	364.7	Facial T-zone	76.1
Nose	59	338.4	115.3	188.2	600.0	305.9	Facial T-zone	66.1
Right Eye	5	331.8	88.2	235.3	400.0	388.2	Facial T-zone	80.0
Calling	1	364.7	0.0	364.7	364.7	364.7	Other Activities	100
Drinking	42	513.4	155.2	329.4	1070.6	464.7	Other Activities	81.0
Eating	53	555.2	270.9	211.8	1223.5	470.6	Other Activities	83.0
Eyeglasses	7	358.0	111.6	247.1	576.5	341.2	Other Activities	0.0
Left Forehead	15	360.0	105.7	247.1	576.5	329.4	Other Activities	80.0
Right Forehead	12	413.7	226.3	211.8	870.6	352.9	Other Activities	58.3
Hair	87	373.1	152.0	223.5	1000.0	329.4	Other Activities	93.1
Left Cheek	53	556.9	238.8	188.2	1094.1	564.7	Other Activities	81.1
Left Ear	15	405.5	116.6	270.6	600.0	376.5	Other Activities	73.3
Right Cheek	21	331.7	76.1	211.8	517.6	317.6	Other Activities	76.2
Right Ear	29	350.9	93.3	164.7	470.6	400.0	Other Activities	75.9

the user-dependent CNN+LSTM models based on two different input lengths, i.g., 6 and 20 image frames, using all data from each participant in Experiment 1 to recognize the gestures in the wild.

5.6.4 Result. We collected 503 FrA gestures in total from ten participants in the in-the-wild experiment. Out of 17 FrA activities we designed in Experiment 1, 15 FrA gestures were recorded including the new activity i.e., adjusting glasses. Three activities such as touching their chin left neck, and right neck did not occur.

On average, the participant performed FrA gestures 49 times per hour (SD: 16.78) (ranging from 24 times to 78 times), indicating that the result seems similar to those of previous hand-facing tracking in the wild [1, 2, 5, 8, 11]. Among these FrA gestures, 46% account for touching the facial areas, and 54% were related to other activities. Each participant touches their face 22.5 times an hour (SD = 9.3) on average.

When touching the face, 70% of the time the participants touched the T-zone, and the remaining 30% of the time they touched the cheeks or forehead. Considering the hand type, 53% is left-hand 44% is right-hand. 3% of the time they use both hands (3%). Although the participants were all right-handed, we found that they often used their left hand to touch the face, while the dominant hand was frequently occupied such as holding a cup or using a smartphone. There was no significant difference for times to touch, $t(563) = 0.3287$, $p = 0.7425$, despite in the in-the-wild condition (M = 422 ms, SD = 194.1 ms) having a higher standard deviation than the controlled setting (M = 425.5 ms, SD = 121.3 ms).

D-Touch was able to distinguish the 15 FrA gestures in the in-the-wild condition with an average of 78.2% (SD = 2.5%) accuracy within about 500 ms (input length = 20 image frames) after the hand first appeared in the camera. The accuracy was lower than that from a controlled setting in Experiment 1 since our models did not train the data such as the new activity, i.e., adjusting glasses, different hand poses, and the use of both hands.

Based on the result of the intervention study, the participants were able to stop their face-touching activity with an average success rate of 72.1%(SD = 4.1 %) within 150 ms notice in advance (input length = 6 image frames). If D-Touch predicts three FrA activities categories in the in-the-wild condition: facial T-zone, other activities, and noises, the accuracy was over 72.3% (SD = 5.9%) within 150 ms after the hand showed up in the camera. Although this figure did not come from a real-time intervention study in the wild, it indicates that D-Touch has the great potential to provide just-in-time intervention to stop the undesirable hand-face touching gestures in the in-the-wild condition.

6 DISCUSSION

In this section, we discuss the limitations of the current system and the challenges and opportunities of applying D-Touch at scale for real-world applications.

6.1 Trade-off Between Accuracy and Prediction Time

In terms of evaluation, we found that there is a trade-off between high prediction accuracy and time. In general, the accuracy is higher when the hand is closer to the face. However, it should be noted that in the case of an intervention, an alarm was less useful if the hands are too close to the face. Considering accuracy and prediction time, the intervention should be set up for its purpose. Furthermore, better hardware equipment will make predictions faster. For example, the new type of LeapMotion, high-performance Stereo IR 170 (formerly known as Rigel), has a wider field of view (170× 170) and a longer tracking range (up to 100cm), allowing it to potentially make even faster predictions.

6.2 Potential Applications of a Neck-mounted Wearable

D-Touch is designed to recognize and predict hand-face touching behavior using a neck-mounted wearable device with cameras. We believe the system can have several other applications related to the face such as human activity recognition (HAR), on-body input techniques, facial expression reconstruction, and silent speech recognition. D-Touch showed the possible potential to recognize other 6 facial-related activities (FrA) since it can capture the images when performing the activities. D-Touch could be improved on HAR by training the model with a massive data set having labels on different activities. Also, the D-Touch system can distinguish fine-grained 11 facial areas so that it be employed as a new input technique such as FaceRubbing[40] and EarBuuddy[69], increasing input space from devices to the body [43]. In addition, capturing the partial facial images from the neck can provide enough information to continuously reconstruct facial expressions [9]. Lastly, the neck-mounted wearable camera can be used as silence speech recognition [73] with images of the neck and face from under the chin.

6.3 In-the-wild Scenarios

Although D-Touch demonstrates promising results in both recognition and prediction of HFT-related activities in in-the-wild conditions, we would like to acknowledge that the study was conducted in a relatively limited in-the-wild setting. We expect that, like all existing technologies, the system will encounter challenges in a less controlled in-the-wild setting.

Firstly, the images captured in the in-the-wild scenarios are likely to be noisier with more diverse lighting conditions and backgrounds. In order to see how much this would impact the image quality, we compared the images captured in indoor and outdoor environments with both sensing devices (IR camera and LeapMotion). LeapMotion is likely to be less impacted by strong sunlight. We discuss this in detail in the section 6.4. For more complicated lighting and background conditions, the issue can potentially be eased with image augmentation and a separate model for background removal.

Secondly, the device stability has not been thoroughly tested in a more complicated application scenario. In the in-the-wild conditions, people are more likely to engage in more vigorous activities such as running. This could cause the device to shift significantly. Also, during these types of activities people are more likely to lean forward, thus tilting the camera angle down, potentially capturing more objects in the background. For the first case, apart from directly addressing the device stability issues, it is also possible to integrate an automatic device status detection that reminds the user to adjust the device position and orientation when needed. For the second case, similar techniques used for the complicated background issue can also be applied.

Thirdly, actual activities, both HFT activities, and other HFT-similar activities can have much greater diversity. For example, people may use both their hands to touch their faces, or people may use their hand to swipe across their faces repeatedly, creating much more complicated cases for the system to recognize. This issue can potentially be addressed by increasing the number of training sets, as well as focusing on the most important information needed for intervention.

However, our results for recognizing or predicting HFT-related activities in the in-the-wild scenarios are promising. The objective of D-Touch is to demonstrate the feasibility of recognizing and predicting fine-grained HFT-related activities using a necklace-mounted camera. We leave further optimization in the wild environment for future exploration.

6.4 Various Lighting Conditions Indoors and Outdoors

The user studies were conducted indoors, during the morning, afternoon, and evening, with varying lighting conditions. Although D-Touch performs well indoors, using it outdoors may result in reduced performance due to bright sunlight. The bright background caused by the infrared spectrum in sunlight makes it challenging to segment skin from the background. In order to investigate the influence of sunlight, we compared photos taken with LeapMotion both indoors and outdoors under sunlight shown in Figure 8. We found that LeapMotion is less affected by the sunlight, indicating that LeapMotion has the potential to be used with D-Touch in outdoor environments. However, the IR camera in LeapMotion can be affected by direct sunlight. To overcome this challenge, one solution is to adjust the camera position and angle. By finding the optimal position and angle for the cameras, it may be possible to minimize the impact of direct sunlight and improve performance. For example, pointing the cameras towards the chin and neck region can help reduce the effects of bright sunlight on the captured images, but may compromise prediction performance.

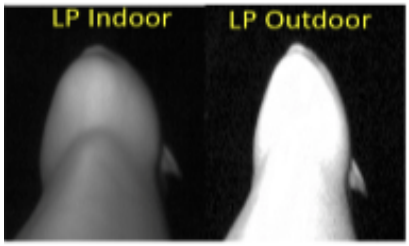

Figure 8: Images captured by Leap Motion camera in indoor and outdoor environment.

6.5 Privacy Concerns

The privacy aspect of wearable camera-based systems is a significant concern while deployed in everyday settings, as they have the potential to capture personal information in different situations. To alleviate this issue, D-Touch has been designed with privacy in mind. The IR camera in D-Touch is positioned at the bottom of the neck and points upward, which restricts the captured information to parts of the ceiling or sky, and the environment in the background is mostly dark, except in the presence of a near-infrared light source. This design, combined with the use of special lighting and filters, could minimize the capture of sensitive information. The captured chin and face from the neck have less privacy information compared to the images captured by frontal cameras. However, we admitted that camera-based systems do capture more environmental information than non-camera wearable systems.

The future work for the D-Touch includes finding ways to remove privacy-related information from the IR images and determining the best method for storing the data, such as feature extraction in real-time.

6.6 Power Consumption

Powering wearable technology has been a persistent challenge in the field. The prototype of D-Touch developed in this paper serves as proof of concept, but is not yet ready for immediately widespread use. The current prototype, which uses LeapMotion, has a relatively high power consumption, meaning its battery is unable to last all day. To overcome this challenge, one possible solution includes optimizing D-Touch with advanced algorithms for image processing and low-power components such as microcontrollers like ESP32 (which consume only 0.79W even with the wireless module activated), low-power cameras (e.g., OV9755 which consumes 100mW at 1280x720@60fps), and reducing the duty cycle of the two LEDs.

6.7 User-independent Model

The evaluation of the classification and prediction in Experiments 1 and 3 was performed using user-dependent models, where the training and testing data came from the same participant. However, this approach has its limitations as it requires new users to spend time collecting data to train the model. To address this, we conducted a leave-one-participant-out evaluation using data from 9 participants to train the models and evaluating them using data from the remaining participant. Each participant had two models for classification (20 input images) and prediction (6 input images), and this process was repeated 10 times so that each participant's data was used once for testing. The results showed that D-Touch could recognize 17 FrA gestures and predict T-Zone vs Other gestures with an average accuracy of 59.2% (SD = 12.3%) and 71.2% (SD = 6.1%) respectively in a controlled lab setting, and 43.4% (SD = 7.3%) for 15 FrA classifications and 56.2% (SD = 5.6%) for prediction in the wild setting. The performance was significantly lower compared to user-dependent models, as the participants' features, including face and hand shapes, hand movement, and neck length, were likely different and led to misclassification of gestures. This suggests that more data is required to develop a general model for new users. We will further explore user-independent models in the future when more data is available.

6.8 Limitations and Future Work

Although our user study results have shown a promising first step towards recognizing and predicting when and where hand-to-face touches occur and how early give the alarm for stopping the behavior, our system still has some limitations that can be improved upon, as is the case with every research prototype.

People may not be willing to wear the device in their daily lives. The main component of the D-touch prototype is a tie-clip device with an IR camera module and an LED mounted on it. An FPC cable is needed to connect the camera module to the laptop. We need to design our device in an aesthetically pleasing way so that people are more likely to wear it. Furthermore, the tie-clip may not work well for all types of clothing. For instance, some shirts may not have a collar to attach the sensor to. Other clothes may partially or fully block the view of the cameras. We plan to also explore other possible form factors around the chest and neck areas in the future to offer more options for users.

The current implementation of D-Touch cannot be immediately applied to commodity wearable devices, as some parts of the hardware are large and not energy efficient. Moving forward, we aim to optimize the hardware design so that the system can fit into one wearable device, eliminating the need to run a wire to another computer. One possible solution is to use the necklace only as the data collection device, which then transmits the data to a cloud server. The server will carry out all the heavy computing work (e.g., machine learning, image processing) and return the prediction results. Additionally, we can also reduce the frame rate of the LED light and camera to prolong battery life.

In Experiment 2, we obtained noteworthy results regarding the timing of D-Touch system alarms to interrupt hand-face touching behaviors. However, it is important to consider that the experiment was conducted in a controlled environment which may have influenced the natural behavior of the participants. For instance, even though we instructed participants to perform hand-face touching gestures in a natural manner, with random alarm timing, the participants might have hesitated or acted abnormally due to the expectation of an alarm. As future work, we plan to conduct a field study with a wireless D-Touch system to gain a more comprehensive understanding of how our system can effectively interrupt hand-face touching in real-life situations.

7 CONCLUSION

In this paper, we present D-Touch, the wearable-based device that can recognize and predict whether and where the hand touches the face. It uses an IR camera to capture the hand activities around the face. These IR images are learned by a customized deep-learning model to recognize and predict hand-face touching behaviors. The user study with 10 participants showed that D-Touch can recognize 17 Facial-related Activity (FrA), i.e., 11 touch positions on the face from other 6 activities with over 92.1% accuracy. Also, D-touch can predict hand-face touching with an average accuracy of 82.12%, 150 ms after the hand shows up in the camera frame. Based on the results, we further discussed the limitations of the current system, as well as the opportunities to deploy the system in real-world applications.

REFERENCES

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. 2006. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence* 28, 12 (2006), 2037–2041.
[2] Jonathan Aigrain, Michel Spodenkiewicz, Séverine Dubuiss, Marcin Detyniecki, David Cohen, and Mohamed Chetouani. 2016. Multimodal stress detection from multiple assessments. *IEEE Transactions on Affective Computing* 9, 4 (2016), 491–506.
[3] Antonis A Argyros and Manolis IA Lourakis. 2004. Real-time tracking of multiple skin-colored objects with a possibly moving camera. In *European Conference on Computer Vision*. Springer, 368–379.
[4] G Bally, J Müller, M Rohs, D Wigdor, and S Kratz. 2012. ShoeSense: a new perspective on hand gestures and wearable applications. In *Proc. CHI*, Vol. 12.
[5] Cigdem Beyan, Matteo Bustreo, Muhammad Shahid, Gian Luca Bailo, Nicolo Carissimi, and Alessio Del Bue. 2020. Analysis of Face-Touching Behavior in Large Scale Social Interaction Dataset. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*.

Association for Computing Machinery, New York, NY, USA, 24–32. https://doi.org/10.1145/3382507.3418876
[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 172–186.
[7] Liwei Chan, Chi-Hao Hsieh, Yi-Ling Chen, Shuo Yang, Da-Yuan Huang, Rong-Hao Liang, and Bing-Yu Chen. 2015. Cyclops: Wearable and single-piece full-body gesture input devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3001–3009.
[8] Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun. 2012. Bayesian face revisited: A joint formulation. In *European conference on computer vision*. Springer, 566–579.
[9] Tuochao Chen, Yaxuan Li, Songyun Tao, Hyunchul Lim, Mose Sakashita, Ruidong Zhang, Francois Guimbretiere, and Cheng Zhang. 2021. NeckFace: Continuously Tracking Full Facial Expressions on Neck-mounted Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–31.
[10] Tuochao Chen, Benjamin Steeper, Kinan Alsheikh, Songyun Tao, François Guimbretière, and Cheng Zhang. 2020. C-Face: Continuously reconstructing facial expressions by deep learning contours of the face with ear-mounted miniature cameras. In *Proceedings of the 33rd annual ACM symposium on user interface software and technology*. 112–125.
[11] Xiang'Anthony' Chen. 2020. FaceOff: Detecting face touching with a wrist-worn accelerometer. *arXiv preprint arXiv:2008.01769* (2020).
[12] Jun Cheng, Can Xie, Wei Bian, and Dacheng Tao. 2012. Feature fusion for 3D hand gesture recognition by learning a shared hidden space. *Pattern Recognition Letters* 33, 4 (2012), 476–484.
[13] Sungman Cho, Minjee Kim, Joonmyeong Choi, Taehyung Kim, Juyoung Park, and Namkug Kim. 2020. Implementation of Face-Touching Action Recognition System based on Deep Learning for Preventing Contagious Diseases. In *Proceedings of the Korean Society of Broadcast Engineers Conference*. The Korean Institute of Broadcast and Media Engineers, 630–633.
[14] Timothy F Cootes and Christopher J Taylor. 1992. Active shape models—'smart snakes'. In *BMVC92*. Springer, 266–275.
[15] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. 1995. Active shape models-their training and application. *Computer vision and image understanding* 61, 1 (1995), 38–59.
[16] Martin Côté, Pierre Payeur, and Gilles Comeau. 2006. Comparative study of adaptive segmentation techniques for gesture analysis in unconstrained environments. In *Proceedings of the 2006 IEEE International Workshop on Imagining Systems and Techniques (IST 2006)*. IEEE, 28–33.
[17] James Crowley, François Berard, Joelle Coutaz, et al. 1995. Finger tracking as an input device for augmented reality. In *International Workshop on Gesture and Face Recognition*. 195–200.
[18] Trevor J Darrell, Irfan A Essa, and Alex P Pentland. 1996. Task-specific gesture analysis in real-time using interpolated views. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18, 12 (1996), 1236–1242.
[19] Martin de La Gorce, David J Fleet, and Nikos Paragios. 2011. Model-based 3d hand pose estimation from monocular video. *IEEE transactions on pattern analysis and machine intelligence* 33, 9 (2011), 1793–1805.
[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.
[21] Nicole D'Aurizio, Tommaso Lisini Baldi, Gianluca Paolocci, and Domenico Prattichizzo. 2020. Preventing Undesired Face-Touches With Wearable Devices and Haptic Feedback. *IEEE Access* 8 (2020), 139033–139043.
[22] Martin Grunwald, Thomas Weiss, Stephanie Mueller, and Lysann Rall. 2014. EEG changes caused by spontaneous facial self-touch may represent emotion regulating processes and working memory maintenance. *brain research* 1557 (2014), 111–126.
[23] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang. 2005. Face recognition using laplacianfaces. *IEEE transactions on pattern analysis and machine intelligence* 27, 3 (2005), 328–340.
[24] Ryosuke Hori, Ryo Hachiuma, Hideo Saito, Mariko Isogawa, and Dan Mikami. 2021. Silhouette-Based Synthetic Data Generation For 3D Human Pose Estimation With A Single Wrist-Mounted 360° Camera. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1304–1308.
[25] Dong-Hyun Hwang, Kohei Aso, Ye Yuan, Kris Kitani, and Hideki Koike. 2020. Monoeye: Multimodal human motion capture system using a single ultra-wide fisheye camera. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 98–111.
[26] Vimal Kakaraparthi, Qijia Shao, Charles J Carver, Tien Pham, Nam Bui, Phuc Nguyen, Xia Zhou, and Tam Vu. 2021. FaceSense: sensing face touch with an ear-worn system. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–27.
[27] Takeo Kanade. 1973. Picture processing by computer complex and recognition of human faces. *Ph. D. Thesis, Kyoto University* (1973).
[28] Michael David Kelly. 1970. *Visual identification of people by computer*. Number 130. Department of Computer Science, Stanford University.
[29] Yen Lee Angela Kwok, Jan Gralton, and Mary-Louise McLaws. 2015. Face touching: a frequent habit that has implications for hand hygiene. *American journal of infection control* 43, 2 (2015), 112–114.
[30] Ivan Laptev and Tony Lindeberg. 2001. Tracking of Multi-state Hand Models Using Particle Filtering and a Hierarchy of Multi-scale Image Features. In *International Conference on Scale-Space Theories in Computer Vision*. Springer, 63–74.
[31] Hyunchul Lim, Yaxuan Li, Matthew Dressa, Fang Hu, Jae Hoon Kim, Ruidong Zhang, and Cheng Zhang. 2022. BodyTrak: Inferring Full-body Poses from Body Silhouettes Using a Miniature Camera on a Wristband. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–21.
[32] H. Lin, M. Hsu, and W. Chen. 2014. Human hand gesture recognition using a convolution neural network. In *2014 IEEE International Conference on Automation Science and Engineering (CASE)*. 1038–1043.
[33] Shang-Hung Lin, Sun-Yuan Kung, and Long-Ji Lin. 1997. Face recognition/detection by probabilistic decision-based neural network. *IEEE transactions on neural networks* 8, 1 (1997), 114–132.
[34] Mona Hosseinkhani Loorak, Wei Zhou, Ha Trinh, Jian Zhao, and Wei Li. 2019. 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* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 32, 12 pages. https://doi.org/10.1145/3338286.3340143
[35] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, Vol. 2. Ieee, 1150–1157.
[36] Tiffany L Lucas, Rachel Mustain, and Robert E Goldsby. 2020. Frequency of face touching with and without a mask in pediatric hematology/oncology health care professionals. *Pediatric Blood & Cancer* 67, 9 (2020), e28593.
[37] Marwa Mahmoud, Tadas Baltrušaitis, and Peter Robinson. 2016. Automatic analysis of naturalistic hand-over-face gestures. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6, 2 (2016), 1–18.
[38] Marwa Mahmoud and Peter Robinson. 2011. Interpreting hand-over-face gestures. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 248–255.
[39] Marwa M Mahmoud, Tadas Baltrušaitis, and Peter Robinson. 2014. Automatic detection of naturalistic hand-over-face gesture descriptors. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 319–326.
[40] Katsutoshi Masai, Yuta Sugiura, and Maki Sugimoto. 2018. 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*. 1–5.
[41] GRS Murthy and RS Jadon. 2009. A review of vision based hand gestures recognition. *International Journal of Information Technology and Knowledge Management* 2, 2 (2009), 405–410.
[42] Mark Nicas and Daniel Best. 2008. A study quantifying the hand-to-face contact rate and its potential application to predicting respiratory tract infection. *Journal of occupational and environmental hygiene* 5, 6 (2008), 347–352.
[43] Aditya Shekhar Nittala, Anusha Withana, Narjes Pourjafarian, and Jürgen Steimle. 2018. Multi-touch skin: A thin and flexible multi-touch sensor for on-skin input. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
[44] Behnaz Nojavanasghari, Charles E Hughes, Tadas Baltrušaitis, and Louis-Philippe Morency. 2017. Hand2face: Automatic synthesis and recognition of hand over face occlusions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 209–215.
[45] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. (2015).
[46] Alex Pentland, Baback Moghaddam, Thad Starner, et al. 1994. View-based and modular eigenspaces for face recognition. (1994).
[47] P Phillips. 1998. Support vector machines applied to face recognition. *Advances in Neural Information Processing Systems* 11 (1998), 803–809.
[48] Laura R Pina, Ernesto Ramirez, and William G Griswold. 2012. Fitbit+: A behavior-based intervention system to reduce sedentary behavior. In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*. IEEE, 175–178.
[49] Juma Rahman, Jubayer Mumin, and Bapon Fakhruddin. 2020. How Frequently Do We Touch Facial T-Zone: A Systematic Review. *Annals of Global Health* 86, 1 (2020).
[50] Siddharth S Rautaray and Anupam Agrawal. 2012. Real time hand gesture recognition system for dynamic applications. *International Journal of UbiComp* 3, 1 (2012), 21.
[51] Siddharth S Rautaray and Anupam Agrawal. 2015. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial intelligence review* 43, 1 (2015), 1–54.
[52] Michael J Reale, Shaun Canavan, Lijun Yin, Kaoning Hu, and Terry Hung. 2011. A multi-gesture interaction system using a 3-D iris disk model for gaze estimation

and an active appearance model for 3-D hand pointing. *IEEE Transactions on multimedia* 13, 3 (2011), 474–486.

[53] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015).

[54] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. 2016. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–11.

[55] Nicola D Ridgers, Melitta A McNarry, and Kelly A Mackintosh. 2016. Feasibility and effectiveness of using wearable activity trackers in youth: a systematic review. *JMIR mHealth and uHealth* 4, 4 (2016), e129.

[56] Hamada Rizk, Tatsuya Amano, Hirozumi Yamaguchi, and Moustafa Youssef. 2022. Smartwatch-based face-touch prediction using deep representational learning. In *Mobile and Ubiquitous Systems: Computing, Networking and Services: 18th EAI International Conference, MobiQuitous 2021, Virtual Event, November 8-11, 2021, Proceedings*. Springer, 493–499.

[57] Enver Sangineto and Marco Cupelli. 2012. Real-time viewpoint-invariant hand localization with cluttered backgrounds. *Image and Vision Computing* 30, 1 (2012), 26–37.

[58] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. 2017. Learning bases of activity for facial expression recognition. *IEEE Transactions on Image Processing* 26, 4 (2017), 1965–1978.

[59] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[60] Lawrence Sirovich and Michael Kirby. 1987. Low-dimensional procedure for the characterization of human faces. *Josa a* 4, 3 (1987), 519–524.

[61] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1891–1898.

[62] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1701–1708.

[63] Orlando Motohiro Tanaka, Robert Willer Farinazzo Vitral, Giulia Yuriko Tanaka, Ariana Pulido Guerrero, and Elisa Souza Camargo. 2008. Nailbiting, or onychophagia: a special habit. *American Journal of Orthodontics and Dentofacial Orthopedics* 134, 2 (2008), 305–308.

[64] Cuong Tran and Mohan Manubhai Trivedi. 2011. 3-D posture and gesture recognition for interactivity in smart spaces. *IEEE Transactions on Industrial Informatics* 8, 1 (2011), 178–187.

[65] Daniel Sáez Trigueros, Li Meng, and Margaret Hartnett. 2018. Face recognition: From traditional to deep learning methods. *arXiv preprint arXiv:1811.00116* (2018).

[66] Hans Van Kuilenburg, Marco Wiering, and Marten Den Uyl. 2005. A model based method for automatic facial expression recognition. In *European conference on machine learning*. Springer, 194–205.

[67] Laurenz Wiskott, Norbert Krüger, N Kuiger, and Christoph Von Der Malsburg. 1997. Face recognition by elastic bunch graph matching. *IEEE Transactions on pattern analysis and machine intelligence* 19, 7 (1997), 775–779.

[68] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. 2008. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence* 31, 2 (2008), 210–227.

[69] Xuhai Xu, Haitian Shi, Xin Yi, Wenjia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K Dey. 2020. Earbuddy: Enabling on-face interaction via wireless earbuds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[70] Liu Yun and Zhang Peng. 2009. An automatic hand gesture recognition system based on Viola-Jones method and SVMs. In *2009 Second International Workshop on Computer Science and Engineering*, Vol. 2. IEEE, 72–76.

[71] Junbo Zhang and Swarun Kumar. 2020. NoFaceContact: stop touching your face with NFC. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*. 468–469.

[72] Ligang Zhang and Dian Tjondronegoro. 2011. Facial expression recognition using facial movement features. *IEEE transactions on affective computing* 2, 4 (2011), 219–229.

[73] Ruidong Zhang, Mingyang Chen, Benjamin Steeper, Yaxuan Li, Zihan Yan, Yizhuo Chen, Songyun Tao, Tuochao Chen, Hyunchul Lim, and Cheng Zhang. 2021. SpeeChin: A Smart Necklace for Silent Speech Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–23.

[74] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. 2018. Single-shot refinement neural network for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4203–4212.

[75] Shibo Zhang, Yuqi Zhao, Dzung Tri Nguyen, Runsheng Xu, Sougata Sen, Josiah Hester, and Nabil Alshurafa. 2020. NeckSense: A Multi-Sensor Necklace for Detecting Eating Activities in Free-Living Conditions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 72 (June 2020), 26 pages. https://doi.org/10.1145/3397313