NeckFace: Continuously Tracking Full Facial Expressions on Neck-mounted Wearables

TUOCHAO CHEN*, Cornell University and Peking University
YAXUAN LI*, Cornell University and McGill University
SONGYUN TAO*, Cornell University
HYUNCHUL LIM, Cornell University
MOSE SAKASHITA, Cornell University
RUIDONG ZHANG, Cornell University
FRANCOIS GUIMBRETIERE, Cornell University
CHENG ZHANG, Cornell University

Facial expressions are highly informative for computers to understand and interpret a person's mental and physical activities. However, continuously tracking facial expressions, especially when the user is in motion, is challenging. This paper presents NeckFace, a wearable sensing technology that can continuously track the full facial expressions using a neck-piece embedded with infrared (IR) cameras. A customized deep learning pipeline called NeckNet based on Resnet34 is developed to learn the captured infrared (IR) images of the chin and face and output 52 parameters representing the facial expressions. We demonstrated NeckFace on two common neck-mounted form factors: a necklace and a neckband (e.g., neck-mounted headphones), which was evaluated in a user study with 13 participants. The study results showed that NeckFace worked well when the participants were sitting, walking, or after remounting the device. We discuss the challenges and opportunities of using NeckFace in real-world applications.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile devices**.

Additional Key Words and Phrases: Facial expressions, Deep learning, Infrared Imaging, Wearable

ACM Reference Format:
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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 58 (June 2021), 31 pages. https://doi.org/10.1145/3463511

1 INTRODUCTION

Facial expressions are highly related to a persons' emotions and activities, e.g., eating. Therefore, to recognize human activities precisely, computing systems need to track and interpret facial expressions continuously.

*First, second, and third authors contributed equally to the paper.

Authors' addresses: Tuochao Chen, 1600012713@pku.edu.cn, Cornell University and Peking University; Yaxuan Li, yaxuan.li@mail.mcgill.ca, Cornell University and McGill University; Songyun Tao, st938@cornell.edu, Cornell University; Hyunchul Lim, hl2365@cornell.edu, Cornell University; Mose Sakashita, ms3522@cornell.edu, Cornell University; Ruidong Zhang, rz379@cornell.edu, Cornell University; Francois Guimbretiere, francois@cs.cornell.edu, Cornell University; Cheng Zhang, chengzhang@cornell.edu, Cornell University, 244 Gates Hall, Ithaca, New York, 14853.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Association for Computing Machinery.
2474-9567/2021/6-ART58 $15.00
https://doi.org/10.1145/3463511

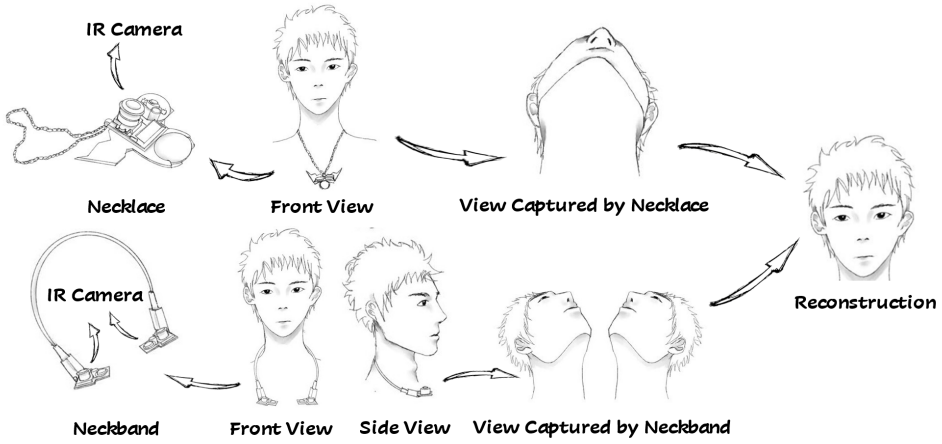

Fig. 1. Overview of Neck-Face

For instance, a person's facial expressions can be used to understand a person's emotions [32, 44]. Facial expressions can also be applied to recognize activities that involve facial movements. For example, researchers have demonstrated that computers can interpret what a person speaks by just analyzing the facial expressions (mostly the movements of lips) extracted from the images of the face, which is also known as silent speech [16, 33]. Furthermore, facial expressions can also help recognize activities with massive facial movements such as eating, drinking, and smoking [12]. To recognize these human activities with facial expressions, the first step is to continuously and accurately track the full facial expressions in daily activities.

In order to recognize and track facial movements, researchers have developed a variety of intelligent sensing systems. The most popular technique utilizes computer vision (CV) techniques to reconstruct facial expressions, analyzing the user's face captured by a frontal camera. This method has provided promising performance[17, 52] in scenarios where the camera can capture the face without any occlusion. However, this method's setup requirements are demanding in real-world scenarios. It significantly limits the user's movement range by requiring the user to stay in front of a camera without collusion, resulting too constraining for continuously monitoring every day activities. Furthermore, it may not work well if the user is in motion or outdoor, where it is hard to set up a frontal camera.

To address the above challenges left by traditional CV-based methods, the wearable research community has developed portable systems to recognize facial expressions in mobile settings using different sensing methods such as Electrical Impedance Tomography (EIT) [43], Electromyography (EMG) [21, 49, 50], and ultrasound [35]. Unfortunately, most of them can only recognize discrete facial expressions instead of continuously tracking the full facial expressions and require obtrusive instrumentation on the user's body (e.g., attaching electrodes on the face [57]), making them impractical to be worn daily. There is a need for a minimally-obtrusive wearable sensing method that can continuously track the full facial expressions. One recently published research project, C-Face [7], has demonstrated a new direction on wearable-based facial expression tracking system. It enables an ear-mounted device by embedding a pair of miniature RGB cameras into both sides of earphones to capture the contours of the face (e.g., earphones). These images of contours are used to train a deep learning model, which could estimate the full facial expressions represented by 42 facial landmarks. However, the intellectual space for tracking facial movements with wearables is still largely open. The users deserve a complementary set of wearables that they can choose from to track facial movements depending on the context. For instance,

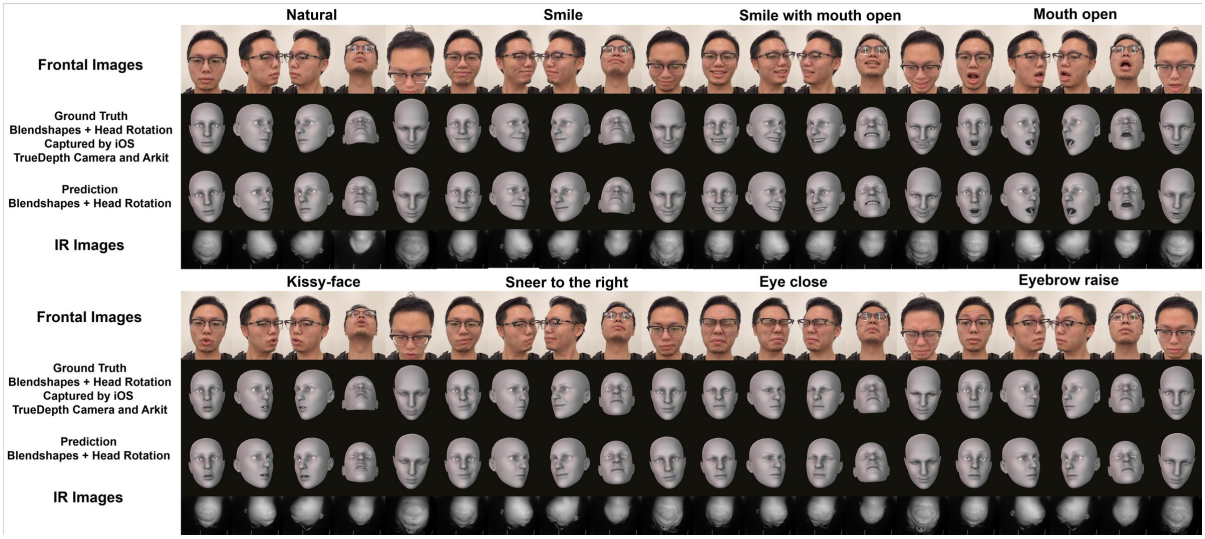

Fig. 2. Facial expressions with head rotations used in NeckFace. 3D mesh model© Apple

to continuously track the facial movements, the users have to always wear the device. Many people may feel uncomfortable wearing an earphone or headphones through daily activities, which may obstruct hearing. Thus, providing a rich set of alternative wearable technologies is critical to satisfying diverse needs from users.

To address the above challenges, we present NeckFace, the first neck-mounted wearable sensing technology that can continuously track the full facial expressions by using infrared(IR) cameras to capture the chin and face beneath the neck. One thing worth to mention is that the face captured from the neck only contains partial info on the nose, cheeks, mouth (no eyes or eyebrows). Traditional CV-based methods (e.g., Dlib) can not directly extract facial expressions using the face's partial/incomplete information. Inspired by C-Face, we believe that by stitching this incomplete information on various facial components together, we could reconstruct the full facial expressions, including the movements of the eyes, eyebrows, cheek, mouth, nose, and chin. To validate this research hypothesis, we developed NeckFace on two common neck-mounted form factors: the necklace and neckband (e.g., neck-mounted headphones), as shown in Figure 1. To ensure high-quality ground-truth data on facial expressions, we adopted a new ground-truth capturing system using the TrueDepth camera on the iPhone X and ARKit API [1]. This technology allows us to collect detailed facial expressions on all facial components, including the cheek, mouth, eyes, eyebrows, chin, and nose, and the three-dimensional head rotation angles. It represents each facial expression with 52 blend shapes, each controlled by a parameter ranging from 0 to 1000 (after multiplying a scale factor of 1000 on the original parameter(0-1)). NeckFace uses a customized deep learning model NeckNet to predict the parameters of these 52 blend shapes and the angles of the head's three-dimensional rotation by learning the IR images of the bottom of the face captured from the neck. To evaluate our systems, we conducted a user study with 13 participants under two real-world scenarios: 1) when the user performs eight kinds of facial expression when the user is sitting while rotating the head, and 2) walking.

To evaluate our systems, we conducted a user user study with 13 participants. Each participant was asked to perform eight facial expressions while sitting and walking as shown in Figure 2. In the sitting scenarios, the participants were also asked to rotate the head while performing the facial expressions, and remounted the device

[1] https://developer.apple.com/augmented-reality/

in one session. The average Mean Absolute Error (MAE) of 52 blendshape parameters (range from 0 to 1000) on necklace and neckband under these three scenarios (sitting without remounting, sitting with remounting, and walking) are 30.293, 34.166, 25.359 and 25.612, 28.418, 22.635 respectively. Furthermore, NeckFace can estimate the three-dimensional rotational angles (roll, yaw, pitch) of the head with an average error of 3.554, 4.336, 2.456 degrees for a necklace, and 3.146, 3.649, 2.584 degrees for a neckband.

The contribution of this paper is:

- We present the first neck-mounted wearable system that can continuously track the full facial expressions. It contains a customized data processing and deep learning pipeline which can estimate the full facial expressions represented by 52 blendshape parameters and three-dimensional head rotations extracted by TruthDepth camera [2] on iPhone and ARkit.
- We evaluated the system on two prevalent neck-worn form factors: neckband and necklace, under three conditions: when the user is static, walking, and remounts the device.
- We discuss the limitations, challenges, and opportunities of NeckFace on applying them in real-world devices and scenarios. And a discussion on the challenges.

2 RELATED WORK

NeckFace is a neck-mounted wearable device that can continuously reconstruct the facial expressions. Therefore, in this section, we define related work as: 1) neck-mounted wearables; 2) non-wearable methods to recognize facial expressions; 3) wearable-based methods to recognize facial expressions. We also highlight the key contributions of NeckFace comparing to prior works.

2.1 Neck-mounted Wearable Devices

We chose neck-mounted piece as the form factor as people are used to wearing neck-piece in daily activities. For example, people wear a necklace or neckband(e.g., Bluetooth speakers). Furthermore, the neck is closer to head, which is a good position to observe and sense human activities without instrumenting heads. Researchers have deployed electrodes [57] or ultrasonic sensors [35] to recognize weak or silent speeches. Neck-mounted sensors also have been used to estimate head posture [34], and recognizing eating activities through capacitive/pressure sensing[10] or jawbone movements[13]. Several other neckbands have been developed throughout the years to detect chewing and swallowing activities [1, 11, 20, 31]. Similar to NeckFace, [9] also adopts an infrared camera around the neck. But it only recognizes the carotid pulse and breathing rate.

Based on our knowledge, we have not identified any neck-mounted device that can continuously track the facial expressions, which clearly distinguishes NeckFace from prior neck-mounted sensing systems.

2.2 Estimating Facial Expressions Using Non-wearables

The most popular method for tracking facial expressions uses cameras (E.g., RGB cameras [53, 58, 67], RGB-D cameras [26, 62], thermal cameras [23]) in front of the user, to capture the entire face. These face images are sent and processed by computer vision algorithms to extract or estimate the facial movements. Traditional computer vision methods estimate the facial movements based on the prior knowledge of the facial structures and analysis of the images [15, 48, 64, 70]. As deep learning has proven to be effective in many ML tasks, a variety of deep learning-based methods[14, 23, 28, 30, 37, 39, 40, 51, 54, 59, 68, 69] have also shown promising performance on tracking facial expressions. More importantly, most of these CV-based methods can work in a user-independent manner, which means the user does not need to provide any training data before using it. Thus, a mature public library can extract facial expressions using RGB cameras [36] or a depth camera (e.g., TrueDepth Camera).

[2] https://support.apple.com/en-us/HT208108

However, despite the impressive performances, these systems do not work well in scenarios where the camera can not be set up or capture the user's face. This drawback has limited the applications of tracking the user's facial expressions in mobile settings. Moreover, the frontal camera always introduces privacy concerns as images may include sensitive private information on the user or the surrounding environment. Especially the methods using frontal cameras could capture the full face of the user, which is highly sensitive private information. NeckFace has a camera with an IR filter, which captures the head from the neck. It significantly reduces the risk of privacy leakage. First, it does not capture the entire face. It is much harder to identify a person from the IR images captured by NeckFace, as shown in Figure 2. Furthermore, it will not capture the surrounding environment, as they will show in black in the IR images. In summary, compared to placing cameras in front of the face, a neck-mounted wearable is less obtrusive, hands-free, and more privacy-aware.

2.3 Facial Movement Reconstruction with Wearable Sensors

Researchers in the wearable community have developed many wearable sensing devices to recognize facial expressions, such that they work without limiting the user's range of movements. Most of these wearable systems are attaching sensors or electrodes on different parts of the head to capture the corresponding signals that are related to facial movements, including acoustic sensors[29], piezoelectric sensors[38, 56], capacitive sensors[49], magnetic sensors[19, 55], barometers[3], electromyography (EMG) signals [21, 49, 50], etc. Unfortunately, most of these systems require heavy instrumentation on the human body (e.g., attaching electrodes on the body). Furthermore, none of them can continuously track the full facial expressions as they recognize a list of discrete facial gestures.

To continuously track the facial expressions in VR, researchers have put cameras on and into VR headset, such that they can capture the critical facial components such as eyes or mouth[5, 24, 63]. However, putting a camera in front of the face is not feasible in daily activities. The latest work C-Face, is the only wearable device that can continuously track facial movements without directly seeing the face(e.g., eyes, mouth)[7]. It attached RGB cameras on earphones and headphones. Using Deep Learning to map the contour of the chin to the pose of the face demonstrated promising facial-movement tracking performance.

As we have discussed in the introduction, NeckFace, as a neck-mounted wearable, was not designed to compete with C-Face, as an ear-mounted wearable. They provide a complementary set of wearables that the users can choose from to track facial movements depending on the context. As a form factor, many people are comfortable wearing a necklace (NeckFace) throughout the day, while not many would keep an earphone(C-Face) on during the day. Offering users with different wearable options are important in the future when wearables are ubiquitous. Beyond the key differences above, NeckFace also presents significant advantages over C-Face on tracking performance and system implementation.

Firstly, the ground-truth resolution on facial expressions is relatively low, which uses 42 landmarks extracted using Dlib library [36] representing the mouth, eyes, and eyebrows. Many complex and extreme facial movements can not be extracted from RGB images of the face using Dlib. For instance, C-Face/D-Lib can not track the cheek's movements, an essential component for facial expressions. NeckFace adopts a more advanced facial expression representation system, which uses 52 land shapes. It can track much detailed and subtle facial movements on all facial components, including eyes, eyebrows, mouth, nose, cheeks, and chin. Secondly, it is hard to segment the human skin from different backgrounds in RGB images, which is one of C-Face's limitations. NeckFace address this issue by using IR cameras, which segmenting the human body from backgrounds more accurately and reliably in IR images. Thirdly, C-Face requires two RGB cameras, while NeckFace can work well with only one camera. Having half the number of cameras not only reduces the energy consumption but also half the size of data. A smaller size of data would make the data transmission and process more efficient, which is another important metric for real-world deployments. Lastly, NeckFace can capture the facial expression and head orientation (in 3D)

simultaneously, which is critical to VR applications. However, additional sensors such as IMU must be deployed to estimate the head pose for the wearables on other positions like ear-mounted or head-mounted.

Overall, NeckFace is the first neck-mounted wearable technology that can continuously track the full facial movements without the need of capturing the entire face.

3 THEORY OF OPERATION

The intuitive sensing principle behind NeckFace is that the images of the head captured from the neck can include incomplete shapes about the neck, chin, cheek, mouth, and nose, as Figure 3 shows. This information is not enough to directly extract facial expressions using traditional CV-based methods (e.g., Dlib library), as none of them captures each facial component's complete picture. However, in our preliminary experiment, we notice that the shapes about the neck, chin, cheek, mouth, and nose captured from the neck-mounted camera present visually distinguishable patterns when the user performs different facial expressions in Figure 3(a),(b),(c),(d) or head rotations in Figure 3(d)(e). This finding encourages us to design and implement a hardware prototype and a machine learning pipeline to realize the complex connections between the shapes of the head (the neck, chin, cheek, nose, and mouth) observed from neck and the full facial expressions.

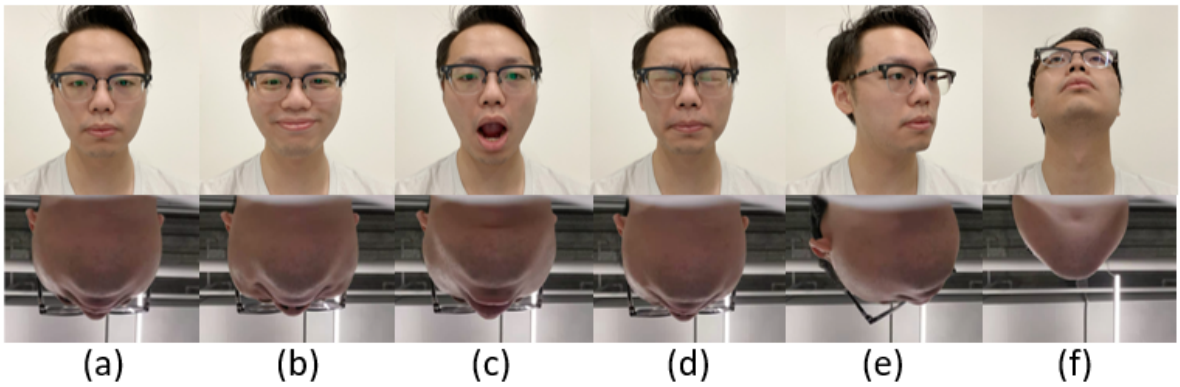

Fig. 3. Some examples of the images from bottom view during different facial expressions and head rotation. (Face movements: (a). Natural. (b). Smile. (c). Mouth open. (d). Eyes close. (e). Head rotates left. (f). Head rotates up.)

4 SYSTEM DESIGN AND IMPLEMENTATION

The initial observation from the preliminary experiment is encouraging. However, some challenges need to be addressed while designing such a neck-mounted facial expression sensing system. In this section, we present the detailed design and implementation of the hardware and algorithm on NeckFace.

4.1 Hardware Design

The first step of designing the hardware is to choose an appropriate sensing device, which: 1) captures detailed information on the shape of the head (including the neck, chin, cheek, nose and mouth) from the neck; 2) provides data that can easily segment the head from different backgrounds; 3) can be embedded into a minimally obtrusive neck-mounted form factor.

We initially used an RGB camera to capture the face's images from the neck. However, we soon realized it is challenging to segment the head from different RGB images backgrounds reliably. Therefore, we look into other

alternative sensing solutions, such as depth cameras, thermal cameras, and proximity sensor arrays that make segmentation easier. However, some of them fail to meet other design considerations we listed above. Despite that depth cameras can provide a high-resolution RGBD image for easy segmentation and depth information, all depth cameras available in the market are bulky and not designed for sensing in a short distance between neck to head (under 10cm).

Considering our ideal sensing device properties, we came up with three candidates: IR, Thermal, and RGB camera. After conducting a pilot study (detailed written in the later section) that compared the three cameras' different results, we discovered that the IR camera met all the requirement and outperformed the thermal and RGB camera; thus, we chose to use an active infrared (IR) sensing approach, which segments the head from the background while providing high-resolution data.

Specifically, our active IR sensing hardware is composed of three parts: OV5647 camera (sensible to infrared 850nm spectrum, FoV 130 degree, 30 FPS, adjustable focus, and auto exposure), IR narrow band-pass filter (850 nm), and NIR (Near-Infrared) LED (TY-850nm3W, light angle 120 degrees) as shown in Figure 4(a). In our active sensing method, the 850nm NIR LED projects the bright IR light on the bottom of the chin with around $9mW/cm^2$ intensity, which is safe for human skin according to the previous research [4]. Then the IR band-pass filter placed on the lens of IR cameras would filter out most of the visible light such that the IR camera can only capture the infrared light mostly reflected by the skin. In this way, the human's chin is much brighter than the background (regardless of the content in the background) in the captured IR images, as shown in Figure 4, which makes it much easier and be more robust to segment the head from the complex backgrounds. Also, since most of the background is invisible to our IR camera, it can better protect the surrounding environment's privacy. We connect the cameras to a Raspberry Pi 4 board to read the IR camera images through a CSI interface. Then the Raspberry Pi uploads the images captured by cameras to the server through WiFi.

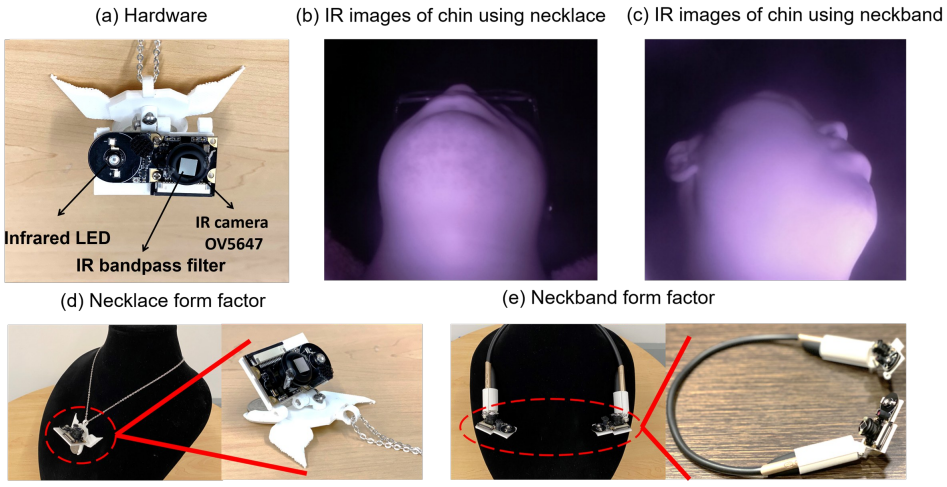

Fig. 4. Hardware and from factor design of NeckFace

4.2 Form Factor Design

After choosing the appropriate sensing device(IR camera), the next step is to design an appropriate form factor to house the IR cameras. There are three factors we consider while designing this neck-mounted form factor. The device should: 1) maintain a relatively stable position to capture the entire chin shape when the users are either

walking or rotating their heads; 2) have an acceptable size and weight which can be worn for daily use; 3) have good practicality and portability.

We design two form factors for NeckFace: a necklace and a neckband since they are popular daily worn accessories(e.g., neckband: neck-mounted Bluetooth speaker).

As shown in Figure 4 (d), the necklace form factor has three major components: 1) A long light-weight silver chain to hang our device; 2) A 3D printed necklace base offers a small hoop for the chain to go through; 3) A 3D printed camera holder to fix for IR cameras.

It is worth mentioning that the necklace base's design went through several iterations to find a balance between size, shape, and stability. When wearing a necklace, a complete smooth plane between the back of the necklace and the cloth could lead to severe swinging; thus, we put an angle between two wings and the necklace (17 degrees) to cancel out the shift of center of mass and provide stability (the human's chest is not flat). Since the 3D printed component's weight is lighter than the camera, we used a quarter dollar coin to balance the mass center.

As shown in Figure 4(e), the neckband form factor has three major components: 1) A flexible neckband blue tooth earphone; 2) A pair of 3D-printed cases fixed on the neckband; 3) A pair of 3D-printed camera holders with a pair of 3D-Printed joints (same as a necklace) connected to the cases. The camera could rotate horizontally and vertically around the 3D-printed case.

4.3 Ground-truth Acquisition Method

We use the TrueDepth camera on iPhone (available on iPhone X or later models) and ARKit offered by Apple to extract the full facial expressions as the ground-truth. Compared to the traditional CV library on extracting facial expressions (e.g., Dlib), this ground-truth acquisition system can extract high-quality and fine-grained details of facial expressions/movements on the eyes, eyebrows, cheeks, mouth, nose, and chin. It extracts facial features, such as facial geometry and rotation, from a depth map created by the TrueDepth camera with 30,000 invisible dots. We chose to use the smartphone as a ground-truth method instead of using larger depth-sensing equipment such as Kinect or RealSense since the ground-truth system can be easily set up and worn without specific hardware requirements.

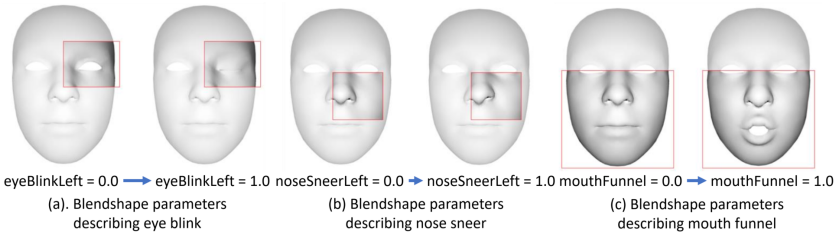

Fig. 5. Three examples of blendshape parameters describing eye blink(a), nose sneer(b) and mouth funnel(c). (The figure is screenshotted from © Apple[3])

Apple ARKit offers the pre-defined 52 blendshapes[3] through which one can achieve complex facial animations. The 52 blendshapes consist of seven features for both left and right eyes, twenty-seven features for mouth and jaw movements, ten features for eyebrows, cheeks, nose, and tongue. For each blend shape parameter, the value ranges from zero to one where the value one represents a specific animation's maximum state. For example, as shown in Figure 5, the image demonstrates three key features and its neutral configuration (0.0) to the maximum

[3]https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation

movement (1.0): eyeBlinkLeft, noseSneerLeft, and mouthFunnel. We multiply the parameters by one-thousand for calculation. Also, we recorded the head's orientation in Euler angles (roll, yaw, pitch) that ARKit provides.

To enable ground-truth collection when the user is in motion, we deploy a wearable supporter that could hold the iPhone so the user's face could be captured, as shown in Figure 6.

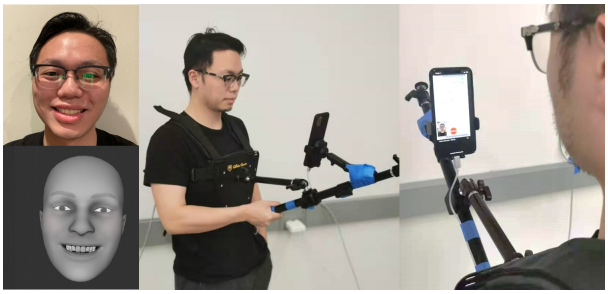

Fig. 6. Left: Camera Captured image and its facial expression reconstruction by ARKit, Left mid: Mobile ground-truth data collection settings, Right mid: User perspective of the mobile settings, Right: iOS data collection application. 3D mesh model© Apple

4.4 Data Processing Pipeline for NeckFace

This section presents NeckNet: a CNN-based deep learning processing pipeline (NeckNet), which learns the IR images of chin and face captured from the neck to estimate full facial expressions and three-dimensional head rotation. We chose a convolutional neural network (CNN) because it has shown excellent performance dealing with 2D image tasks like classification, retrieval, and segmentation compared with other traditional ML algorithms [25]. Another important reason on why approach deep learning model as the first attempt to address this problem is that the mapping between the IR images of the face observed from the neck and the full facial movements are much less intuitive compared to directly extracting facial movements from the images of the face. It is even challenging for human eyes to directly find the connection. Therefore, we think a more complex ML model such as deep learning would better handle the problem and find the hidden connections between the IR images of the neck and chin with the full facial movements. Overall, we believe CNN is a good candidate for our task, which involves 2D images of the human body. Furthermore, even with CNN which is hard to deploy directly on a wearable computing platform, we shared a similar solution with other wearable devices, which use the wearables as the sensing and data collection unit, and defer all heavy computation work (e.g., ML training and processing) to another more powerful computers in the cloud. We also plan to further explore other ML models with lower-requirement for resources in the future.

4.4.1 Overview of Proposed Pipeline. Figure 7 presents the deep neural network pipeline to estimate the blendshape parameters p_{exp} and head orientation R_{head} from IR images of chin and face. Specifically, we firstly preprocess the raw IR images with the CV-based algorithm and then feed the processed images to a deep neural network called NeckFace to simultaneously reconstruct facial expression (represented by 52 blendshape parameters) and head rotations).

4.4.2 Preprocessing. The preprocessing step is designed to reduce the influence of different types of noises brought into the IR image (e.g., images shifting caused by the camera's motion, image background noises caused by other bright light sources). The preprocessing can also help the neural networks generalize better and more robust when the images are captured in different settings, backgrounds, and scenarios. Our preprocessing process

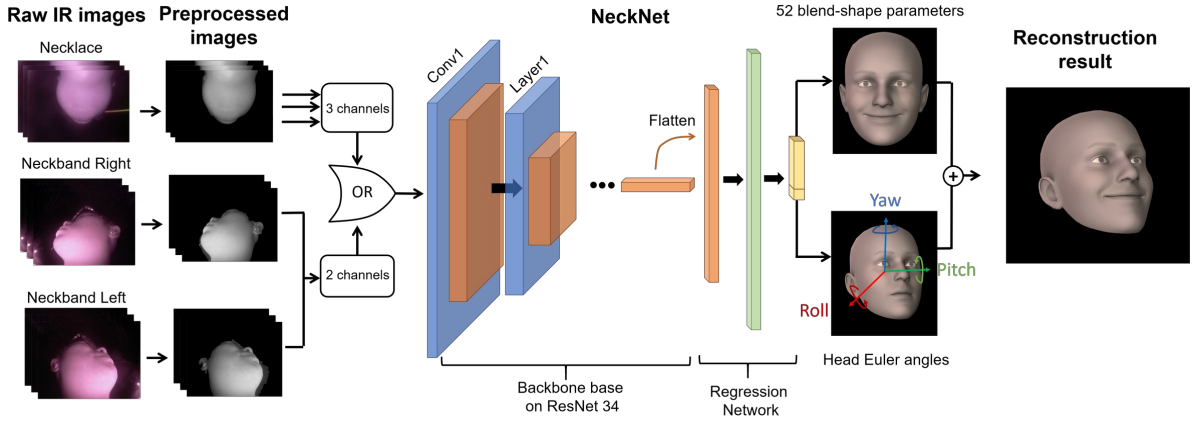

Fig. 7. Deep neural network pipeline for Neckface. 3D mesh model© Apple

is composed of three steps: color conversion, background segmentation, and data augmentation, as shown in Figure 8.

First, we convert the raw IR images into gray-scale images to eliminate potential color variance. Since the IR filter only allows monochrome light into the camera, the color in the images carries no actual meaning, while only the brightness carries the valid information. Then, we separate the head from the backgrounds in IR images. Since we adopt infrared sensing method as mentioned in section 4.1, the human skin (neck, chin and face) in the IR image is much brighter than the background. However, in practice, some light sources containing IR light in the background would still introduce some noises in the background as shown in the red circle in the Figure 8. In order to remove these noises, we set a brightness threshold (threshold of brightness is 50) to binarize the gray-scale image, then keep the maximum connected region and remove all other regions (light sources), as shown in Figure 8. This solution works well in an indoor environment; however, this may be problematic if the sunlight is too strong (e.g., when the user is in a sunny outdoor environment). We will discuss this issue and propose our verified solutions in the discussion section 8.2.

Lastly, we apply data augmentation to make the system more robust to motion noise. In a real-world setting, the camera's position and angle can not be constants, as human activities can easily shift the position of the cameras. For instance, if a user is walking, the camera's position would shifts slightly around the neck. Or if a user takes off the camera and puts back on again, the camera's positions and angles may not be exactly the same compared to the last worn position. Different camera positions would introduce significant different on view angle and areas in the images for the same facial expressions. Even we have designed the form factor to increase the stability of the form factor while wearing. The small shifts on camera positions is inevitable during our daily activities. One straightforward way is to have the user to collect training data in each of these conditions (e.g., walking, remounting), which is time-consuming and influencing user experience. To mitigate this issue without adding burdens to the user, we use data augmentation to synthesize the training data sets under these conditions. More specifically, we set a probability of 60% to conduct three types of image transformations, which can be caused by camera shifting: translation, rotation, and scaling on the images. We deploy 3 Gaussian Models to generate the parameters for the translation($\mu = 0, \sigma^2 = 30$), rotation($\mu = 0, \sigma^2 = 10$), scaling($\mu = 1, \sigma^2 = 0.2$) on the synthesized training data. We will run the data augmentation on all the images in the training dataset during each training epoch before feeding images into deep learning. Data augmentation can improve our deep learning model's ability to confront camera shifting and avoid over-fitting during model training.

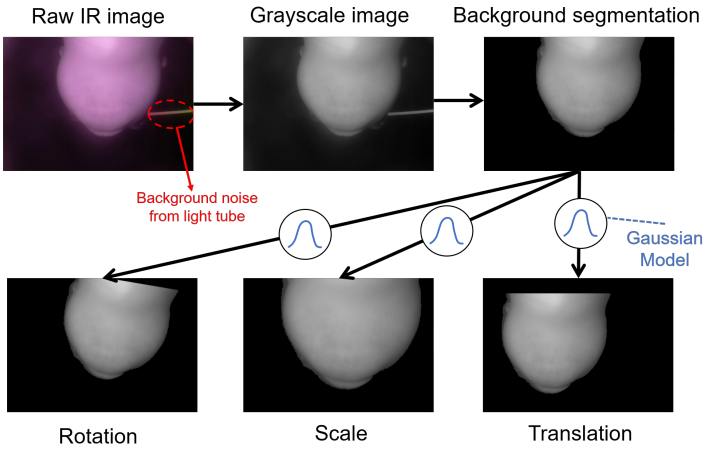

Fig. 8. Preprocessing for IR images

4.4.3 NeckNet Architecture. After preprocessing, the necklace and neckband generate one and two streams of grey-scale pictures, respectively. Before we send them directly to the deep learning model, we duplicate the input channels for necklace data because research [42] has demonstrated that duplicating grey-scale images to three channels can improve the expressiveness and the ability to extract features of the model. Therefore, we duplicate the grey-scale images of necklace three times to three channels and set the input channel of the first convolutional layer as 3. We do not apply similar duplication on neckband, as it already has two channels.

Since Residual learning [22] has shown promising performance on image recognition, we design the structure of NeckNet structure based on a Residual Network (ResNet34)[22] as the backbone, which extracts global features from the input IR image as shown in Figure 7. In ResNet block, there is a batch normalization layer [22] and a rectified linear unit (ReLU) following each convolutional layer. At the end of the 34-layer residual blocks, we deploy an adaptive average pooling layer to obtain a feature vector representation of each image. After that, a regression network, composed of two full-connected layers and a dropout layer [61] (p = 0.5) between them, is placed behind the pooling layer. The last fully-connected layer comprises two parts: 52 blendshape parameters p_{exp} and three Euler angles of head R_{head}. Once we retrieved the estimated parameters on facial expressions and rotations, we used a program in Unity to visualize a 3D mesh model of the human head, as shown in Figure 7.

4.4.4 NeckNet Loss Function. In NeckNet, we chose to use the Huber loss function [27] in our regression layer, as shown in Equation (2). The biggest challenge we encountered when designing the loss function is that the data set is imbalanced between inactive frames (natural facial expressions), where the user does not perform facial expressions, and active frames, where the user is performing facial expressions. During data collection, the samples of one particular facial expression are usually much smaller than the samples of the natural facial expressions. In other words, the active frames of each facial expression are usually fewer than the inactive frames. Therefore, if we directly apply Huber loss function and assign equal weights on all frames, the model would tend to predict facial expressions closer to the natural facial expression to lower the total cost. Intuitively, it functions as an average filter on a data stream, where the active frames are the peaks, and the inactive frames the flat curve. If given equal weights, the predicted curve after applying the average filter would be flat. In other words, the model would not take the risk and predict facial expressions with larger movements. To address this issue, we introduce a customized loss function based on Huber Loss function by assigning higher weights of the active

frames and lower weights to the inactive frames to address this issue. We first classify all the frames into inactive frames and active frames using parameter threshold (more details in subsection 4.5). Then we assign different weights (w_a for active frames and w_n for inactive frames) to their loss function according the proportion of two types of frames. Specifically, $w_a = K \times N_n/(N_a + N_n)$ and $w_n = 0.5 \times N_a/(N_a + N_n)$, where K is the type number of facial expressions except natural expression, and 0.5 is the coefficient selected by experiment, N_a is the number of active frames including natural facial expression, N_n is the number of inactive frames including other K facial expressions. The following equation describes the details of our customized weight loss function.

$$L_B = \frac{1}{B}\sum_{i=1}^{B} w_n \cdot I_n(\hat{p}_i)(L_{Huber}(p_i, \hat{p}_i) + L_{Huber}(R_i, \hat{R}_i)) + \frac{1}{B}\sum_{i=1}^{B} w_a \cdot I_a(\hat{p}_i)(L_{Huber}(p_i, \hat{p}_i) + L_{Huber}(R_i, \hat{R}_i)) \quad (1)$$

$$L_{Huber}(y, \hat{y}) = \begin{cases} \frac{1}{2}[y - \hat{y}]^2 & \text{for } |y - \hat{y}| \le 1, \\ |y - \hat{y}| - \frac{1}{2} & \text{otherwise,} \end{cases} \quad (2)$$

where L_B is the loss function for one batch data with batch size B. The p_i and R_i is the prediction result of blendshape parameters and head rotation of frame i, while $\hat{p}_i$ and $\hat{R}_i$ is the ground-truth of frame i. I_a and I_n are indicator functions which specify whether the frame i is active or inactive, and w_a and w_n are the weights for the loss of active frames or inactive frames respectively.

4.4.5 NeckNet Training. We deploy and train the NeckNet with Pytorch [46]. During training, we use the standard mini-batch stochastic gradient descent (SGD) optimizer with momentum (0.9) and weight decay (1e-4) [41]. We set the initial learning rate as 0.01 with Cosine learning rate annealing. For all experiments, our model is trained for 80 epochs. During each training epoch, before fed into NeckNet, the full training dataset would be randomly shuffled and divided into several batches (batch size = 30). Finally, after training, the NeckNet would be evaluated on each frame from the hold-out non-overlapping testing dataset.

4.5 Evaluation Metrics

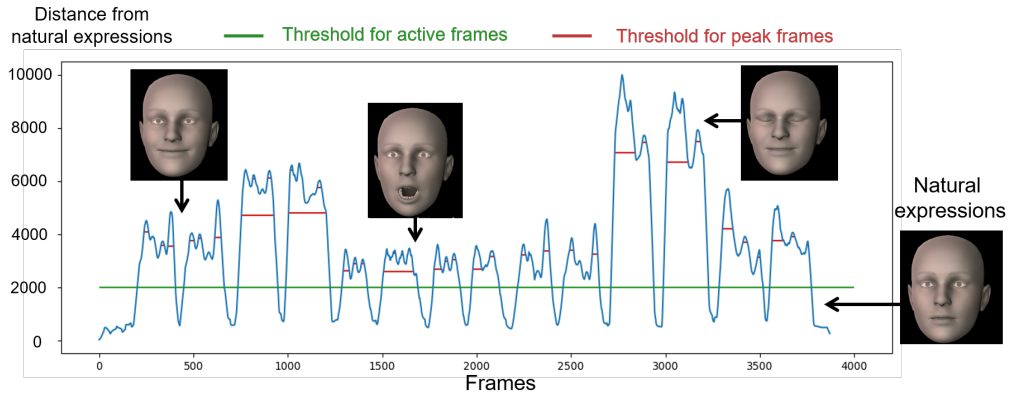

Fig. 9. The illustration of active frames and peak frames. 3D mesh model© Apple

We use Mean Absolute Error (MAE) to evaluate our result, which is widely used in evaluating the performance of deep learning models. Specifically, we calculate the MAE of 52 parameters between ground-truth (p) and our

prediction $\hat{p}$ across N frames as shown in the equation below:

$$MAE_{parameter} = \frac{1}{N}\sum_{1}^{N}\frac{1}{52}||p - \hat{p}||_1$$

The MAE of three rotation angles between our prediction (R) and ground-truth ($\hat{R}$) across N frames is calculated using the equation below:

$$MAE_{angle} = \frac{1}{N}\sum_{1}^{N}\frac{1}{3}||R - \hat{R}||_1.$$

MAE is the average error across all the frames which include a lot of the frames with natural expressions. As the reason we discussed above, it may not best represent the true performance of the model. Therefore, besides MAE of all frames, we propose two other metrics: active MAE and peak MAE, which are representing the performance on active frames and peak frames.

Active frames mean the frames when people conduct large facial expressions rather than natural facial expressions. In order to determine whether one frame is active or inactive, we use the ground-truth data to calculate the difference of the 52 parameters' value between this frame and the inactive frame with natural expressions (the blue curve in Figure9). Then we derive a threshold (2000, the green line in Figure 9). When the difference between the current frame and the inactive frame is larger than the threshold, it is an active frame; otherwise, it is inactive frames. The peak frames are the frames when the degree of one expression reaches around the peak maximum. Specifically, we utilized the peak seeking algorithm (min-height = 2, prominence=0.3, width=0.1, relative height=0.3) to find the peaks and regard the frames on the peak as the peak frames. Based on active frames and peak frames, we calculate active MAE and peak MAE, where the active MAE calculates the MAE of the active frames only, while the peak MAE is the MAE across the peak frames. Even though the active MAE and peak MAE are usually larger than MAE, we believe presenting these results would help the readers to comprehensive understand the true performance of our system.

MAE is a number, which is hard to interpret in terms of how different it would reflect on the reconstructed facial expressions. Therefore, we generate a list of facial expressions using Unity, with different MAE ranging from 20 to 140, as shown in Figure 10. As the Figure 10 shows, when the MAE is under 40, the predicted facial expressions is visually highly similar to the ground-truth facial expressions. When MAE is between 40 to 80, the prediction and ground-truth is close with visually noticeable difference. When MAE is larger than 80, it is hard to link the ground-truth and predicted facial expressions as the same expression.

5 PILOT STUDY

Before we conducted the full user study, we conducted a pilot study on three co-authors to examine the hardware configurations and camera positions on different form factors. The pilot study has two parts: 1) We compared the performance among different cameras (IR, thermal and RGB), so that we can make an informative decision on camera types;2) We evaluated different camera positions on two form factors to understand how would the position of cameras impact the performance of NeckFace.

5.1 Comparisons between Different Types of Cameras

We conducted a preliminary pilot study to select a suitable camera focusing on different spectral domains. More specifically we considered:

- an IR camera (OV5647: adjustable focus, auto exposure, 640x480 pixels, 30 FPS, FOV of 130 degrees) fitted with an LED illuminator;

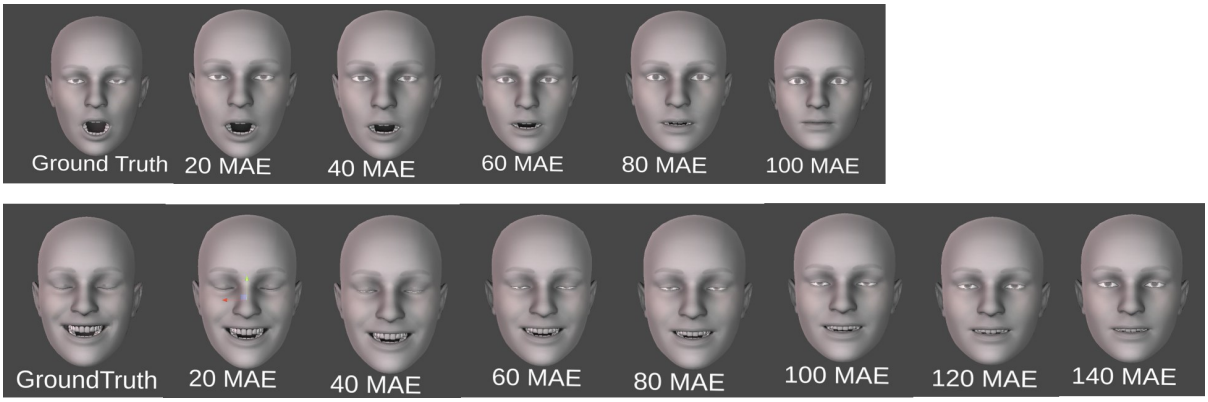

Fig. 10. The visualization of two types of facial expressions with different MAE. 3D mesh model© Apple



Fig. 11. Camera's position on neckband in pilot study. The result of wider neckband beats the narrow neckband; the lower necklace indicates a better result. 3D mesh model© TurboSquid

- a small RGB camera (OV5647 with IR cut-off filter, a fixed focus, auto exposure, 640x480 pixels, 30 FPS, FOV of 120 degrees);
- a thermal camera (mlx90640: fixed focus, 32x24 pixels, 30 FPS, FOV of 90 degrees);

When testing each camera, each researcher repeated five facial expressions (smile, open mouth, kissy face, eye close, eyebrow raise) with head rotation from right to left slowly for six times (around $60{\times}30FPS{\times}3.5min = 6300frames$ for each camera). To collect testing data, for each camera, each researcher repeated these five expressions twice (total around $60 \times 30FPS \times 1.12min = 2100frames$ for each camera). We adjusted the lighting condition in the data collection process to simulate real-world scenarios. Three cameras were placed at the same location during this process. We found that the IR camera was the best with an average MAE of 26.28, followed by the thermal camera (51.928) and the RGB camera (63.697). This led us to select the IR camera for the rest of the evaluation.

5.2 Comparison between Different Hardware Settings

It is to be expected that people might wear the neckband in different shapes and choose different lengths for their necklace chain, leading to the camera position change. These changes could significantly influence the performance of our system. We hypothesis that:

- Wider neckband (within a reasonable range) may have better reconstruction results since the IR cameras placed on the end could grasp more information about the cheek motion around the eye and mouth;

- Lower necklace position may have better reconstruction results since the IR camera would have a wider field of view, improving recognition during head rotation. A lower setting will also benefit the overall scene illumination;
- The further the camera is from the chest the better the recognition in part because less of the chest and more off the face are visible to the camera.

We present our settings and results in 11. Note that for all settings, each researcher repeated the same five expressions (smile, open mouth, kissy face, eye close, eyebrow raise up) with same head rotation from right to left six times (6300 frames) for training data collection, and two times for testing (2100 frames).

As shown in Figure 11 right, the effect of distance to the chin met our expectations; however, the performance did not improve significantly when we make the camera away from the chest. It might be that we did not move the camera far enough, yet one has to consider that in this position, the necklace seems very obtrusive and not practical. We will continue our evaluation with the best setting found here.

With respect to the neckband, Figure 11 left shows that the wider setting is better as expected. We will continue our evaluation with the best setting found here.

6 USER STUDY

In order to evaluate how would the NeckFace perform when the user is sitting (with head rotation), after remounting the device, and when the user is walking, we recruited 13 participants (5 females) with an average age of 25. 5 participants have long hair, 2 participants have a beard, and 8 participants wear glasses. The user study was conducted in a large and bright room, with six windows letting sunlight and several light tubes on the ceiling. Our study was conducted across different times of the day, including morning, afternoon, and evening.

We used the TrueDepth camera on iPhone 10 Pro Max to collect the ground-truth data of the participant's facial expressions. The iPhone was mounted on the user using a chest mount, as shown in 6, which help point the camera steadily towards the face. Besides collecting ground-truth of facial expressions, iPhone was also used to play the demo video, which instructs the users on how and when to perform which facial expressions. We used an Ethernet cable to communicate the data from the iPhone to our server. The resolution and frame rate of our IR cameras was 640x480 and 30 FPS. These cameras are connected to a Raspberry Pi, which communicates to the server using WiFi.

6.1 User Study Procedure

At the beginning of the user study, one researcher first helped the participant to wear our ground-truth capturing device (TrueDepth camera on iPhone) on the chest and adjust the angle accordingly, as shown in Figure 6. The researcher then assisted participants in wearing our prototype (necklace or neckband) and adjusting the prototype at a fixed position according to the results from our pilot study. When data collection began, there is no further adjustment to the device position. In the user study, the participants were asked to perform eight facial expressions(natural, smile, smile with teeth, mouth open, kissy-face, sneer to the right, eye close, eyebrows raise, shown in Figure 2) under two scenarios: 1) when the user is sitting, 2) when the user is walking.

6.1.1 The Sitting Scenario. In the first part of the study, the participants sit on a chair while wearing both the chest-mounted iPhone (ground-truth) and our prototypes(necklace and neckband). The participants were asked to perform the aforementioned eight facial expressions following the demonstration video shown on the iPhone. To simulate the real-world scenarios, where the user may rotate the head while performing the facial expressions, we asked the participants to move the head in four directions (up, down, right, left) following the demo video. To perform each facial expression, the participants first mimicked the facial expression and then moved the head to one of the four directions while maintaining that facial expression. The covering angles of rotation are about 80°

along yaw direction, 78° along pitch direction, and 12° along roll direction. Performing eight facial expressions in one head direction once took approximately 3 minutes. This part can be divided into three sessions:

In the first session, we collected the training data. Each participant performed eight facial expressions six times in each of the four head direction for each form factor (in a total of 2), which resulted in 192 ($8 \times 6 \times 4$) training facial expression samples for each form factor. The dataset collected in these sessions contains around 12600 frames(60s/min × 30FPS × 7min).

In the second session, we collected testing data without remounting the device. After the participant completed the first session, they were asked to take a break and then continue this second session. In this session, each participant performed eight facial expressions twice in each of the 4 head directions for each form factor (in a total of 2), which resulted in 64 ($8 \times 2 \times 4$) testing facial expression samples for each form factor. The dataset collected in these sessions contains around 4500 frames(60s/min × 30FPS × 2.5min).

In the third session, we collected testing data after remounting the device. Remounting devices can cause a position shift on the cameras, which may impact the system's performance. Therefore, after session two, we asked the participants to remove the form factor and put the form factor back. Then the participant was asked to perform exactly the same procedure as the second session. Each participant performed 8 facial expressions twice in each of the 4 head directions for each form factor (in a total of 2), which resulted in 64 ($8 \times 2 \times 4$) testing facial expression samples after remounting for each form factor. The dataset collected in these sessions contains around 4500 frames(60s/min × 30FPS × 2.5min).

6.1.2 The Walking Scenario. After testing our system's performance in the first scenario where participants were sitting on a chair, we evaluated them in a more complex and dynamic situation - walking in the second scenario.

In this part of the study, the hardware setting remained the same. The only difference is that the participants were asked to carry a portable battery in a bag to supply power to NeckFace hardware. We marked large-sized glowing arrows on the floor to guide participants walking to easily find the direction using the corner of the eyes without the need to look down at the floor. The participants were instructed to keep their head naturally forward while walking. This section can be divided into two sections: training and testing.

In the first session, the participant provided training data while keep walking along the marked path. Each participant performed 8 facial expressions 6 times in one head direction (forward) for each form factor(in a total of 2), which resulted in 48 ($8 \times 6 \times 1$) training facial expression samples for each form factor. The dataset collected in these sessions totally contains around 7200 frames (60s/min × 30FPS × 4min).

In the second session, the participant provided testing data while keep walking along the marked path. Each participant performed 8 facial expressions twice in one head direction (forward) for each form factor(in a total of 2), which resulted in 16 ($8 \times 2 \times 1$) training facial expression samples for each form factor. The dataset collected in these sessions totally contains around 2340 frames (60s/min × 30FPS × 1.3min).

7 THE RESULTS ON ESTIMATING THE FACIAL EXPRESSIONS AND HEAD ROTATION ANGLES

7.1 Result for Sitting Scenario without Remounting the Device

We used the training data collected in the first training session in the sitting scenario to train our model, and used the data from the second session in the same scenario (testing data without remounting) as the testing data. The average MAE over thirteen participants across all frames on the data collected from testing sessions is 30.293($SD = 6.336$) for necklace and the 25.612($SD = 5.069$) for the neckband, as shown in Table. 1 and Figure 13. The range of the ground-truth and prediction results for blendshape parameters is 0 to 1000 as we multiply the raw parameters with 1000 as mentioned in 4.3. We also calculated the active MAE (MAE of frames excluding natural facial expressions) and the peak MAE (MAE of frames when the degree of one expression reaches around the peak maximum). The average active MAE and peak MAE are 34.857($SD = 6.891$) and 40.416($SD = 9.294$) for the necklace, and 29.651($SD = 5.875$) and 34.337($SD = 8.986$) for the neckband. The performance on P2 was the

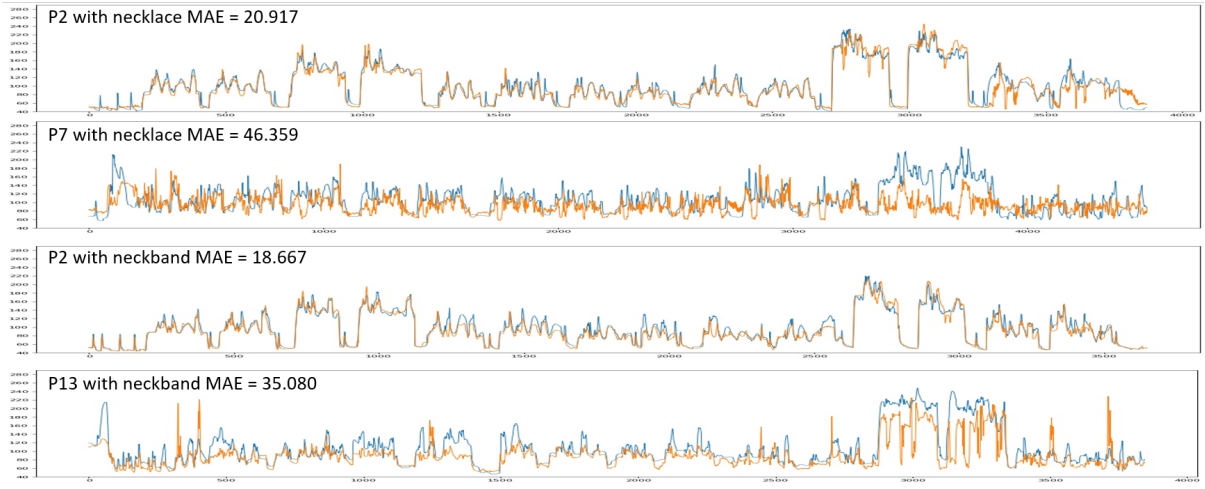

Fig. 12. Ground-truth and prediction blendshape MAE of participants wearing necklace and neckband with highest and lowest performance in sitting scenario without remounting. (Blue line is ground-truth and yellow line is predicted result)

best with an MAE of 20.917 with the necklace; the performance on P7 was the worst with an MAE of 46.359. The reason for the bad performance on P7 with the necklace is that the long hair blocks the camera seriously in this session of the experiment. As we have presented in Figure 10, an MAE under 40 indicates the predicted facial expressions are highly similar to the ground-truth facial expressions acquired using the TrueDepth camera. Therefore, we think NeckFace provided promising facial expressions tracking performance and worked very well if the device was not remounted.

To provide the readers with more comprehensive details on the performance of NeckFace throughout the testing session, we plotted the sum of values on 52 parameters from both the ground-truth and our predication for participants wearing necklace and neckband with highest and lowest performance, as shown in Figure 12. The yellow line is the values of predicted parameters, and the blue line is the values of the parameters acquired by ground-truth.

As for estimating the rotation angles of the head, the average MAEs were $3.258°(SD = 1.414°)$, $2.574°(SD = 0.900°)$ and $4.829°(SD = 1.614°)$ in yaw, pitch, and roll for necklace. The average MAEs for neckband were $2.641°(SD = 0.570°)$, $2.474°(SD = 0.459°)$ and $4.322°(SD = 0.861°)$ MAE in yaw, pitch and roll.

To compare the necklace and neckband performance, we conduct a one-way ANOVA study on the MAE of blendshape parameter reconstruction between two form factors. We found a statistically significant effect ($F(1,24)=4.32$ and $p=0.048<0.05$). We also conduct the one-way ANOVA study on the MAE of head rotation reconstruction between two form factors and find no significant effect ($F(1,24)=1.38$ and $p=0.25>0.05$). The results of ANOVA analysis indicate that neckband has a statistically better performance than necklaces as for the facial expression reconstruction. One possible reason is that the neckband used two cameras covering a wider area of the chin and face, which helped achieve better performance. As for head rotation reconstruction, necklace and neckband performance do not have a significant difference.

7.2 Result for Sitting Scenario after Remounting the Device

We used the training data collected in the first training session in the sitting scenario to train the model and used the data from the third session in the sitting scenario (testing data collected after remounting) as the testing data.

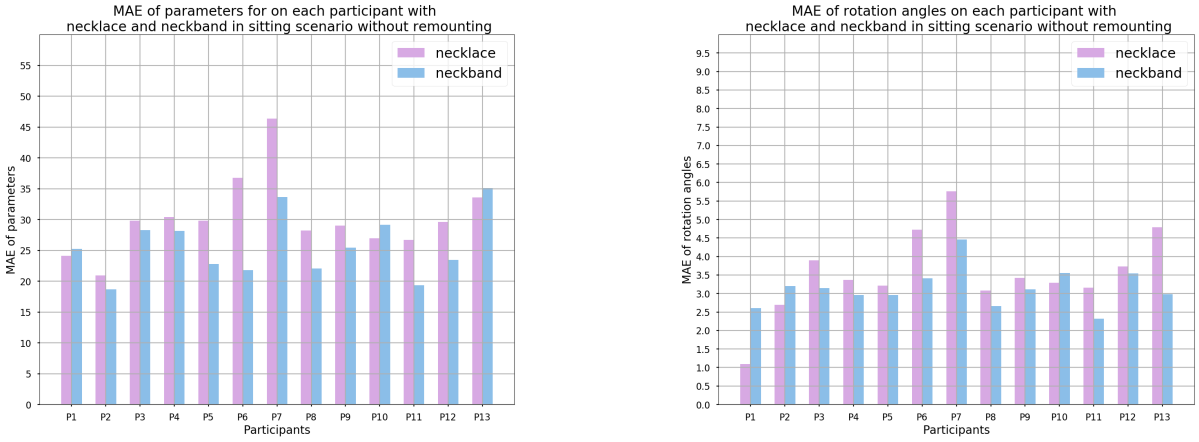

Fig. 13. MAE of parameters and rotation angles on each participant with necklace and neckband in sitting scenario without remounting

Table 1. Average results of thirteen participants in sitting scenario without remounting (The results are shown with [Mean | Standard Deviation])

Scenario	Blendshape / Form factor	MAE	active MAE	peak MAE
Sitting scenario without remounting	Necklace	30.293 | 6.336	34.857 | 6.891	40.416 | 9.294
	Neckband	25.612 | 5.069	29.651 |5.875	34.337 | 8.986
	Angle (°) / Form factor	Yaw MAE	Pitch MAE	Roll MAE
	Necklace	3.258 | 1.414	2.574 | 0.900	4.829 | 1.614
	Neckband	2.641 | 0.570	2.474 | 0.459	4.322 | 0.861

As for the facial expression reconstruction, the average MAE for necklace and neckband over 13 participants is 34.166 (SD = 9.518, max = 57.598 for P7, min = 23.802 for P2) and 28.418 (SD = 7.859, max = 49.245 for P3, min = 18.427 for P11) respectively as shown in Figure 14. Although there was a slight performance drop compared to the previous session where the testing data was collected without remounting, all MAEs are still under 40, which indicates our predicted results after remounting was still highly similar to the ground-truth. Regarding the head rotation angle estimation, the average MAE over three angles increase from 3.554° to 4.336° with the necklace and from 3.146° to 3.649° with neckband as shown in Figure 14.

We also ran a two-way ANOVA test to study possible effects between remounting and form factors. We found the main effect on form factors ($F(1,48) = 6.48$ and $p = 0.014$), which indicates that the type of form factor has a significant effect on facial expression reconstruction performance. Meanwhile, we did not find a main effect on the performance before remounting and after remounting ($F(1,48) = 2.65$ and $p = 0.11$), meaning that our system's performance does not significantly differ before remounting and after remounting. There were no interaction effect between these two variables ($F(1,48) = 0.068$ and $p = 0.79$). Also, we conducted the two-way ANOVA on the MAE of head rotation reconstruction on two variables: remounting and form factors. The result also shows that neither of remounting ($F(1,48) = 3.07$ and $p = 0.086$) and form factors ($F(1,48) = 2.22$ and $p=0.142$) have significant

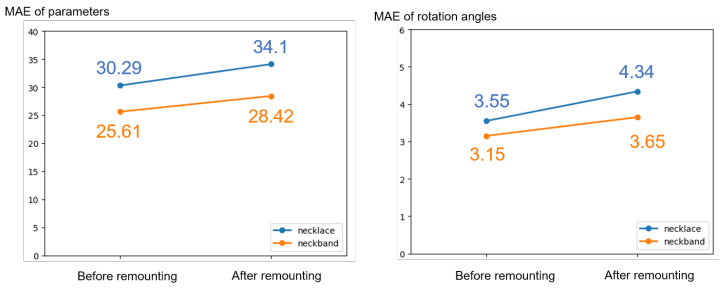

Fig. 14. The change of the MAE of parameters and rotation head before remounting and after remounting

effect on the head rotation reconstruction result. We found no interaction effect between these two variables ($F(1,48) = 0.145$ and $p = 0.7$).

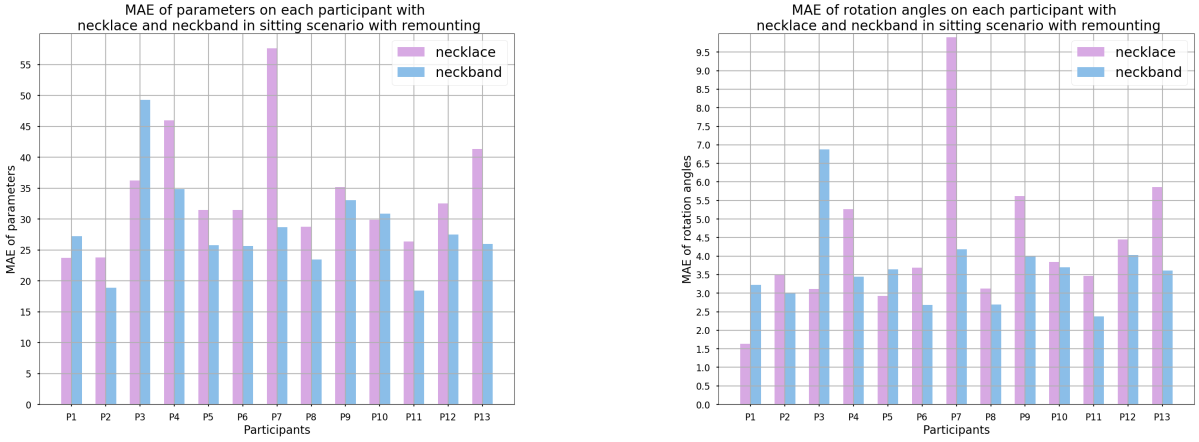

Fig. 15. MAE of parameters and rotation angles on each participant with necklace and neckband in sitting scenario with remounting

Table 2. Average results of thirteen participants in sitting scenario with remounting (The results are shown with [Mean | Standard Deviation])

Scenario	Blendshape / Form factor	MAE	active MAE	peak MAE
Sitting scenario with remounting	Necklace	34.166 | 9.518	40.025 | 11.119	46.194 | 13.559
	Neckband	28.418 | 7.859	32.535 | 8.051	35.999 | 9.597
	Angle (°) / Form factor	Yaw MAE	Pitch MAE	Roll MAE
	Necklace	4.733 | 4.156	3.368 | 1.464	4.908 | 1.929
	Neckband	3.415 | 1.721	3.061 | 1.228	4.472 | 1.051

Table 3. Average results of thirteen participants in walking scenario (The results are shown with [Mean | Standard Deviation])

Scenario	Blendshape / Form factor	MAE	active MAE	peak MAE
Walking scenario	Necklace	25.359 | 6.351	30.123 | 7.693	33.291 | 11.863
	Neckband	22.635 |5.196	29.109 |7.660	30.735 | 9.550
	Angle (°) / Form factor	Yaw MAE	Pitch MAE	Roll MAE
	Necklace	1.906 | 0.691	1.860 | 0.686	3.604 | 0.858
	Neckband	1.667 | 0.495	1.544 | 0.340	4.541 | 0.743

7.3 Result for Walking Experiment

We used the training data collected in the training session from the walking scenario to train our model, and used the data from the second session (testing session) in the same scenario as the testing data to evaluate the system performance. The detailed results were shown in Table 3. Our system achieved an average MAE of 25.359(SD = 6.351) for necklace and 22.635 (SD = 5.196) for neckband on estimating facial expressions. From the experiment results, the necklace on P2 performs best with 13.963 MAE and worst on P5 with 36.480 MAE of parameters; the neckband on P11 performs best with 14.790 MAE and worst on P9 with 32.166 MAE of parameters.

We conducted one-way ANOVA on the MAE of facial expressions and head rotation angles between two different form factors. As for the facial expression reconstruction ($F(1,24) = 1.43$ and $p=0.24$), the form factors did not significantly affect either, which was different from the sitting scenario without remounting. As for head rotation reconstruction, the form factors did not have a significant effect ($F(1,24) = 0.87$ and $p =0.358$), which was similar to result of the sitting scenario without remounting.

The walking scenario results are very encouraging and even surprising, as it worked better than when the participants were sitting. One possible explanation is that we did not ask participants to rotate the hand while walking, making the reconstruction task easier than the sitting scenario, where the participant was instructed to rotate the head in four directions. We will further discuss this in 8.2.

7.4 Summary

According to the results, we found that the neckband generally performs better than the necklace in all three studies. One possible explanation is that the cameras on the neckband could capture more information from both sides. Additionally, the neckband is more stable as it fits the body closely as it is larger, which provides more support and stability.

In our user study, there are eight users wearing glasses (P3, P5, P6, P7, P8, P10, P11, and P12). The blendshape MAE is 31.761 with necklace and 25.054 with neckband in the sitting scenario without remounting. The result of participants with glasses does not have an apparent decrease. Therefore, we conclude that glasses will not have an obvious impact on our results.

Furthermore, we found P7 as the outlier in our study. After reviewing the data, we found that the long hair of P7 blocked a significant portion of the camera's view for both necklace and neckband in the experiment. The recorded images had little information on the chin and neck, which causes our system to fail. This is one limitation of NeckFace, that the system does not work well if the camera is blocked by cloth or hairs.

8 DISCUSSION

In this section, we discuss the practical challenges and opportunities for applying NeckFace on practical applications based on the study results.

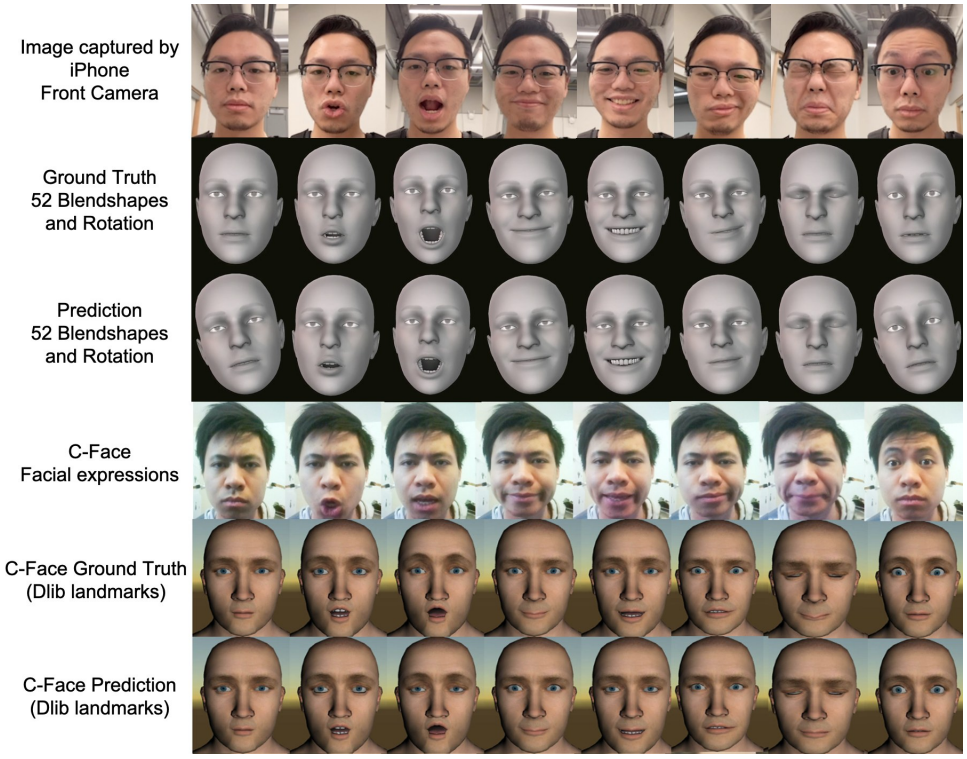

Fig. 16. Comparison between C-Face and NeckFace. Row 2,3: 3D mesh model© Apple. Row 5,6: 3D mesh model© TurboSquid

8.1 Performance Comparison between C-Face and NeckFace

As we have discussed in the introduction, C-Face is the other wearable device that can continuously track facial expressions. In this section, we want to compare the performance between C-Face and NeckFace. It is hard to compare the MAE between the two papers directly, and the ground-truth acquisition method is different. C-Face uses 42 landmarks extracted from Dlib, while NeckFace uses 52 Blendshape parameters extracted from TrueDepth camera by ARKit.

We decided to compare the performance by directly putting the predicted facial expressions from C-Face and NeckFace side by side as shown in Figure 16, as the facial expressions in NeckFace overlap with the facial expressions used in C-Face such that the readers can directly interpret the visual difference between the results. We contacted the authors of C-Face, who granted us permission to use C-Face figures in our paper. The C-Face reconstructs the face with unity from those 42 landmarks on the eyes, mouth, and nose. While NeckFace has 52 parameters and three angles to fully reconstruct the facial movement and head rotation in detail. These parameters are able to capture the subtle changes of not only mouth and eyes but also face parts like cheek and eye-brown, which can not be represented by the 48 landmarks. As Figure 16 shows, the predicted facial expressions using NeckFace are visually better or closer to the real facial expressions, especially when the facial expressions require more movements. For example, in the sixth column from the left (sneer to the right), the reconstructed facial expression using NeckFace is significantly more vivid than C-Face.

There are many reasons attributed to this: Firstly, NeckFace uses a much better quality of ground-truth, which can capture the entire facial movements including eyes, eyebrows, cheeks, mouth, nose, chins. DLib used by

C-Face can only capture eyes and eyebrows, and mouth. Secondly, NeckFace used a customized loss function to assign different weights to frames, designed to secure better performance when the facial movement is significant. Thirdly, the camera on the neck captures more information, such as chin, cheeks, nose, mouth, neck. C-Face only captures incomplete side contours from the ear. The grey-scale images used in NeckFace contain 2D information, while C-Face only has 1D information (contour line).

Based on our experience, NeckFace can apparently represent a much richer set of subtle facial movements in higher accuracy than C-Face. However, this is based on our own subjective perception. To acquire more substantial proof of the performance difference, we plan to conduct qualitative studies to have participants compare the estimated facial expressions by the two techniques in the future.

Beyond these differences, it is essential to remember that these two systems have very different form factors to satisfy the diverse needs of the users in the future. The goal of this paper(NeckFace) is not to beat the performance on C-Face. Instead, we vision C-Face and NeckFace are complementary to each other. They together can form an ecosystem on the wearable-based facial tracking system, which offers the users different options in a variety of contexts.

8.2 Outdoor Scenario and Hardware Improvements

Like any camera-based system, sunlight's inference is a potential issue for our system, as sunlight contains the 850nm spectrum light, which may interfere with the reflected IR lights in the captured IR images. The top left of Figure 17 displays the IR images of chin and face captured using NeckFace under the indoor scenario and outdoor scenario with strong sunlight. Due to sunlight's impact, the background in IR images outdoors is brighter than indoors, making it more challenging to segment the human body in the IR images.

One possible solution is to use the more advanced IR light source and the IR filter on the camera to address this issue, such that the overlap on the light spectrum between the sunlight and project IR light source is minimal. We tried different IR cameras to explore this option and found that the IR camera in Leap Motion can effectively filter out the natural light even in an outdoor environment.

The sensing principle of Leap Motion is very similar to our methods: both project the pattern-less IR light on the object and use the IR camera to capture the IR images. However, Leap Motion used more advanced hardware (two monochromatic IR cameras and three brighter IR LEDs) and a better exposure strategy and algorithm to better handle the sunlight issues.

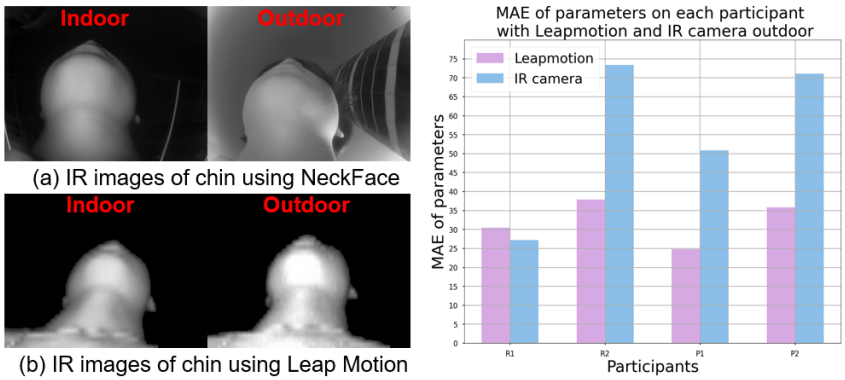

Fig. 17. Left: The IR images of chin indoors and outdoors using NeckFace and Leap Motion. Right: The blendshape MAE of 2 participants and 2 researchers with Leapmotion and our IR camera

To compare the facial expression reconstruction performance using Leap Motion and our NeckFace in the outdoor scenario, we conducted a preliminary study with 4 participants (two were coauthors). We only chose five facial expressions (smile, open mouth, kissy face, eye close, eyebrow raise). The procedure is similar to the previous study; these participants repeated the five facial expressions six times indoors (6300 frames) as the training data and then repeated twice outdoors as testing data (2100 frames). In Figure 17 right, the MAE for NeckFace and LeapMotion is 55.6 and 31.8, respectively. Per Figure 10, the results showed that the current NeckFace camera did not work well in the outdoor environment while leap motion provided a promising direction. Therefore, although the current hardware set in NeckFace may not immediately work well in an outdoor environment, one possible solution is to adopt advanced filters and IR light source to filter out the natural light in the background.

Another possible solution is to implement a pattern in the IR light source. We are inspired by the FaceID system from the iPhone, which provides Neckface with another optional solution to sunlight issues. Currently, our IR LED justly projects the infrared light on our chin, while the structured light module in iPhone can project the IR pattern encoded with spatial and temporal information. Such a carefully designed pattern can help segment the object that the pattern projected from the background. Besides, these patterns can also provide extra information, like depth, leading to even better performance. With the development of Microelectromechanical systems (MEMS) technology, we believe that the smaller IR projector has the potential to be widely used in wearable devices in the future. We plan to explore this in the next step.

8.3 The Influence of Head Rotation on Facial Expression Reconstruction

In the sitting scenario, participants were asked to move the head in four directions while performing the facial expressions. To explore the impact of different head rotation angles on the performance, we did another analysis. We only analyzed the direction on pitch and yaw as they are the major moving direction during the study.

We first calculated the average of blendshape parameter MAE across 13 participants at different head rotation angles (represented by a 2D coordinate (Yaw, Pitch) as shown in Figure 18(a)). We plotted two heatmaps to visualize the MAE distribution for the necklace and neckband in Figure 18(a),(b) respectively. In these figures, the brighter color means higher MAE and worse performance. If the figures' area is black, it means no participants reached this angle combination on yaw and pitch in the study.

As we can observe, the higher MAE tends to be distributed on the area with higher absolute value on Yaw and Pitch, especially with large values on Yaw (rotate head to the right and left). Besides, if we compare the MAE distribution of the necklace and neckband, it is visually apparent that the MAE of the necklace increases much higher when the absolute value of yaw increased compared to the neckband.

To explain the results, we further analyzed the raw IR image of three conditions using necklace and neckband: 1) head pose kept in the middle, 2) Pitch is large 3) Yaw is large as shown in Figure 18(c). These pictures show that when users rotate their head to the right, left, or up, some part of chin, mouth, and cheek becomes invisible to the IR camera. Such information loss tends to decrease the facial expression reconstruction performance, especially when the absolute value of yaw increases. For the necklace, since it only has one camera directly below the chin, more parts of the cheek and mouth would be blocked during head rotation. However, when it comes to the neckband, because of the two IR cameras, it covers a larger area of the chin and face, even when the user rotates their head to one side. Therefore, the neckband is more robust in estimating facial expressions when the head moves to extreme positions.

8.4 Facial Reconstruction with Different Skin Tone

Given that our system is based on active IR sensing, the different surface and skin colors may impact how the IR lights are reflected on the body. In our current user study, we did not systematically evaluate this issue. However, we found multiple previous research studies that investigated how IR lights with different wavelengths would

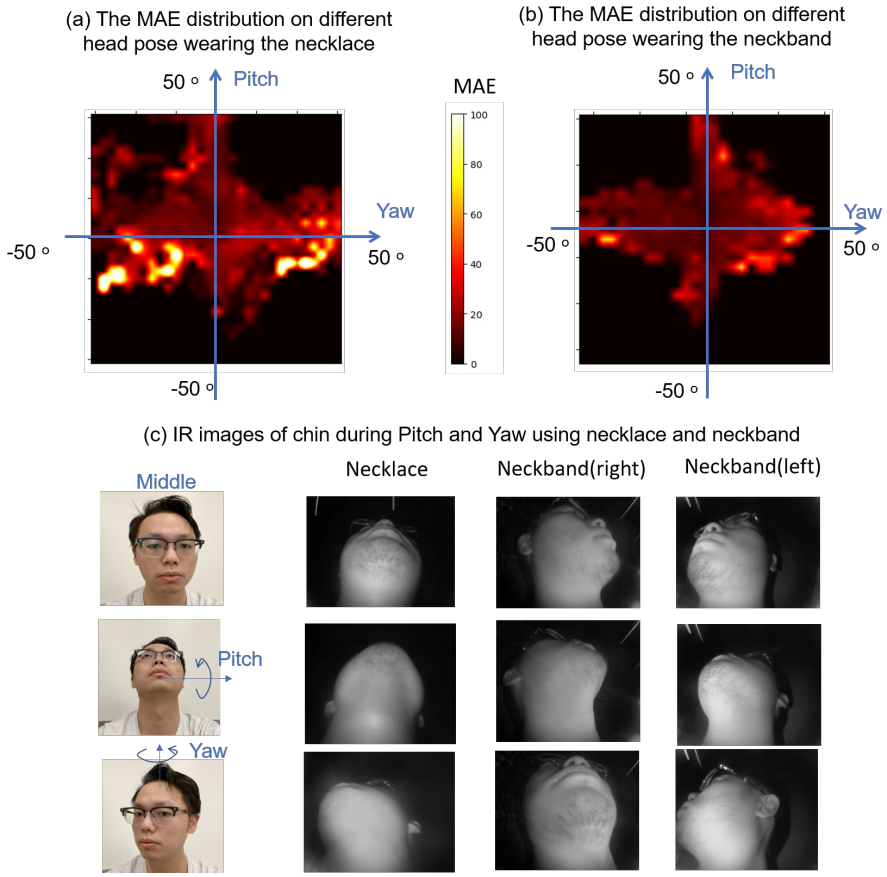

Fig. 18. The raw IR image captured by necklace and neckband during head rotation. The MAE distribution with different head rotation wearing necklace and neckband

reflect different types of skins (e.g., color). In [6, 8], the researchers explored the optical reflectivity and emissivity of infrared wavelengths on different skin colors. They found that the skin pigmentation has little impact on the IR (from 0.76μm - 10μm) reflectivity and emissivity of skin because the IR light cannot penetrate the Stratum Basale, which contains the pigmentation. Based on this finding, we believe our IR camera should work well with different skin colors. In addition, other papers [18, 45, 65] also utilize the infrared camera (such as Leap Motion) on human skin (such as the human hand) in a similar way as our NeckFace. In these papers, they confirmed that the infrared camera (e.g., Leap Motion) could work well with different skin colors [18, 45, 65]. In Summary, based on the results from the above prior research, we think it is likely that the different skin colors would not introduce a significant effect on the performance of NeckFace with 850nm infrared sensing. However, to make a substantial claim, we plan to conduct more studies in the future to investigate this issue systematically.

8.5 Power Consumption Analysis

Power consumption is critical for any wearable device. In this section, we offer the estimation of power consumption on NeckFace. We apply the USB Digital Power Meter [2] to connect to the Raspberry Pi power supply

wire to test the total power consumption. For the Raspberry Pi 4, its no-load power is around $1.75W$. When the camera is connected to the Raspberry Pi, and it is collecting the data (the frame rate of the camera is 30 FPS), the total power consumption is around $2.7W$. When the IR LED turns on, the total power is around $4W$. Taking the EEMB battery with $30 \times 40 \times 6mm$ size, $15g$ and $3.7V$ voltage weight as an example[4], its capacity is $850mAh$, widely used, and suitable to power mobile and wearable devices. When the camera frame rate is $30FPS$, and IR LED is driven by $3.3V$, the working time for one $850mAh$ battery is around 1 hours. Admittedly, the power consumption of our existing prototype now makes it a bit challenging to be immediately adopted on commodity wearables. However, the purpose of the paper is to demonstrate the feasibility of this idea.

We can further optimize the power consumption in the future. In NeckFace, the main power consumption consists of three parts: Raspberry Pi 4, IR camera and IR LED. Firstly, we can deploy other low-power MCU equipped with a wireless module such as Raspberry Pi Zero ($700mW$) or ESP ($210mW$) instead of the Raspberry Pi 4, which can help save plenty of power. Secondly, as for IR camera, we can either lower the frame rate and resolution to decreases the energy, or replace OV5647 with low-power image sensors, such as OV9712[5] (1280x800@30FPS, 110mW), OV9755[6] (1280x720@60FPS, 100mW). Finally, IR LED is consuming a lot of energy. We also have two plans to reduce its power further. (1) Currently, the voltage we use to drive the IR LED module is 3.3V. However, we may not need such a high voltage to power the LED, and reducing the drive voltage can help save much energy. For example, we have measured that when the drive voltage decreases from 3V to 2V, the power consumption of IR LED decreases from $627mW$ to $168.6mW$. (2) Secondly, the IR LED is always turned on, and it is a waste of power while the shutter of the camera is closed. Hence, we can use pulse to drive the LED to lower the power consumption. Specifically, we can turn on the LED just before the camera shutter opens and captures the new frames, while we can keep the LED off at other time. For example, we found that the LED consumption decreases from $627mW$ to $192.6mW$ when the duty cycle decreases from 100% to 30% (frame rate is 30FPS and LED drive voltage is $3V$). So, if we can deploy the C-Face with ESP32 ($210mW$), OV9755($100mW$), and IR LED with 30% duty cycle ($192.6mW$), the battery life ($3.7V$, $850mAh$) can be extended from 1 hour to around 6.25 hours. We also plan to explore the relationship between new low-power hardware setting (LED's duty cycle) and the performance of the facial expression reconstruction model in the future. Nevertheless, we believe that the power load is neither a fundamental limitation of NeckFace nor a significant problem for demonstrating the feasibility of our research idea.

We will explore different options for low-power sensing as the immediate next step before it can be deployed at scale in real-world scenarios.

8.6 Deploying NeckFace on Low-resource Devices

The system we used in the user study, recorded the data first, which were then process on a workstation offline. However, moving forward, it is critical to understand how and whether NeckFace can be deployed on a low-resource platform to make real-time facial movement classification. We explored two different system settings that could make a real-time prediction on the edge device (e.g., wearables). We discuss the system requirement and our implementations for these two settings in the following paragraphs, respectively.

In the first setting, we defer all heavy computing workloads to a remote computer which is more powerful to process the image data. The wearable device only collects the sensing data and transmits the data to the remote computer through wireless communication (e.g., WiFi). In this setting, given most of the edge device can transmit data in real-time, the bottleneck is how fast the data can be transmitted to the remote cloud server, and how much time the cloud need to process the data and return the results.

[4] https://www.amazon.com/EEMB-653042-Rechargeable-Connector-Certified/dp/B082152887?ref_=ast_sto_dp
[5] https://www.ovt.com/sensors/OV9712-1D
[6] https://www.ovt.com/sensors/OV9755

The resolution of each IR image is 480 × 360 pixels (around 53 KB each frame). Given a frame rate of 30 FPS, the required bandwidth to transmit the data is 53 KB/Frame x 30FPS = 1.59 MB/s, which is much lower than the maximum bandwidth in most WiFi networks nowadays. To estimate the processing time on the server, we develop a real-time demo system by deploying the deep learning model on a laptop (GPU: RTX 2080Ti, CPU: Intel Core i7-9700, RAM: 64G.) as the remote server. We transmitted the data through WiFi to this laptop, which real-time processes the data and return the prediction results. The deep learning model needs around 0.0001 to 0.0003s to process the data for results for each frame on the laptop. We have also measured the wireless transmission latency ranges from 10ms to 30ms. Our deep learning model takes 162 MB in the memory of the laptop. The results of the facial expressions were visualized in real-time using a program in Unity with 30 FPS. Although further optimization can be done to improve the FPS or the system requirement further, this setting proved that deploying the model on the cloud is one feasible approach to run NeckFace on low-resource devices.

Another solution to deploy NeckFace is to deploy the machine learning model on the edge device, which is much more challenging than the first setting. To explore this option, we deployed our model using PyTorch on Jetson Nano (which has 128 CUDA Cores for GPU) for real-time classification. The IR camera was connected to the CSI port in the Jetson Nano for data transmission. Upon receiving the new frame of IR image, the Jetson Nano loads the image to the GPU on board, processing and predicting the images on the fly. Our deep learning model takes 162 MB in its memory. Since the whole computing is finished on edge, there is no bandwidth requirement. As for the computation burden, we can achieve a prediction speed of 13 FPS in Jetson Nano. Considering Nano's size are still too large for a practical wearable system, we believe the reasonable next step is to use Google Coral USB accelerator for model prediction. According to the Coral official benchmark, the TPU could achieve 56 FPS for ResNet-50, and 151 FPS for ResNet-152 [7]. Since Coral only allows a quantization model, the speed could be dramatically accelerated and lead to faster prediction. We plan to further explore this option in the future.

8.7 Bring Temporal Information to the Deep Learning

Currently, our deep learning model based on ResNet only utilizes one frame of IR image to reconstruct facial expression and head rotation. However, if we also feed the adjacent images to the deep learning model, which would provide extra-temporal information and improve performance. Especially when one frame is blocked, we can also use the adjacent frames to estimate the reconstruction result in this blocked frame. In the field of human body reconstruction (such as full-body tracking [47, 60], hand track [66]), the temporal deep learning model, such as CNN-LSTM [60] and Temporal Convolutional Network (TCN) [47] has be demonstrated to outperform the traditional CNN. Thus, in the future, we can also deploy a similar structure in NeckFace to further improve our system's reconstruction ability and robustness.

8.8 Applications

After evaluating our system in both sitting and walking scenarios, we believe it has a wide range of application fields and could benefit people's daily lives. These include:

(1) Virtual conferencing and remote collaboration: Mainstream video call requires the user to sit in front of the camera; however, this could be inconvenient if the call is about debugging a piece of hardware on a bench. NeckFace provides the exact solution for this problem: with NeckFace, the user could efficiently conduct other activities while having a "virtual" call with other people.
(2) Facial expression detection in VR: When wearing a VR helmet, the upper part of the face is usually blocked, thus preventing acquiring facial expression information. Our system might be the solution since it only uses the chin, nose, and cheek images from the neck (apparently not blocked by the VR helmet) to reconstruct the facial expressions.

[7] https://coral.ai/docs/edgetpu/benchmarks/

(3) Silent speech: This interface allows the user to give input and command without using the sound, which is usually to help people who cannot speak. Since we could reconstruct the mouth movement, we believe it is feasible for our system to do silent speech word recognition.
(4) Eating detection: People's eating behavior is an essential indicator for their health, and it usually requires the monitoring device to detect when, what and how the food is eaten. Our system seems to be an excellent choice to achieve such goals: our reconstruction could indicate the food intake pattern by detecting the mouth movements and analyzing the eating time.
(5) Mental health monitoring: Facial expressions could also be an essential indicator of mental health. Our system might also help to monitor people's emotions, which helps analyze mental health, thus providing people with better recommendations and advice.

8.9 Limitations and Future Work

All research prototypes have limitations, so does NeckFace. In this section, we discuss the limitations of the current prototype and potential solutions.

8.9.1 Hardware Optimization. Admittedly, the current implementation of NeckFace can not be immediately adopted as a wearable commodity device due to its size because the current system implementation requires multiple hardware components, which have not been fitted into one piece of neck-mounted form factor such as the processor (Raspberry Pi). However, we would like to highlight that the prioritized goal of the paper is to demonstrate the feasibility of the idea of using a neck-mounted IR camera to estimate the full facial movements, in which NeckFace is the first of its kind. As all research prototypes do, our implementation can be further improved towards a more practical wearable device. We do believe that with further engineering efforts, the system has the potential to be implemented on a much smaller piece of hardware. For instance, Raspberry Pi is apparently an overkill for the function needed in NeckFace, as it is only used to collect and transmit the information to the server. The IR sensors can also be smaller with more customized PCBs and modules. We made the hardware choices for fast prototyping as a proof-of-concept. We believe that it is possible to design a customized PCB that can host the processor, WiFi module, lighting module, battery, and IR camera together in one piece, which can be put into a necklace. So, we have designed a customized PCB for the IR camera module, as shown in Figure 19 to reduce the size of the hardware. It is visually much smaller than the version used in the user study. With more interactions, we are confident that the PCB size can become even smaller with advanced engineering efforts to fit into neck-mounted wearables, as there is clear space in the PCB. Continuing optimizing the hardware setting to make it more practical is one of the most important targets moving forward.

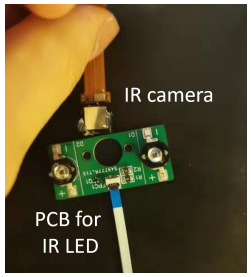

Fig. 19. Smaller hardware configuration

8.9.2 Hair/beard Blocking. One apparent limitation with NeckFace is that if the camera is blocked significantly by the hair, NeckFace may not work well in these scenarios, as shown by P7 in the user study. One possible solution is to explore other locations under the chin, making the camera less venerable to being blocked by the hair. Another potential improvement is to adopt temporal deep learning models(e.g., LSTM) to use the temporal information to make the predicted results more robust to occasional noise. For example, if the hair only occasionally blocked part of the camera, a temporal model may provide a more stable performance than CNN, which provides a prediction on each frame independently.

If a user has a beard that covers the face, it may not impact the performance of NeckFace as much as the long hair if the beard is not directly blocking the camera. Because we think the beard can be recognized as part of the

shapes on the chin, which also moves as the user performing facial expressions. We will investigate this issue in the future.

8.9.3 Strenuous Exercise. In our walking scenario study, we did not control the walking speed of the participants. We just asked them to walk. However, we observed that most participants walk at a speed that was a bit slower than the normal walking speed in the user study. The potential reason is that the participants have to watch the instruction videos as they walk, which slowed their walking speed. Therefore, the results of our study in the walking scenario only indicated a possibility of using NeckFace in walking. However, it is unclear how well the system would perform if the users walk at a much higher speed, running or jumping. These situations may be challenging for the current NeckFace prototype to address, especially for the necklace form factor because the necklace would swing and oscillate during running or jumping. A potential solution is to revise the design of the form factor and fix the position of the camera on the body or the cloth. We will explore other design options to address this issue in the future.

8.9.4 User-dependency. Another limitation of the current system is that our deep learning model is user-dependent, which means each user needs to provide the training data before the system can start tracking their facial expressions. The training process on NeckFace took around 7 minutes, and the user does not recollect training data after remounting.

To evaluate the feasibility of building an independent user model, we use the training data from 12 participants to train a model, evaluate it on the testing data from the remaining participant, and use the remaining one participants. The MAE of the facial expressions was 69.514 and 71.189 for necklace and neckband, respectively. The MAEs for the rotation angles are 9.047 and 8.144 for necklace and neckband, respectively. The results indicate that our system could not perform well for user-independent study to estimate facial expressions, which was not surprising at all. Because different people have different chin and face shapes, it is hard to expect a model with such a small training set to generalize well on new user data. However, if provided with a huge amount of data as speech recognition tasks, it would be interesting to explore how to make the NeckFace user-independent model. This data collection process is highly demanding for resources (e.g., funding, workers). If we have a chance, we would like to partner with industry leaders to explore this possibility.

9 CONCLUSION

This paper presents NeckFace, the first neck-mounted wearable that can continuously track the full facial expressions. It uses neck-mounted IR cameras to capture the chin and face shapes under the neck, which is learned by a customized CNN model to predict facial expressions. A user study with 13 participants showed that NeckFace could track facial expressions well on both necklace and neckbands when the participants were sitting, walking, or remounting the device. We also discussed the opportunities and challenges of applying NeckFace in real-world scenarios.

ACKNOWLEDGMENTS

This work is supported by Information Science Department at Cornell University. We thank all study participants for participating in the study, reviewers for their comments, and lab mates in Cornell SciFi Lab for their support and feedbacks.

REFERENCES

[1] O. Amft and G. Troster. 2009. On-Body Sensing Solutions for Automatic Dietary Monitoring. *IEEE Pervasive Computing* 8, 2 (2009), 62–70. https://doi.org/10.1109/MPRV.2009.32

[2] Amonzon. [n.d.]. Musou USB Safety Tester,USB Digital Power Meter Tester Multimeter Current and Voltage Monitor DC 5.1A 30V Amp Voltage Power Meter, Test Speed of Chargers, Cables, Capacity of Power Banks,Black. [EB/OL]. https://www.amazon.com/Musou-Digital-Multimeter-Chargers-Capacity/dp/B071214RD8 Accessed Oct 4, 2020.

[3] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017. CanalSense: Face-Related Movement Recognition System Based on Sensing Air Pressure in Ear Canals. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 679–689. https://doi.org/10.1145/3126594.3126649

[4] Alper Bozkurt and Banu Onaral. 2004. Safety assessment of near infrared light emitting diodes for diffuse optical measurements. *biomedical engineering online* 3, 1 (2004), 1–10.

[5] Jaekwang Cha, Jinhyuk Kim, and Shiho Kim. 2016. An IR-based facial expression tracking sensor for head-mounted displays. *2016 IEEE SENSORS* (2016), 1–3.

[6] Matthew Charlton, Sophie A Stanley, Zoë Whitman, Victoria Wenn, Timothy J Coats, Mark Sims, and Jonathan P Thompson. 2020. The effect of constitutive pigmentation on the measured emissivity of human skin. *Plos one* 15, 11 (2020), e0241843.

[7] 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* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 112–125. https://doi.org/10.1145/3379337.3415879

[8] Tenn F Chen, Gladimir VG Baranoski, Bradley W Kimmel, and Erik Miranda. 2015. Hyperspectral modeling of skin appearance. *ACM Transactions on Graphics (TOG)* 34, 3 (2015), 1–14.

[9] Weixuan Chen, Javier Hernandez, and Rosalind W. Picard. 2018. Estimating Carotid Pulse and Breathing Rate from Near-infrared Video of the Neck. *CoRR* abs/1805.09511 (2018). arXiv:1805.09511 http://arxiv.org/abs/1805.09511

[10] Jingyuan Cheng, Bo Zhou, Kai Kunze, Carl Christian Rheinländer, Sebastian Wille, Norbert Wehn, Jens Weppner, and Paul Lukowicz. 2013. Activity recognition and nutrition monitoring in every day situations with a textile capacitive neckband. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. 155–158.

[11] Jingyuan Cheng, Bo Zhou, K. Kunze, C. C. Rheinländer, S. Wille, N. Wehn, J. Weppner, and P. Lukowicz. 2013. Activity recognition and nutrition monitoring in every day situations with a textile capacitive neckband. In *UbiComp '13 Adjunct*.

[12] H. Chidananda and Dr. T Hanumantha Reddy. 2017. Human eating/drinking activity recognition using hand movements to monitor and assist elderly people.

[13] Keum San Chun, Sarnab Bhattacharya, and Edison Thomaz. 2018. Detecting eating episodes by tracking jawbone movements with a non-contact wearable sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–21.

[14] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine* 18, 1 (2001), 32–80.

[15] Arnaud Dapogny, Kevin Bailly, and Séverine Dubuisson. 2015. Dynamic facial expression recognition by joint static and multi-time gap transition classification. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1. IEEE, 1–6.

[16] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg. 2010. Silent speech interfaces. *Speech Communication* 52, 4 (2010), 270–287.

[17] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. 2017. End-to-end 3D face reconstruction with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5908–5917.

[18] AS Elons, Menna Ahmed, Hwaidaa Shedid, and MF Tolba. 2014. Arabic sign language recognition using leap motion sensor. In *2014 9th International Conference on Computer Engineering & Systems (ICCES)*. IEEE, 368–373.

[19] Michael J Fagan, Stephen R Ell, James M Gilbert, E Sarrazin, and Peter M Chapman. 2008. Development of a (silent) speech recognition system for patients following laryngectomy. *Medical engineering & physics* 30, 4 (2008), 419–425.

[20] Muhammad Farooq, Juan M Fontana, and Edward Sazonov. 2014. A novel approach for food intake detection using electroglottography. *Physiological measurement* 35, 5 (2014), 739.

[21] Anna Gruebler and Kenji Suzuki. 2010. Measurement of distal EMG signals using a wearable device for reading facial expressions. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 4594–4597.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[23] Shan He, Shangfei Wang, Wuwei Lan, Huan Fu, and Qiang Ji. 2013. Facial expression recognition using deep Boltzmann machine from thermal infrared images. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 239–244.

[24] Steven Hickson, Nick Dufour, Avneesh Sud, Vivek Kwatra, and Irfan A. Essa. 2017. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. *CoRR* abs/1707.07204 (2017). arXiv:1707.07204 http://arxiv.org/abs/1707.07204

[25] Samer Hijazi, Rishi Kumar, and Chris Rowen. 2015. Using convolutional neural networks for image recognition. *Cadence Design Systems Inc.: San Jose, CA, USA* (2015), 1–12.

[26] Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. 2015. Unconstrained Realtime Facial Performance Capture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[27] Peter J Huber. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics*. Springer, 492–518.
[28] E. P. Ijjina and C. K. Mohan. 2014. Facial Expression Recognition Using Kinect Depth Sensor and Convolutional Neural Networks. In *2014 13th International Conference on Machine Learning and Applications*. 392–396.
[29] Yasha Iravantchi, Yang Zhang, Evi Bernitsas, Mayank Goel, and Chris Harrison. 2019. Interferi: Gesture Sensing Using On-Body Acoustic Interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300506
[30] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, et al. 2013. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 543–550.
[31] Haik Kalantarian, Nabil Alshurafa, and Majid Sarrafzadeh. 2014. A wearable nutrition monitoring system. In *2014 11th International Conference on Wearable and Implantable Body Sensor Networks*. IEEE, 75–80.
[32] Leo Kanner. 1931. Judging emotions from facial expressions. *Psychological Monographs* 41, 3 (1931), i.
[33] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. Alterego: A personalized wearable silent speech interface. In *23rd International Conference on Intelligent User Interfaces*. 43–53.
[34] T. Kim, S. Chen, and J. Lach. 2011. Detecting and Preventing Forward Head Posture with Wireless Inertial Body Sensor Networks. In *2011 International Conference on Body Sensor Networks*. 125–126.
[35] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: an ultrasound imaging-based silent speech interaction using deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.
[36] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10, 60 (2009), 1755–1758. http://jmlr.org/papers/v10/king09a.html
[37] Ying-Hsiu Lai and Shang-Hong Lai. 2018. Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 263–270.
[38] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. 2015. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–9.
[39] M. Liu, S. Shan, R. Wang, and X. Chen. 2014. Learning Expressionlets on Spatio-temporal Manifold for Dynamic Facial Expression Recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 1749–1756.
[40] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. 2014. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1805–1812.
[41] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).
[42] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. 2017. The Expressive Power of Neural Networks: A View from the Width. *CoRR* abs/1709.02540 (2017). arXiv:1709.02540 http://arxiv.org/abs/1709.02540
[43] Denys J. C. Matthies, Bernhard A. Strecker, and Bodo Urban. 2017. EarFieldSensing: A Novel In-Ear Electric Field Sensing to Enrich Wearable Gesture Input through Facial Expressions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1911–1922. https://doi.org/10.1145/3025453.3025692
[44] Albert Mehrabian. 2008. Communication without words. *Communication theory* 6 (2008), 193–200.
[45] Chetna Naidu and Archana Ghotkar. 2016. Hand gesture recognition using leap motion controller. *International Journal of Science and Research (IJSR) ISSN (Online)* (2016), 2319–7064.
[46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*. 8026–8037.
[47] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
[48] Subramanian Ramanathan, Ashraf Kassim, YV Venkatesh, and Wu Sin Wah. 2006. Human facial expression recognition using a 3D morphable model. In *2006 International conference on image processing*. IEEE, 661–664.
[49] Ville Rantanen, Pekka-Henrik Niemenlehto, Jarmo Verho, and Jukka Lekkala. 2010. Capacitive facial movement detection for human–computer interaction to click by frowning and lifting eyebrows. *Medical & biological engineering & computing* 48, 1 (2010), 39–47.
[50] Ville Rantanen, Hanna Venesvirta, Oleg Spakov, Jarmo Verho, Akos Vetek, Veikko Surakka, and Jukka Lekkala. 2013. Capacitive measurement of facial activity intensity. *IEEE Sensors Journal* 13, 11 (2013), 4329–4338.
[51] Marc'Aurelio Ranzato, Joshua Susskind, Volodymyr Mnih, and Geoffrey Hinton. 2011. On deep generative models with applications to recognition. In *CVPR 2011*. IEEE, 2857–2864.

[52] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. 2016. Learning Detailed Face Reconstruction from a Single Image. *CoRR* abs/1611.05053 (2016). arXiv:1611.05053 http://arxiv.org/abs/1611.05053

[53] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. 2012. Disentangling Factors of Variation for Facial Expression Recognition. In *Computer Vision – ECCV 2012*, Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 808–822.

[54] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. 2012. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*. Springer, 808–822.

[55] Himanshu Sahni, Abdelkareem Bedri, Gabriel Reyes, Pavleen Thukral, Zehua Guo, Thad Starner, and Maysam Ghovanloo. 2014. The tongue and ear interface: a wearable system for silent speech recognition. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers*. 47–54.

[56] Jocelyn Scheirer, Raul Fernandez, and Rosalind W Picard. 1999. Expression glasses: a wearable device for facial expression recognition. In *CHI'99 Extended Abstracts on Human Factors in Computing Systems*. 262–263.

[57] Tanja Schultz. 2010. ICCHP keynote: Recognizing silent and weak speech based on electromyography. In *International Conference on Computers for Handicapped Persons*. Springer, 595–604.

[58] Nicu Sebe, Michael S Lew, Yafei Sun, Ira Cohen, Theo Gevers, and Thomas S Huang. 2007. Authentic facial expression analysis. *Image and Vision Computing* 25, 12 (2007), 1856–1863.

[59] Inchul Song, Hyun-Jun Kim, and Paul Barom Jeon. 2014. Deep learning for real-time robust facial expression recognition on a smartphone. In *2014 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 564–567.

[60] Shengli Song, Haitao Huang, and Tongxiao Ruan. 2019. Abstractive text summarization using LSTM-CNN based deep learning. *Multimedia Tools and Applications* 78, 1 (2019), 857–875.

[61] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[62] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.* 34, 6 (2015), 183–1.

[63] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. FaceVR: Real-Time Facial Reenactment and Eye Gaze Control in Virtual Reality. arXiv:1610.03151 [cs.CV]

[64] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence* 23, 2 (2001), 97–115.

[65] Babak Toghiani-Rizi, Christofer Lind, Maria Svensson, and Marcus Windmark. 2017. Static gesture recognition using leap motion. *arXiv preprint arXiv:1705.05884* (2017).

[66] Mathias Wilhelm, Jan-Peter Lechler, Daniel Krakowczyk, and Sahin Albayrak. 2020. Ring-based finger tracking using capacitive sensors and long short-term memory. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 551–555.

[67] Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. An Anatomically-Constrained Local Deformation Model for Monocular Face Capture. *ACM Trans. Graph.* 35, 4, Article 115 (July 2016), 12 pages. https://doi.org/10.1145/2897824.2925882

[68] Huiyuan Yang, Zheng Zhang, and Lijun Yin. 2018. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 294–301.

[69] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. 2018. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3359–3368.

[70] X. Zhao, D. Huang, E. Dellandréa, and L. Chen. 2010. Automatic 3D Facial Expression Recognition Based on a Bayesian Belief Net and a Statistical Facial Feature Model. In *2010 20th International Conference on Pattern Recognition*. 3724–3727.