BioFace-3D: Continuous 3D Facial Reconstruction Through Lightweight Single-ear Biosensors

Yi Wu
University of Tennessee, Knoxville
Knoxville, TN, USA
ywu83@vols.utk.edu

Vimal Kakaraparthi
University of Colorado Boulder
Boulder, Colorado, USA
venkata.kakaraparthi@colorado.edu

Zhuohang Li
University of Tennessee, Knoxville
Knoxville, TN, USA
zli96@vols.utk.edu

Tien Pham
University of Texas at Arlington
Arlington, Texas, USA
tienan.pham@uta.edu

Jian Liu
University of Tennessee, Knoxville
Knoxville, TN, USA
jliu@utk.edu

Phuc Nguyen
University of Texas at Arlington
Arlington, Texas, USA
vp.nguyen@uta.edu

ABSTRACT

Over the last decade, facial landmark tracking and 3D reconstruction have gained considerable attention due to their numerous applications such as human-computer interactions, facial expression analysis, and emotion recognition, etc. Traditional approaches require users to be confined to a particular location and face a camera under constrained recording conditions (e.g., without occlusions and under good lighting conditions). This highly restricted setting prevents them from being deployed in many application scenarios involving human motions. In this paper, we propose the first single-earpiece lightweight biosensing system, *BioFace-3D*, that can unobtrusively, continuously, and reliably sense the entire facial movements, track 2D facial landmarks, and further render 3D facial animations. Our single-earpiece biosensing system takes advantage of the cross-modal transfer learning model to transfer the knowledge embodied in a *high-grade* visual facial landmark detection model to the *low-grade* biosignal domain. After training, our *BioFace-3D* can directly perform continuous 3D facial reconstruction from the biosignals, without any visual input. Without requiring a camera positioned in front of the user, this paradigm shift from visual sensing to biosensing would introduce new opportunities in many emerging mobile and IoT applications. Extensive experiments involving 16 participants under various settings demonstrate that *BioFace-3D* can accurately track 53 major facial landmarks with only 1.85 mm average error and 3.38% normalized mean error, which is comparable with most state-of-the-art camera-based solutions. The rendered 3D facial animations, which are in consistency with the real human facial movements, also validate the system's capability in continuous 3D facial reconstruction.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

MobiCom '21, Oct 25-29, 2021, New Orleans, LA
© 2021 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

KEYWORDS

mobile computing, wearable sensing, 3D facial reconstruction, single-ear biosensing

ACM Reference Format:
Yi Wu, Vimal Kakaraparthi, Zhuohang Li, Tien Pham, Jian Liu, and Phuc Nguyen. 2021. *BioFace-3D*: Continuous 3D Facial Reconstruction Through Lightweight Single-ear Biosensors. In *Proceedings of The 27th Annual International Conference On Mobile Computing And Networking (MobiCom '21).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Serving as a major role in human interactions, the face conveys both verbal and non-verbal information, such as intention, engagement, and emotion, which would allow a more credible interaction loopback. Facial landmark tracking and 3D reconstruction thus have been becoming fundamental in various emerging applications which require facial analysis. For instance, facial landmark tracking can be used for driver attentiveness monitoring to detect drowsiness and abnormal behaviors [8]. Continuous 3D facial reconstruction can enable a fully immersive user experience by increasing the awareness of the user's real-time facial expressions and emotional states in virtual reality (VR) scenarios [34]. Moreover, recognizing facial movements can enable silent-speech interfaces for convenient human-computer interactions [16].

Prior Research. Traditional vision-based approaches (e.g., [12, 13, 60]) can localize facial landmarks and produce high-quality facial animations, however, they require a camera positioned in front of the user's face and constrained recording conditions, such as requiring an entire view of the face without occlusions and in good lighting environments. This highly restricted setting makes them not applicable to many application scenarios where users are likely to engage in three-dimensional head movements. A recent work C-Face [10] uses two ear-mounted cameras to reconstruct facial landmarks. Although it enables a relatively more flexible deployment setting, this work still fails to capture the face in inadequate lighting conditions and raises privacy concerns due to the mobility and proximity of the cameras to surrounding people. In addition, deploying energy-hungry cameras on the wearable device may prevent it from practical deployment.

[1] Our rendered facial animation samples can be found at https://mosis.eecs.utk.edu/bioface-3d.html.

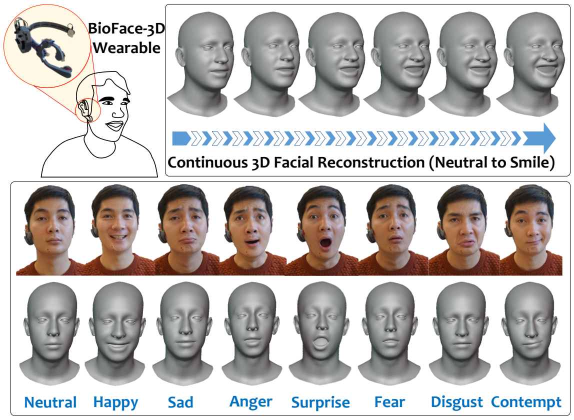


Figure 1: Illustration of the reconstructed 3D facial avatar with various facial expressions[1].

Alternatively, there exist several audio-driven approaches (e.g., [20, 21, 46]) that rely on speech to reconstruct speaking-associated facial movements. However, they neither distinguish between expressions while talking (e.g., talking in an enthusiastic or sad manner) nor can be applied to the scenarios that do not involve human speaking (e.g., silent-speech gestures). Additionally, a lot of wearable-sensor-based methods have been proposed to recognize user's facial gestures, such as magnetic sensing [9], capacitive sensing [37], and electromyography (EMG)-based sensing [40, 41, 58]. However, all these studies can only distinguish a small set of pre-defined facial gestures. To the best of our knowledge, there has been no prior work that can continuously track the positions of facial major landmarks (e.g., the mouth, nose, eyes, and eyebrows) and reconstruct 3D facial animations using camera-free and unobtrusive wearable technology.

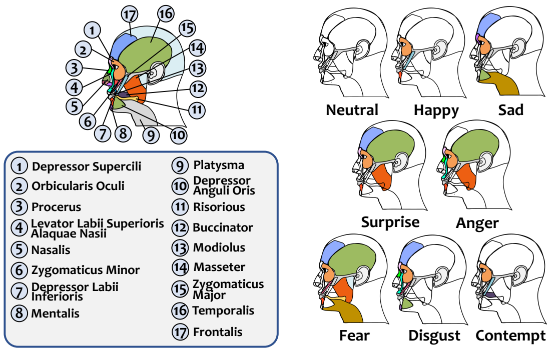
System Objective and Challenges. To circumvent all the limitations of existing approaches, this paper aims to provide a wearable biosensing system that can unobtrusively, continuously, and reliably sense the entire facial movements, track 2D facial landmarks, and further render 3D facial animations through fitting a 3D head model to the 2D facial landmarks. Although existing studies (e.g., [41, 58]) have shown the success of using biosensors, such as EMG and electrooculography (EOG), to detect facial muscle activities and eye movements, realizing such a system is still very challenging:
(1) *Biosensing-based Facial Landmark Tracking:* Tracking facial landmarks via biosensing is an unexplored area. Although the captured biosignals can potentially sense expressive facial deformations, it remains unclear how to learn the spatial mapping between the biosignals and facial landmarks.
(2) *Unobtrusive Facial Sensing:* To allow a long-term facial sensing with minimal impact on the user's mobility and comfort, the obtrusiveness and social awkwardness caused by our designed wearable device should be minimized.
(3) *Continuous 3D Facial Reconstruction:* A compelling 3D facial avatar animation requires the rendered 3D faces to be continuous and smooth over time, and the animation should be generated in a timely manner for real-time applications.

***BioFace-3D* Design.** To address these challenges, we explore a novel point in the design space and propose a single-earpiece biosensing system, *BioFace-3D*, as illustrated in Figure 1. Specifically, *BioFace-3D* uses two-channel biosensors (i.e., surface electrodes) attached to a very small area around one side of the user's ear to capture both EMG and EOG bioelectrical signals. This sensor position ensures the sensing capability of *BioFace-3D* in providing sufficient information for the entire facial reconstruction while still remaining a minimized obtrusiveness level to the wearer. To enable 3D facial reconstruction beyond the confines of cameras, we build a cross-modal transfer learning model that can learn vision-biosignal correspondences in a supervised manner, which pushes the limits of biosensing to enable rich sensing capabilities that are currently infeasible. More specifically, our designed transfer learning model consists of a visual landmark detection network and a biosignal neural network, enabling facial landmark detection knowledge to be transferred across modalities during training time. During testing, the well-trained biosignal network can directly localize 2D facial landmarks from the biosignals, without any visual input. The recognized 2D facial landmarks will be further processed with a Kalman filter and fitted into a generalized 3D head model to render continuous 3D facial animations. Note that our system mainly focuses on continuously tracking facial movements, and the rendered 3D facial avatar uses a generic face model, which lacks detailed features of the individual's face. Our main contributions are summarized as follows:

- To the best of our knowledge, *BioFace-3D* is the first single-earpiece biosensing system that can unobtrusively, continuously, and reliably sense the entire facial movements, track 2D facial landmarks, and further render 3D facial animations through fitting a 3D head model to the 2D facial landmarks.
- Through a thorough anatomical analysis of human facial muscles and elaborate experiments, we identify optimal biosensor placement positions on the face to maintain a minimized obtrusiveness level of *BioFace-3D* to the wearer.
- Relying on the transfer learning across multiple modalities, we push the limits of biosensing to make it possess the capability of other *high-grade* modalities (e.g., vision). This significantly extends its sensing capabilities beyond the common form of biosensing and introduces new opportunities for many emerging applications.
- Extensive experiments involving 9 participants and various settings demonstrated the effectiveness and robustness of the system. The results show that *BioFace-3D* can accurately track 53 facial landmarks with only 1.85 mm average error and 3.38% normalized mean error, which is comparable with most camera-based solutions.

2 PRELIMINARIES

Facial Muscles and Eye Movements. Facial muscles, as illustrated in Figure 2 (a), are striated skeletal muscles lying underneath the skin of the face and scalp to perform important functions for daily life, such as mastication and facial expressions. Different facial movements or expressions are produced by the contraction of a different set of facial muscles [19, 68]. For instance, *smile* involves a person pulling their lip corners up, thereby, raising their cheeks towards the eyes, making the eyelids come closer. These micro-facial movements are mainly driven by zygomaticus major, orbicularis oris, and orbicularis oculi. Differently, *surprise* involves raising eyebrows, widening eyes, opening the mouth, etc., which are usually



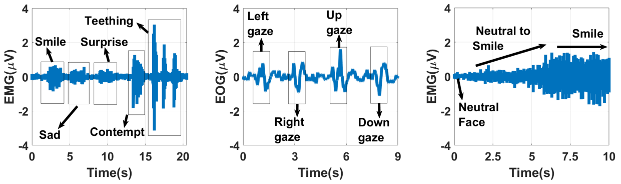
(a) Location of important facial muscles (b) Activated facial muscles during expressions

Figure 2: Illustration of facial muscles.

associated with frontalis, depressor labii inferioris, temporalis, masseter, and orbicularis oris, etc. Figure 2 (b) shows a common set of the activated facial muscles for seven universal expressions of emotion [68]. In addition, the eyeball acts as a dipole with a positive pole oriented anteriorly (cornea) and a negative pole oriented posteriorly (retina) [5]. This shows the potential of tracking the entire facial movements and eye movements through sensing the contraction of corresponding facial muscles and the bioelectrical signals caused by eye movements.

Sensing Facial Muscle Contractions via Single-ear Biosensors. Whenever a muscle contracts, a burst of electric impulses is generated which propagates through adjacent tissue and bone and can be recorded from neighboring skin areas [35]. These bursts of electricity can be captured by surface electrodes using electromyography (EMG) measurements if the electrodes are placed close to or on top of the activated muscles. Although the electrical potentials may pass through the connected muscles to be captured by an electrode, it remains unclear whether we can use the surface electrode attached to a least-obtrusive area, such as the area around one side of the ears, to sense the entire facial movements. We thus conduct an experiment where a surface electrode is attached to one side of the masseter around the ears while a participant performs multiple facial expressions including smile, sad, surprise, contempt, and chewing. As can be seen in Figure 3 (a), multiple events are generated corresponding to different muscle contractions. While some of them are not visually distinguishable due to the wide range frequency response of EMG, the events caused by facial activities can be clearly captured. We prove in Section 6 that the signals of each expression are indeed unique as validated by the Principal Component Analysis (PCA) presented. In the same setting, we ask the participant to look in different directions, and we observe that a unique voltage fluctuation is caused in the electrooculography (EOG) signals depending on the direction and duration of the movement, as shown in Figure 3 (b). These observations confirm the possibility of using single-ear biosensors to sense the entire facial movements.

Continuously Sensing Muscle Contractions. To render continuous and smooth facial animations, the biosensors must be able to continuously track the muscle activities during the transitions between facial events. To validate the feasibility, we conduct an experiment to track the user facial expression while the participant is asked to change their face from neural to smiling with a slower speed than normal (around 10 seconds). The purpose of this experiment is to validate whether the biosensor can capture muscle biosignals generated continuously during facial expression. Figure 3 (c) shows the EMG signals obtained from the experiment, clearly validating the capability of surface electrode in continuously sensing facial activities.



(a) EMG signals of facial activities (b) EOG signals of eye movements (c) EMG signals of a slow smiling

Figure 3: Biosignals collected from a side of masseter around the ears.

3 CHALLENGES & SYSTEM OVERVIEW

3.1 Challenges

Facial Landmarks Tracking via Non-visual Biosignals. Compared with camera-based solutions (e.g., [13, 60, 69]), *BioFace-3D* relies on single-ear biosensors which can only provide non-visual one-dimensional bioelectrical signals (e.g., EMG and EOG). Although the captured biosignals can potentially sense facial skin deformation and facial expression changes, there is no direct spatial mapping relation between the biosignals and facial landmarks. Moreover, some facial expressions (e.g., smiling and sneering) are produced by a similar facial muscle group, making the system even harder to accurately track facial landmarks and animate 3D avatars, regardless of the user's actual facial expressions.

Non-Intrusive Electrode Placements. Most of the facial gestures involve the muscular activities of frontal facial muscles. To make our system less intrusive, we attempt to attach the biosensors to the area around the user's ears. This sensor placement, however, can only sense passive motor activities from the side facial muscles instead of the active frontal muscles. Thus, we have to elaborately select the electrode placement positions via a thorough analysis of facial muscle anatomy, making the system less intrusive to users while still providing sufficient information for the entire facial reconstruction.

Continuous and Smooth 3D Facial Reconstruction. To create a well replicated animated 3D avatar, the reconstructed 3D facial animation from the biosignals should be continuous and smooth, consistent with the user's facial movements. To tackle this, we need to achieve accurate and smooth facial landmark tracking across a range of uncontrolled poses. This requires the system to accurately capture both the spatial changes in facial appearances and the temporal dependencies between adjacent time frames to reconstruct dynamical transitions of facial movements. Moreover, the reconstructed faces should be generated in a timely fashion for real-time applications.

3.2 System Overview

As shown in Figure 4, the proposed *BioFace-3D* has two phases: the *training phase* in which our system uses the biosignals and visual information in a supervised manner to learn the real-time behavioral

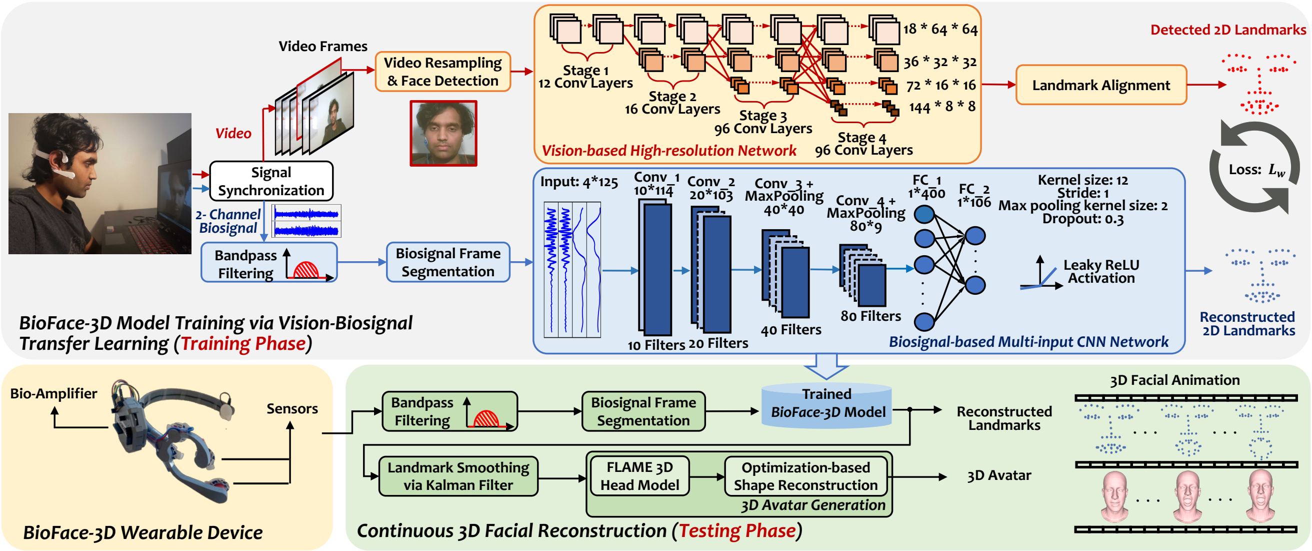


Figure 4: *BioFace-3D* system overview.

mapping from biosignal stream to facial landmarks, and the *testing phase* where the well-trained biosignal network can work independently to perform continuous 3D facial reconstruction, without any visual input. Specifically, during training, we collect visual and biosignal streams using an off-the-shelf camera (e.g., a laptop's built-in camera) and our designed *BioFace-3D* wearable device (Section 6), respectively. We then perform *Signal Synchronization* to ensure the synchronization between the streamed biosignal and the video frames. After that, the visual and biosignal streams are separately processed as follows:

Visual Stream in Training. We first conduct *Video Resampling* to make the recorded videos from different camera types to be resampled in a uniform frame rate, which allows the vision network to take any visual input regardless of its actual frame rate in recording. Next, we perform *Face Detection* for each video frame, and crop the frame to only preserve the detected face. The cropped image frames are then fed into the pre-trained *Vision-based High-resolution Network* for 2D facial landmarks detection. Furthermore, we employ *Landmark Alignment* to eliminate the effect caused by head poses (i.e., scale, rotation, and translation). The detected 2D facial landmarks are then warped and transformed to a uniform aligned coordinate space, which will serve as the ground truth to guide the training of the biosignal network.

Biosignal Stream in Training. *BioFace-3D* collects two biosignal streams from the biosensors integrated into our single earpiece wearable. Each biosignal stream is first processed to obtain both EOG and EMG biosignal streams via *Bandpass Filtering* [56]. We then apply *Biosignal Frame Segmentation* to segment the filtered biosignal stream into frames, each corresponding to a re-sampled video frame. The signal segments are then fed into *Biosignal-based Multi-input CNN Network* to reconstruct 2D facial landmarks. To transfer knowledge from the vision network into the biosignal domain, we utilize the Wing loss [23] to enhance attention of the landmarks which are important but less active (e.g., pupils) and to help the biosignal network learn an accurate spatial mapping between biosignals and facial landmarks.

Biosignal Stream in Testing to Continuously Reconstruct 3D Faces. During testing, the biosignal stream first passes through the same pre-processing procedures in training. Then the fine-tuned biosignal network can continuously reconstruct 2D facial landmarks from the biosignal stream, without any visual input. To ensure a fluent 3D avatar animation, we then apply *Landmark Smoothing via Kalman Filter* to stabilize the facial landmark movement across successive frames. Next, we generate 3D facial animation from the stabilized landmarks using the FLAME (Faces Learned with an Articulated Model and Expressions) model [38]. The generated sequence of fitted head models can then be used for rendering a 3D facial animation that recovers the user's facial movements.

4 BIOSIGNAL-BASED FACIAL LANDMARK RECONSTRUCTION VIA KNOWLEDGE TRANSFER

In this section, we describe the detailed training procedure and the designed knowledge transfer learning network across multiple sensing modalities (i.e., vision and biosignals).

4.1 Signal Synchronization

To guarantee the synchronization between the two modalities' data streams, the user needs to tap the earpiece near the bottom measurement sensor at the beginning of the training phase. This way, a sharp and sizeable peak will be generated in the biosignal stream due to the *skin-electrode contact variation*, while such an event can also be tracked in the video stream with quantifiable accuracy (e.g., through detecting the user's hand using a pre-trained hand keypoint detection model [54]). To detect such a peak in the biosignal stream, we implement a z-score peak transformation algorithm [45], which calculates if any data point of the biosignal stream deviates from a moving average by a given threshold τ. In our implementation, we use a moving window size of 40 milliseconds across all users, which is sufficient to detect the signal peak caused by the finger tap. The threshold τ is set to $\mu_w \pm 0.4\sigma_w$, where μ_w and σ_w are the mean and standard deviation of the sliding window. This z-score based method has been shown to be effective and accurate throughout our system evaluation.

4.2 Data Pre-processing

Visual Stream - Video Resampling & Face Detection. To make our system compatible with various recording devices of different frame rates, we first downsample the recorded video to a uniform frame rate $f_v = 20$, which can also reduce the computational cost for real-time facial reconstruction while maintaining the fluency of the video. Specifically, given the frame rate of the original video f_o, we only keep $\frac{f_v}{f_o}$ of the frames equally distributed in the video buffer, and the timestamps of these frames are then re-scaled to the new timebase (i.e., $\frac{1}{f_v}$). After resampling, we apply a pre-trained Haar Cascade Classifier [59], which provides high accuracy in object detection under varied lighting conditions, to each downsampled video frame for face detection. To meet the required input size of the following vision network, we then make the detected face centered, crop the corresponding square area, and resize the cropped frame to 256 × 256 pixels.

Biosignal Stream - Bandpass Filtering & Biosignal Frame Segmentation. On the biosignal side, we first apply two band-pass filters to extract the main structure of the bio-electrical signals, i.e., EMG and EOG bioelectrical signals [56]. Moreover, in order to transfer knowledge from the vision-based facial landmark detection model into the biosignal modality, we need to match the visual input (i.e., resampled video frames) with the time-series biosignal input. To match with each video frame, we segment the biosignal streams (i.e., both EMG and EOG signals) into overlapped short frames starting at each video frame's timestamp. Given that the gap between adjacent frames is $\frac{1}{f_v} = 0.05s$ and the sampling rate of biosignal is 250 Hz, the length of each biosignal frame is set to $l = 0.5s$ for all experiments, which creates massive overlapped data samples between adjacent biosignal frames as well as sufficient data for the CNN network. This setting makes the subsequent transfer learning model better capture the temporal dynamics and dependencies among continuous biosignal streams to ensure smooth frame transitions in the rendered animation.

4.3 Vision-based High-resolution Network

Conventional image-processing networks for facial landmark detection either rely on low-resolution features built by gradually reducing the size of the feature maps (e.g., TCNN [65]), or utilize a 2-stage high-to-low and low-to-high process to first extract low-resolution features and then rebuild high resolution features through deconvolution and unpooling operations (e.g., encoder-decoder [44]). However, the important spatial and semantic information embedded in the initial high-resolution features might be lost during this process and is hard to recover. To address this and improve the recognition accuracy, we adopt a high-resolution network (HRNet) [60] which maintains high-resolution through the whole process. As shown in Figure 4, the whole network consists of four stages, in which low-resolution convolution streams are added gradually during the training process.

Specifically, the first stage only has a single high-resolution (64 × 64) stream with 12 convolutional layers, and the depth is set to 18. The subsequent stages decrease the resolution to $\frac{1}{2}$ of the resolution of the previous stage and double its depth. Stage 2 adds a lower resolution stream and the number of layers is increased to 16, while Stage 3 and Stage 4 handle more streams in parallel using 96 convolutional layers, with 16 × 16 and 8 × 8 resolution, respectively. Each stage processes a number of convolution streams with different resolutions in parallel. At the end of each stage, information is exchanged among different resolutions via repeated multi-resolution fusions, where low-resolution representations are up-sampled and concatenated with the high-resolution representation. Specifically, we use a pre-trained model on the WFLW dataset [64], which has a total of 98 landmarks, as shown in Figure 5. To reduce computational complexity, we only keep 53 landmarks that cover major facial components such as eyes, eyebrows, nose, and mouth. The output of the vision-based facial landmark detection network provides biosignal modality with transferable knowledge for training the biosignal network.

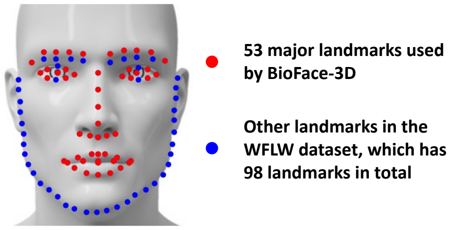

● 53 major landmarks used by BioFace-3D
● Other landmarks in the WFLW dataset, which has 98 landmarks in total

Figure 5: Major facial landmarks used in *BioFace-3D*.

4.4 Landmark Alignment

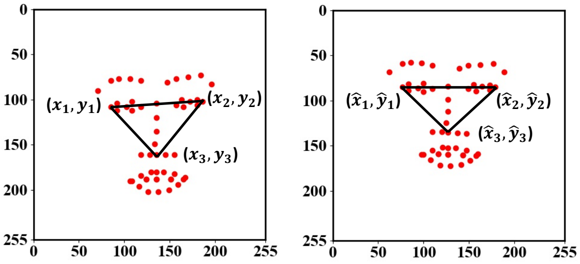
The detected landmark positions can be impacted by large head pose variations caused by head motions, facing directions, and camera angles and positions. To eliminate the impact of these irrelevant factors, we attempt to obtain a canonical alignment of the face based on affine transformations including translation, rotation, and scaling. Specifically, given the coordinate of the i_{th} facial landmark (x_i, y_i), the transformed landmark $(\hat{x_i}, \hat{y_i})$ can be obtained by:

$$\begin{bmatrix} \hat{x_i} \\ \hat{y_i} \\ 1 \end{bmatrix} = \mathbf{R} \cdot \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}, \qquad (1)$$

where **R** is the affine matrix. To derive **R**, we fix the positions of three aligned landmarks (i.e., the left canthus $(\hat{x_1}, \hat{y_1})$, the right canthus $(\hat{x_3}, \hat{y_3})$, and the tip of the nose $(\hat{x_2}, \hat{y_2})$) which are supposed to be static in the aligned coordinate space, as shown in Figure 6. To be more specific, given the video frame size of $w \times w$, the coordinates of two lateral canthus are fixed to $(\lfloor \frac{3w}{10} \rceil, \lfloor \frac{w}{3} \rceil)$ and $(\lfloor \frac{7w}{10} \rceil, \lfloor \frac{w}{3} \rceil)$, respectively. According to ideal facial proportions [22], the tip of the nose is fixed to $(\lfloor \frac{w}{2} \rceil, \lfloor \frac{8w}{15} \rceil)$. With the three fixed landmarks' coordinates and the coordinates before alignment, we can derive all the unknown entries in **R** through solving a set of six-variable linear equations. We can then use Equation 1 to align all the remaining facial landmarks.

4.5 Biosignal-based CNN Network

During the training of the biosignal network, we take the aligned 2D facial landmarks from the vision network as ground truth and train a 1D CNN network to regress the facial landmarks directly from four channels of time-series biosignals (i.e., two EMG and two EOG streams). Other network architectures (e.g., TDNN and LSTM) may also work, but 1D CNN is more suitable for end-to-end



(a) Before landmark alignment (b) After landmark alignment

Figure 6: Illustration of facial landmark alignment.

learning of raw time-series data and has relatively lower computational cost [33]. Specifically, given the default sampling rate of the biosignal f_s = 250 Hz and the length of each biosignal frame l = 0.5 s, the input size of the biosignal network is 4 × 125. The output of the network is the 2D coordinates of 53 facial landmarks. As shown in Figure 4, the network has 4 1D convolutional layers and 2 fully-connected layers. Each convolutional layer has a kernel length of $\lfloor \frac{f_s}{f_v} \rceil$), which is the time gap between adjacent frames. Additionally, the number of filters is doubled when the network is processed to the subsequent convolutional layer, which is initially set to 10. Two max-pooling layers are added to the last two convolutional layers to obtain a more compressed feature map.

Loss Function. Training our learning model comes down to minimizing the designed loss functions to decrease the error between predicted landmark positions and the corresponding ground truths. As landmark position loss treats each individual landmark independently, some important but less-active landmarks, such as pupils compared with lips, may not achieve good attention during training because all the landmarks share an equal weight. To address this issue, we adopt the wing loss function [23], and the loss for each facial landmark is defined as:

$$Loss(x_i) = \begin{cases} w * ln(1 + |x_i|/\epsilon), & \text{if } |x_i| < w \\ |x_i| - C, & \text{otherwise.} \end{cases} \quad (2)$$

where $|x_i|$ is the L2 distance between the ground truth and the reconstructed coordinate for the i_{th} landmark. w represents the threshold of the small error, which is set to 20 in our case. ϵ means the curvature in the small error range, and $C = w - wln(1 + w/\epsilon)$ which links the linear part and non-linear part together. This way the small range errors would obtain more attention when training a regression network, thereby significantly improving the network training capability for the small-scale error landmarks.

Optimization. In addition, the network is trained using the Adam optimizer [32], and the learning rate is set to 0.1 with a decay of 0.9 every 10 epochs. The stride and dilation are all set to 1, and each layer has a dropout rate of 0.3 to avoid over-fitting.

5 CONTINUOUS 3D FACIAL RECONSTRUCTION

In this section, we mainly introduce the testing phase of *BioFace-3D*. Specifically, the well-trained biosignal network takes as input each pre-processed biosignal frame to reconstruct 2D facial landmarks. Then, a Kalman filter and a 3D head model are used to stabilize landmarks and generate 3D facial animation, respectively.

5.1 Landmark Smoothing via Kalman Filter

We observe that the reconstructed facial landmarks regressed directly from the biosignal network are inevitably jittery, which may be caused by the instability of the network as well as the noises introduced in the biosignal. To guarantee the smoothness of the reconstructed 2D facial landmarks over time, we adopt a Kalman filter [47] to stabilize the landmark outputs. Specifically, given a facial landmark in the frame t, we define its state vector $\mathbf{s_t} = [x^t, y^t, v^t_x, v^t_y, a^t_x, a^t_y]^T$, where x^t, v^t_x, a^t_x represents the location, velocity, and acceleration of the landmark, respectively, along x axis, while y^t, v^t_y, a^t_y stands for y axis. A state-space model describing this landmark movement thus can be represented as $\mathbf{s_t} = \mathbf{As_{t-1}}$, and the landmark coordinates $\mathbf{z_t} = \mathbf{Hs_t}$, where the state transition matrix $\mathbf{A}$ and the observation matrix $\mathbf{H}$ can be defined as:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \Delta t & 0 & \frac{1}{2}\Delta t^2 & 0 \\ 0 & 1 & 0 & \Delta t & 0 & \frac{1}{2}\Delta t^2 \\ 0 & 0 & 1 & 0 & \Delta t & 0 \\ 0 & 0 & 0 & 1 & 0 & \Delta t \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \mathbf{H} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}^T, \quad (3)$$

where $\Delta t = \frac{1}{f_v}$ represents the time interval between two adjacent frames. Given the known constant variable Δt and the frame size of 256 × 256, based on the relationships between the six variables in $\mathbf{s_t}$, the process and measurement noise covariances, $\mathbf{Q}$ and $\mathbf{R}$, are set to:

$$\mathbf{Q} = \begin{bmatrix} \frac{1}{4}\Delta t^4 & 0 & \frac{1}{2}\Delta t^3 & 0 & \frac{1}{2}\Delta t^2 & 0 \\ 0 & \frac{1}{4}\Delta t^4 & 0 & \frac{1}{2}\Delta t^3 & 0 & \frac{1}{2}\Delta t^2 \\ \frac{1}{2}\Delta t^3 & 0 & \Delta t^2 & 0 & \Delta t & 0 \\ 0 & \frac{1}{2}\Delta t^3 & 0 & \Delta t^2 & 0 & \Delta t \\ \frac{1}{2}\Delta t^2 & 0 & \Delta t & 0 & 1 & 0 \\ 0 & \frac{1}{2}\Delta t^2 & 0 & \Delta t & 0 & 1 \end{bmatrix}, \mathbf{R} = \begin{bmatrix} 12.5 & 0 \\ 0 & 12.5 \end{bmatrix}. \quad (4)$$

The process noise covariance is a covariance matrix associated with the errors in the state vector s_t, where the noise of acceleration is initialized to 1. This covariance will automatically get updated to achieve a good state. The values in the measurement noise covariance matrix are set to a relatively large value, which ensures that jittery landmarks with larger errors can still be effectively smoothed. The smoothed landmark coordinate in the frame t can then be derived as $\mathbf{H\hat{s}_t}$, where $\mathbf{\hat{s}_t}$ is the optimal state estimate.

We calculate the average standard deviation of all mouth-related landmarks as the evaluation metric to validate the effectiveness of the Kalman filter. Specifically, we select 4 minutes of reconstructed landmarks in which the user repeatedly performs the *surprise* expression. In addition to the Kalman filter, we also implement a simple linear interpolation technique, in which the average of each adjacent frame pair is compensated between them. Specifically, the average standard deviation of all mouth-related landmarks is 8.66 if no smoothing techniques are applied, 8.61 when simple linear interpolation is utilized, and 7.73 when the Kalman filter is implemented. The results demonstrate the effectiveness of the Kalman filter on landmark smoothing.

5.2 3D Avatar Generation

To improve system usability and reduce modeling complexity, we seek a compact head model that can be easily fitted to data while preserving enough details to generate expressive facial animations.

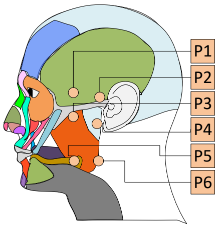


Figure 7: Potential electrode placements.

FLAME 3D Head Model. The FLAME (Faces Learned with an Articulated Model and Expressions) model [38] is a statistical 3D head model that uses a learned shape space of identity variation and articulated jaw, neck, and eyeballs to achieve accurate, expressive, and computationally efficient 3D face modeling. The model is based on linear blend skinning and corrective blendshapes, and contains 5023 vertices and 4 rotary joints (neck, jaw, and eyeballs). The modeling process can be viewed as a function: $M(\vec{\beta}, \vec{\theta}, \vec{\psi}) : \mathbb{R}^{|\vec{\beta}|\times|\vec{\theta}|\times|\vec{\psi}|} \rightarrow \mathbb{R}^{3N}$, that takes shape $\vec{\beta} \in \mathbb{R}^{|\beta|}$, pose $\vec{\theta} \in \mathbb{R}^{|\theta|}$, and expression coefficients $\vec{\psi} \in \mathbb{R}^{|\psi|}$ and return N vertices. The model is composed of a template mesh of a neutral pose, shape blendshapes, pose blendshapes, and expression blendshapes, which are used to account for variations caused by identity, pose deformation, and facial expressions, respectively.

Optimization-based Shape Reconstruction. To generate a 3D head model that reflects the user's facial movements and expressions, we exploit a 2-stage optimization process to fit the generic 3D head model to the 2D landmarks extracted from biosignals. In the first stage, we conduct camera calibration by optimizing the parameters for rigid transformation, including scale, rotation, and translation, to minimize the L_2 distance between the landmarks and the corresponding 3D head model vertices projected into the 2D space. In the second stage, we optimize the model parameters (e.g., pose, shape, and expression) by optimizing the L_2 distance while regularizing the shape coefficients $\vec{\beta}$, pose coefficients $\vec{\theta}$ (including neck, jaw, and eyeballs), and expression coefficients $\vec{\psi}$ by penalizing their L_2 norms. After optimization, we can generate a 3D head model that recovers the user's facial expressions. We note that the generated avatar is based on a generic head model that aims to capture the user's facial movements rather than to reconstruct user-specific facial details (e.g., pores and moles). Potential ways for further improving the personality of the generated 3D avatar are discussed in Section 8.

6 SYSTEM IMPLEMENTATION

6.1 Electrode Placements

A traditional bio-electrical sensor channel includes three types of electrodes: *reference electrode*, *measurement electrode*, and *ground electrode*. To provide a relatively stable reference point and driven ground, the reference electrode and ground electrode should be attached to bony areas to keep all the underlying muscular signals minimized. Thus, we attach these two types of electrodes to the back of the ears (i.e., mastoid bone) in the design of *BioFace-3D*. Regarding the measurement electrode, from our analysis in terms of the unobtrusiveness and the capability of sensing, it could be placed at six locations P1-P6 as illustrated in Figure 7. In particular, P1 is on the temporalis, proximity to orbicularis oculi; P2 is on the temporalis and temporal bone, proximity to deeper head; P3 is on the masseter (on the zygomatic bone); P4 is at the junction of the mandible and temporal bones with proximity to temporalis and masseter; P5 is at the junction of risorius, masseter, platysma, on the mandible bone; and P6 is on the lower side of the masseter, proximity to risorius and platysma. To find the most suitable location for the measurement electrode, we perform both SNR and Principal Component Analysis (PCA) analyses below.

Table 1: SNR results from the six facial locations in decibels (dB).

	Happy	Sad	Angry	Surprise	Fear	Disgust	Contempt
P1	12.89	4.24	7.00	4.87	7.07	1.98	16.50
P2	8.49	2.95	7.06	7.65	5.23	1.86	8.55
P3	10.18	5.50	7.00	4.87	7.07	1.98	16.50
P4	3.70	***1.58***	3.19	5.31	2.50	***1.27***	12.06
P5	6.58	11.97	4.19	3.14	2.60	***0.89***	16.24
P6	3.82	3.47	3.86	3.13	***1.54***	***0.56***	11.18

SNR Analysis. We calculate the Signal-to-Noise Ratio (SNR) of the signals generated by each of the universal facial expressions using the six measurement electrode locations. To ensure the acceptable quality of the measured biosignals, the SNR should be greater than 1.2 db [61]. However, from our experiments we observe that a value more than 1.6 dB is acceptable to withstand the baseline noise variations. The results are shown in Table 1. We observe that P4-P6 have a relatively low SNR (< 1.6 dB) for some of the expressions. For instance, *sad* has a low magnitude at P4 because the location is situated outside the masseter, which has loose connections to the depressor anguli oris that facilitates the facial gesture. *Fear* has low magnitude at P6 as it is at the lower end of the masseter that has no connection to the muscles deforming the mouth. Through this analysis, we found that P1, P2, P3 locations perform well with all the universal facial expressions.

PCA Analysis. Although SNR is a great indicator for detecting facial activities, it does not provide sufficient details on the quality of the signals in distinguishing different facial activities. To analyze the distinguishability of the captured signals of different facial movements, we transform the gestural signals from P1, P2 and P3 locations to the frequency domain using Discrete Fourier Transform (DFT). The DFT signals are then projected into new dimensional space for feature engineering via Principal Component Analysis (PCA) separability scores [31]. The overall separability scores at P1, P2, P3 are 94.25%, 93.53%, 92.27%, respectively. This result affirms the fact that each facial movement generates a unique physiological signature at each of these facial locations. Specifically, P1 and P2 are affected by an overlap between *fear* and *anger* gestures while P3 is affected by an overlap between *smile* and *contempt* gestures. P3 can distinguish *fear* and *anger* due to its connections to frontal face muscles while P1 and P2 can separate *smile* and *contempt* as they can capture buccinator and zygomatic major activations in a fine grained manner. Due to the intrusive nature of P1, we choose to use two measurement electrodes at P2 and P3, which can complement to each other to sense the entire facial activities.

6.2 Prototype

Single Ear-piece Design. From our experiments, the gestural signals generated on both sides of the face are observed to be very similar in magnitude, shapes, etc. In particular, there are no significant changes to the dimensional space and separability scores

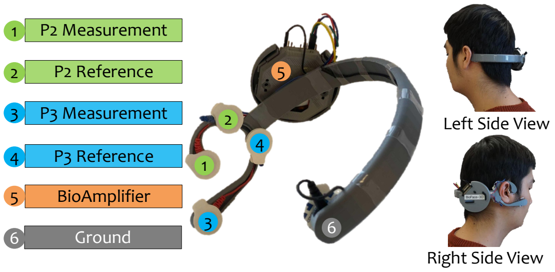


Figure 8: *BioFace-3D* prototype.

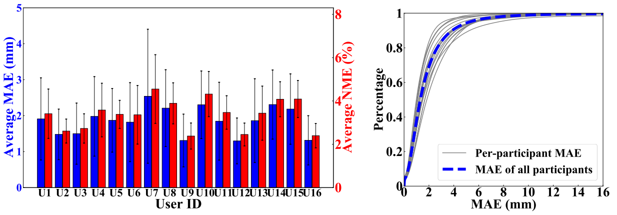
of the gestural signals after PCA. For P2, the dimensional space has 270, 272, and 272 components when we use the data from left side, right side and both sides of the face, respectively. For P3, the dimensional space has 153, 156, and 158 components. Hence, there are almost no unique features that can be added by the data from the second side of the face. The dimensional space explains 95% of the variance of the dataset and the separability scores for each case does not vary by more than 1% while remaining higher than 92%. Thus, universal gestures that involve muscle groups from both sides of the face can be captured with equal detail from electrode channels being placed on just one side of the face.

Prototype. The *BioFace-3D* wearable device is customized based on (a) dimensions of the user's head (b) preference for the side of the earpiece. The earpiece design is dictated by the facial locations of measurement electrodes P2 and P3 as described previously. The reference electrodes are placed on a bony surface behind the ear such that those electrodes are sufficiently away from the facial muscle activity that the measurement electrodes capture. The earpiece provides slots for measurement, reference and ground electrode placements at precise locations as illustrated in Figure 8. This earpiece is integrated with a headband that goes around the neck. We designed three sizes of prototypes that place the sensors in appropriate facial locations for three adult population groups: Large, Medium, and Small. For each of the sizes we designed two variants based on which side the earpiece is present. This allows for a wearable device that suits a large population. This headpiece also houses a circuit box to contain the hardware. All of the components in the headset are manufactured by 3D printing of PLA to ensure that the prototype is lightweight. *BioFace-3D* uses an ADS1299 based bio-amplifier circuit, i.e., OpenBCI [6, 29], and Ag/AgCl surface electrodes [1] that stick to the user's skin, as illustrated in Figure 8. A Bluetooth module is integrated for data streaming.

7 PERFORMANCE EVALUATION

7.1 Experimental Methodology

Experimental Setup & Data Collection. We recruited 16 participants to evaluate the performance of *BioFace-3D*[2]. Particularly, the participants include 11 males and 5 females, aging from 21 to 34 years old. Six of them wore glasses during the data collection as usual. To evaluate the performance of tracking 53 facial landmarks, we focus on seven universal facial expressions of emotion [18] involving *happy*, *sad*, *anger*, *surprise*, *fear*, *disgust*, and *contempt*, as shown in Figure 1. The participants were asked to sit in front of a camera (for training and ground truth recording purposes) and repeatedly perform the aforementioned seven expressions while wearing our implemented *BioFace-3D* prototype. Each expression was separated by a *neutral* facial expression (i.e., relaxed facial expression). To assist participants with their data collection, seven pictures were displayed on a screen portraying the corresponding faces for them to imitate. The pace and to which extent each expression was performed were not controlled throughout the experiments. To show the generalizability of our system in using various types of cameras for training, we used a variety of cameras of different resolutions and recording frame rates (e.g., 720P, 1080P resolutions, and 25, 30 fps), including the webcam of a Lenovo ThinkPad X1, a Lenovo Ideapad Y700, a MacBook Pro 2019, an EMeet C960 Webcam on a desktop, and the built-in rear camera of an iPhone 8.



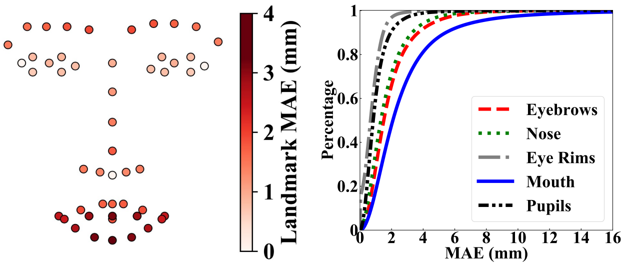
(a) Per-participant landmark tracking error (b) MAE CDF for each participant

Figure 9: Performance of continuous facial landmark tracking for each participant.

Particularly, each participant was asked to repeatedly make each facial expression for 4 minutes, which leads to about 40 to 50 rounds of facial expressions. The data collection lasts for 28 minutes (7 facial expressions in total) for each participant, and their eye movements were not constrained during the data collection. Unless mentioned otherwise, for each participant, we use the first 20 minutes of data for training and the remaining 8 minutes of data for testing. The default sampling rate of biosignals per channel was set to 250 Hz. The impact of sampling rate on performance will be discussed in Section 7.3.1. After data collection, we also asked participants to complete a questionnaire on their experience with *BioFace-3D*, which is elaborated in Section 7.4. We also extended our experiments to other types of facial movements (i.e., speaking) with 5 participants involved, which is detailed in Section 7.2.3. We further collected 4 additional datasets with one participant involved to study the impact of facial occlusion and bursty head movements. Three participants separately evaluated the performance of eye tracking and tested the system's temporal stability when training and testing data are separated by multiple days. The data collection details for these tests are elaborated in Section 7.2.2 and Section 7.3.

Evaluation Metrics. *1) Mean Absolute Error (MAE)* is the absolute error between the reconstructed landmarks and groundtruth landmarks, which are converted from pixels to a physical unit (millimeter). The MAE of a single landmark can be calculated as $MAE = ||g - r||_2 \times \frac{l_r}{l_f}$, where g and r represent the groundtruth and reconstructed landmark coordinates, respectively. l_f is the distance between the two lateral canthus in the frame, which is $\lfloor \frac{2w}{5} \rceil$ as aforementioned in Section 4.4, while l_r is the distance between the two lateral canthus of the participant we measured; *2) Normalized Mean Error (NME)* is the mean error between the groundtruth

[2] The study has been approved by our Institutional Review Board (IRB).



(a) Visualization of the average MAE (b) MAE CDF for each facial feature

Figure 10: Performance of continuous facial landmark tracking for each facial landmark and facial feature.

and reconstructed landmark coordinates, normalized by the inter-ocular distance, which is a commonly used metric in camera-based solutions for facial landmark tracking. Given the groundtruth and reconstructed coordinates of landmark as g and r, the NME can be calculated as $NME = \frac{||g-r||_2}{||g_{lp}-g_{rp}||_2}$, where g_{lp} and g_{rp} denote the groundtruth of left pupil and right pupil, respectively.

7.2 Overall System Performance

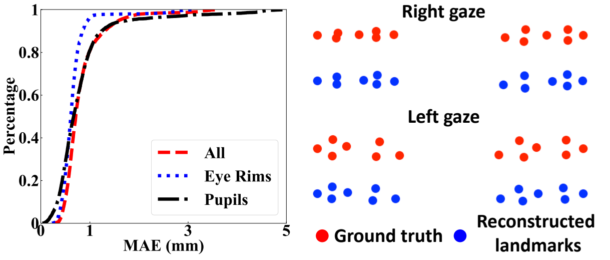
7.2.1 Facial Landmark Tracking (Facial Expression). Figure 9 (a) illustrates the average MAE & NME and corresponding standard deviations for all the 53 facial landmarks of each participant. We observe that all the participants can achieve comparable low errors. Specifically, *BioFace-3D* obtains an average of 1.85 mm MAE and 3.38% NME with average standard deviations of 0.99 mm and 0.90%, respectively, indicating that mm-level accuracy could be achieved in our system. Among all the participants, *U12* achieves the best reconstruction results with only 1.29 mm MAE and 2.45% NME, while *U7* has the largest error (i.e., only 2.54 mm MAE though). Figure 9 (b) depicts the Cumulative Density Function (CDF) of the MAE errors for each individual participant as well as cross-participant cases. 80% of the reconstructed landmarks have a low MAE of < 2.66 mm, which demonstrates the promising capability of *BioFace-3D* in tracking human 2D facial landmarks.

In addition, distinct landmarks may have different scales of errors due to their movement variability. Figure 10 (a) visualizes the average MAE for the entire 53 major landmarks, and Figure 10 (b) shows the CDF of the categorized landmarks. We find that reconstructed landmarks on the mouth have a relatively larger error, but 80% of them are still within an acceptable range (i.e., < 3.87 mm). Eye rims (12 landmarks in total without pupils) and pupils (2 landmarks only) achieve a relatively lower MAE error. Specifically, 80% of the reconstructed eye-related landmarks are within 1.17 mm, indicating *BioFace-3D* can accurately track the unconstrained eye movements of the participants during data collection.

As NME is a commonly used metric in vision-based facial landmark tracking, we directly compare our landmark tracking results with several state-of-the-art vision-based solutions [60, 67, 70] in Table 2. These vision-based solutions were evaluated using multiple public image datasets (e.g., WFLW [64], 300-W [50]) which have manually labeled groundtruths and different numbers of facial landmarks to be reconstructed. Although it might not be a fair comparison as our dataset is self-collected and we use a pre-trained camera-based network to generate landmark groundtruths instead of human labeling, the comparable NME accuracy shows the promising performance of *BioFace-3D*, even compared with vision-based solutions.

Table 2: Comparison with vision-based solutions.

Methods	Dataset	# of Landmarks	NME
SDM [70]	300-W	68	7.52
	LFPW	68	5.67
CFSS [67]	300-W	68	5.76
	LFPW	68	4.87
HRNet [60]	300-W	68	2.87
	WFLW	98	4.60
BioFace-3D	**Self-collected**	**53**	**3.38**

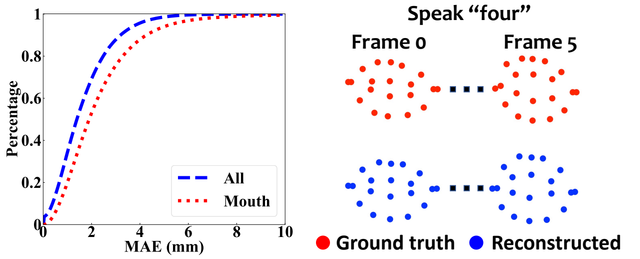


(a) MAE CDF for eye-related landmarks (b) Ground truth & reconstructed landmarks

Figure 11: Performance of continuous eye-tracking.

7.2.2 Eye Movement Tracking. To better evaluate the performance of gaze tracking, we collected another dataset involving three participants, who were asked to repeatedly look into four different directions (i.e., left, right, up, and down) for 300 seconds. Each gazing activity lasts for 2 seconds and was separated by 1 second *looking straight ahead*, which results in a total of 100 gaze movements. For each participant, we use the first 4 minutes for training and the remaining 1 for testing. The MAE CDF is shown in Figure 11 (a), in which we achieve an average MAE of 0.82 mm for all eye-related landmarks, 0.73 mm for eye rims, and 0.95 mm for pupils. We found that 80% of the pupil landmarks have an error lower than 0.98 mm, which shows the promising capability of *BioFace-3D* for gaze tracking even in this active eye-moving setting. Examples of the reconstructed landmarks (right/left gaze) are shown in Figure 11 (b).

7.2.3 Facial Landmark Tracking (Speaking). To comprehensively evaluate our system, we extended our experiments to other types of facial movements (i.e., speaking) by involving five participants who were asked to repeatedly speak nine digits (i.e., one to nine). During experiments, each digit was repeatedly spoken for 4 minutes, which results in a total of 36 minutes of data. We used 26 minutes of data for training and the remaining 10 minutes data for testing. The CDF curves for MAE are shown in Figure 12 (a), in which we achieve an average MAE of 1.63 mm for all facial landmarks, while 2.39 mm for mouth-related landmarks. We found that 80% of the mouth landmarks have an error lower than 3.29 mm, which shows the promising capability of *BioFace-3D* in tracking facial movements of speaking. Examples of the reconstructed mouth landmarks at a interval of five frames of speaking *four* are shown in Figure 12 (b). The promising results demonstrate the capability *BioFace-3D* of tracking the users' mouth movements while they are speaking, potentially extending our system to other usage scenarios such as speech enhancement. We leave this as our future work.



(a) MAE CDF for mouth-related landmarks (b) Ground truth & reconstructed landmarks

Figure 12: Performance of continuous mouth movement tracking while the user is speaking.

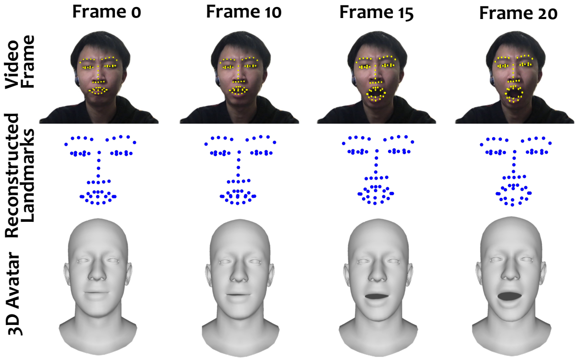


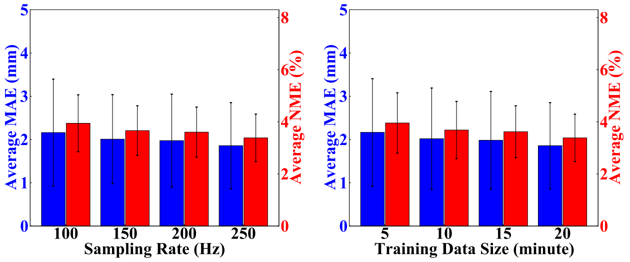
Figure 13: Example of the rendered facial animation.

7.2.4 *Continuous 3D Facial Reconstruction.* To test *BioFace-3D*'s ability of continuous 3D facial reconstruction, we show the video frames, reconstructed landmarks, and rendered 3D avatar frames at a interval of 5 frames in Figure 13. We can observe that *BioFace-3D* is able to capture the user's facial gestures from biosignals in a continuous manner and further render a smooth 3D facial animation that includes the entire facial changes. Our rendered facial animation samples can be found at [2] with the password *mobicom2021*.

7.3 Micro-benchmark Tests

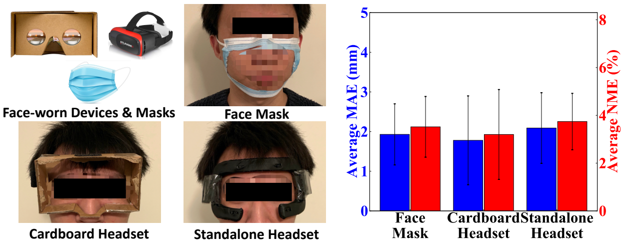
7.3.1 *Impact of Biosensor Sampling Rate.* To evaluate the impact of sampling rates on our system, we down-sample the frequency of the biosignal collected at 250 Hz to 50-200 Hz. Figure 14 (a) presents the average MAE and NME when varying the sampling rate from 100 Hz to 250 Hz. We observe that high sampling rate slightly improves the performance, and *BioFace-3D* is not very sensitive to changes in the sampling rate, given the range from 100 Hz to 250 Hz. Even if the sampling rate is decreased to 100 Hz, *BioFace-3D* still achieves an average MAE of 2.16 mm and NME of 3.94%, with average standard deviations of 1.23 mm and 1.09%, respectively. These results show that our system can also provide good performance even with a lower sampling rate, which can further reduce the computational complexity and power consumption.

7.3.2 *Impact of Training Data Size.* We then evaluate the system robustness with different training data sizes to seek the potential of further reducing training efforts. Figure 14 (b) presents the overall system performance when varying the training data size from 5 minutes to 20 minutes for each participant, while all the remaining data is used for testing. We observe that even if the size of training data is decreased to 5 minutes, *BioFace-3D* still achieves an average MAE of 2.17 mm and NME of 3.95%. A larger training size would lead to better accuracy, but it remains operable if a user intends to have a quick enrollment process.



(a) Impact of sampling rate (b) Impact of training data size

Figure 14: Performance of facial landmark tracking with different sampling rate & training data size.

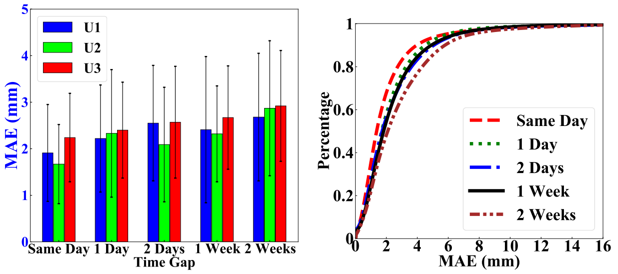


(a) Different facial occlusions (b) Landmark tracking error

Figure 15: Performance of continuous facial landmark tracking under the presence of facial occlusions.

7.3.3 *Impact of Face-worn Devices and Masks.* Wearing face-worn devices/masks involve external forces (e.g., rubber bands for face coverings), which would tighten facial muscles and add additional pressure on the prototype, potentially introducing noises to biosensor readings. We further test the system performance with the presence of a face mask or a VR headset (a cardboard headset or a standalone headset), as shown in Figure 15 (a). Specifically, we asked a participant to wear a face mask and two types of VR headset (i.e., a cardboard headset and a standalone headset) respectively while using our system. The training data is the 20 minutes data with no occlusion involved. To obtain the ground truth from the vision-based network, we cut off the front side of the mask to expose the mouth of the user, and tear off the headsets to reveal the user's eyes & eyebrows, as shown in Figure 15 (a). Figure 15 (b) presents the system performance when wearing face masks and head-worn VR headsets. Although wearing face-worn devices/masks decreases the performance, the overall performance remains within an acceptable range, e.g., average MAE of 1.93 mm, 1.78 mm, and 2.09 mm while wearing a face mask, a cardboard headset, and a standalone headset, respectively. These results demonstrate the robustness of *BioFace-3D* with different facial occlusions.

7.3.4 *Resilience to Bursty Head Movements.* We are also interested in how bursty head movements impact our system. Specifically, the participant was asked to regularly rotate & shake his head



(a) MAE for cross-day setting (b) CDF for cross-day setting

Figure 16: Performance of landmark tracking over time.

during testing data collection, while ensuring the head could be captured by camera for the ground truth acquisition. The training data is the 20 minutes data with no head movements involved. With an average MAE of 1.79mm and an NME of 3.25%, BioFace-3D is resilient against active head movements, making it applicable to many practical scenarios involving active head movements.

7.3.5 Temporal Stability. The sensor measurement would be influenced by the day-by-day change of the users' body status, uncontrollable impurities on the skin surface, and the sensor displacement as the prototype won't be worn in exactly the same way. As time passes by, these issues may become more serious and therefore affect the sensor measurements at a greater scale. It is thus important to validate the system's temporal stability to prevent repetitive training. We asked three participants to collect another four sets of testing data (10 minutes each) which is separated from training data by 1 day, 2 days, 1 week, and 2 weeks. As shown in Figure 16 (a), we found that in the worst case, *BioFace-3D* stills reaches an MAE of 2.87 mm over two weeks and there is no significant performance change in two-week period, as illustrated in Figure 16 (b). These results affirm the fact that the sensitivity to sensor placement positions, which tend to differ minutely with each usage, have a negligible effect on the system outputs.

7.3.6 Computational Cost & Power Consumption. The inference time of 53 facial landmarks is measured on a single NVIDIA GTX 2080Ti GPU, and our designed transfer learning model only takes around 0.033 ms to reconstruct a single frame, which is sufficient for real-time applications. In addition, we use a power monitoring device (i.e., Monsoon High Voltage Power Monitor [3]) to measure the power consumption of *BioFace-3D.* All measurements are conducted at 60°F with a normal Lipo battery voltage (3.7V). Specifically, if the system is in the idle state where MCU is working in the idle mode without streaming data via Bluetooth, *BioFace-3D* consumes 118 mW on average. If *BioFace-3D* is sensing and streaming biosignal data via Bluetooth, the whole system's power consumption is 138 mW. This indicates that *BioFace-3D* can provide continuous data logging for 8.2 hours using a 500 mAh Lipo battery, which meets the requirements of most applications.

7.4 User Study

We asked the participants to fill a questionnaire, as shown in Figure 17, on their experience with *BioFace-3D* after the experiments. We found that 81.3% of the participants are willing to use *BioFace-3D* and 75% of the participants feel it's comfortable to wear. 50% of the participants think *BioFace-3D* is easy to use and 31.3% of the participants feel it's very easy to use. We only got one negative feedback towards *BioFace-3D* , simply because the specific participant "doesn't want to have anything around his head". Additionally, 81.3% of the participants prefer *BioFace-3D* rather than traditional camera-based solutions, mostly due to the reason that *BioFace-3D* is more privacy preserving, can detect facial expression independently of body movements, and is reliable when parts of the face are blocked. Finally, all participants can use it for more than 30 minutes and 81.3% of the participants can use it more than 1 hour, which is sufficient for many usage scenarios. Specifically, the major reason which made 18.8% of the participants only choose to wear it for 30 minutes is the lack of adjustability. Due to the size of the prototype being fixed at the current stage, sometimes it cannot fit the user's head very appropriately and will cause displacement as time passes by, which may downgrade user experience. We plan to address this issue by utilizing more flexible materials to enhance the size variability. This is considered as our future work.

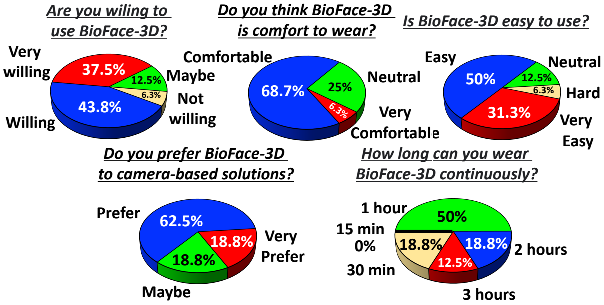


Figure 17: Results of user study questionnaire.

8 DISCUSSION

User-independent Model. As the signal strength, response, and sensitivity of biosignals may vary from user to user, we currently adopt user-specific training to mitigate this variance and improve system accuracy. However, this might reduce the usability as new users have to undergo the enrollment phase before using the system. To improve usability, we can potentially train a generic user-independent model using data collected from a large set of users. When new users are introduced, the generic model can be adapted to the users with few calibration samples via meta-learning-based few-shot adaptation [24, 43]. We leave this as our future work.

Effects of Body Movements. Motion artifacts are a formidable noise that occurs in all Electrogram measurements [62]. It is a low-frequency noise occurring in the EOG frequency range. They occur due to two reasons: 1) relative motion between the surface electrodes and the skin surface; and 2) connection quality fluctuations between the wires and electrodes. We ensured that motion artifacts are mitigated by the design of the prototype which maintains the contact quality of the surface electrodes and keeps the connecting wires very short. This is evidenced by the evaluation of the system under rapid head movements in Section 7.3.4. The body movements such as walking would have significantly less impact on the results as they introduce less relative motion between electrodes and skin as well as electrodes and wires when compared to rapid head movements. We note that the participants were allowed to move freely during our experiments as long as they can be captured in the video.

Improving Landmark Granularity and Animation Quality. Aiming to reduce device weight and improve user comfort, the current design of *BioFace-3D* uses 2-channel biosignal and can provide a landmark resolution of ∼ 2 mm, which is sufficient for general applications. For applications that require higher-precision landmark sensing, the landmark granularity can be improved by increasing the number of electrodes. In addition, the system accuracy is also confined to the performance of the visual recognition model, and therefore using a high-precision visual model (e.g., multi-camera 3D imaging systems [48]) can also help to improve the model's granularity. The FLAME model used in this paper is a generic 3D face model for 3D avatar construction and animation generation. Despite its ease of use, generic face models can be coarse and lack personal details. The animation quality and expressivity can be further improved by involving personalized blendshape models which can be built via user-specific calibration or training process [28, 63].

Reducing Power Consumption. The current prototype can support up to 8.2 hours of continuous usage if paired up with a 500 mAh Li-ion battery. In our future work, we seek to further improve the system's energy efficiency by designing a customized data collection board using a more compact analog-to-digital converter with fewer channels for the biopotential measurements (e.g., ADS1299-4 consumes 43% less power than ADS1299-8 used in the current bio-amplifier circuit [29]).

Potential Applications. Without requiring a camera positioned in front of the user, our system would introduce new opportunities in various emerging applications. For instance, through increasing the awareness of the user's real-time facial expressions and emotional states, our system can enable a more immersive user experience for existing AR/VR applications (e.g., face-to-face interactions), assess student engagement for online courses, and assist with driver fatigue detection to monitor abnormal behaviors, etc. In addition, our system can serve as a silent-speech interface for human-computer interaction. Through performing different facial gestures, people can interact with smart home appliances (e.g., turn down the volume of a smart speaker) and disabilities can control their handicap equipment (e.g., a wheelchair) more conveniently. We plan to develop an API library which is compatible with major AR/VR platforms (e.g., OpenVR) and a mobile app to support various mobile devices in our future work.

9 RELATED WORK

Camera-based Facial Landmark Detection. Conventional camera-based solutions can be categorized as holistic methods, Constrained Local Model (CLM)-based methods, and deep-learning-based methods. Traditional holistic methods [12, 13] detect facial landmarks by iteratively mapping a statistical facial model to the video frames. CLM-based methods [15, 52] build independent local shape models for each landmark, making them more robust to illumination and occlusion. Differently, deep-learning-based methods [17, 60, 65, 66, 69] extract high-level features from images and further learn a mapping to landmark locations via deep learning. However, these solutions require users to face a camera all times without occlusions and under good lighting conditions, which largely restricts their application scenarios. Additionally, users' facial expressions can be detected using a RGB-D camera and strain gauges [36]. A more recent work, C-Face [10], reconstructs facial landmarks using facial contours captured via two ear-mounted cameras. Although it enables more flexible deployment, the system is still constrained by illumination conditions and may raise privacy concerns. In addition, deploying energy-hungry cameras on the wearable device largely limits its battery life.

Speech-driven Facial Animation. Early works [7, 11, 14, 57] utilize hidden Markov model (HMM) to generate speech-driven facial animations. Recent studies show a success of generating facial animations from audio spectrograms using 2D CNN [46] and from raw audio waveform using 1D CNN [21]. Moreover, LSTM-based methods have also been deployed for synthesizing mouth animations [55] or reconstructing full facial landmark [20]. However, these studies are not able to recognize silent facial gestures or expressions while talking (e.g., talking enthusiastically or sadly).

Wearable-sensor-based Facial Movement Classification. Some studies recognize the user's facial movements using wearable sensors. For instance, speech-related movements can be sensed using capacitive sensors [37] or magnetic sensors attached to the tongue surface [9, 51], facial expressions can be identified using smart glasses with piezoelectric sensors [53] or optical sensors [39], and facial gestures can be sensed using earphone microphone [4], or acoustic interferometry [30]. However, all these studies are classification-based methods and cannot be used for continuous 3D facial reconstruction.

Biosensor-based Facial Movement Classification. EMG & EEG signals have been shown effective in distinguishing a limited set of pre-defined facial gestures. Through attaching sensors around the user's eyes and forehead, previous studies can perform 5-class [25], 9-class [49], 10-class [26], 11-class gesture recognition [27], but these systems were not designed for continuous facial tracking. More recently, Matthies *et al.* use tiny biosensors placed inside the ear canal to distinguish a set of 5 facial gestures [40], and Nguyen *et al.* use EMG signals captured behind the user's ears to sense tongue movements [42]. In addition to EMG, a few studies (e.g., [41, 58]) propose to use EOG signals to track eye movements to interact with machines. Similar to all the other wearable-sensor-faced approaches, the aforementioned studies can only identify a small set of facial gestures, and the sensor placement is quite obtrusive in most of these studies.

10 CONCLUSION

In this paper, we propose *BioFace-3D*, the first single-earpiece lightweight biosensing system for continuous 2D facial landmarks tracking and 3D facial animation rendering. We design a novel cross-modal transfer learning framework to leverage high-precision camera sensor to guide the training of the biosensing model. We conducted extensive experiments involving 16 participants under various settings. The results demonstrated that the proposed *BioFace-3D* can accurately track major facial landmarks in a continuous manner with only 1.85 mm average error and 3.38% normalized mean error. The rendered 3D facial animations are smooth, continuous, and highly consistent with the real human facial movements, showing the system's promising capability.

ACKNOWLEDGMENTS

We would like to thank our anonymous shepherd and reviewers for their insightful feedback.

REFERENCES

[1] 2021. Covidien Kendall Disposable Surface EMG/ECG/EKG Electrodes 1" (24mm). https://bio-medical.com/covidien-kendall-disposable-surface-emg-ecg-ekg-electrodes-1-24mm-50pkg.html

[2] 2021. Demo Video for BioFace-3D. https://mosis.eecs.utk.edu/bioface-3d.html

[3] 2021. Monsoon High Voltage Power Monitor. https://www.msoon.com/high-voltage-power-monitor

[4] Takashi Amesaka, Hiroki Watanabe, and Masanori Sugimoto. 2019. Facial expression recognition using ear canal transfer function. In Proceedings of the 23rd International Symposium on Wearable Computers. 1–9.

[5] Anwesha Banerjee, Shreyasi Datta, Monalisa Pal, Amit Konar, DN Tibarewala, and R Janarthanan. 2013. Classifying electrooculogram to detect directional eye movements. Procedia Technology 10 (2013), 67–75.

[6] Open BCI. 2021. Cyton Biosensing Board (8-channels). https://shop.openbci.com/products/cyton-biosensing-board-8-channel?variant=38958638542

[7] Matthew Brand. 1999. Voice puppetry. In Proceedings of the 26th annual conference on Computer graphics and interactive techniques. 21–28.

[8] Sandra Carrasco and Miguel Ángel Sotelo UAH. 2020. D3. 3 Driver Monitoring Concept Report. (2020).

[9] Lam Aun Cheah, James M Gilbert, José A González, Phil D Green, Stephen R Ell, Roger K Moore, and Ed Holdsworth. 2018. A Wearable Silent Speech Interface based on Magnetic Sensors with Motion-Artefact Removal.. In BIODEVICES. 56–62.

[10] Tuochao Chen, Benjamin Steeper, Kinan Alsheikh, Songyun Tao, François Guimbretière, and Cheng Zhang. 2020. C-Face: Continuously Reconstructing Facial Expressions by Deep Learning Contours of the Face with Ear-mounted Miniature Cameras. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology. 112–125.

[11] Kyoungho Choi, Ying Luo, and Jenq-Neng Hwang. 2001. Hidden Markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system. Journal of VLSI signal processing systems for signal, image and video technology 29, 1 (2001), 51–61.

[12] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. 2001. Active appearance models. IEEE Transactions on pattern analysis and machine intelligence 23, 6 (2001), 681–685.

[13] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. 1995. Active shape models-their training and application. Computer vision and image understanding 61, 1 (1995), 38–59.

[14] Darren Cosker, Dave Marshall, Paul L Rosin, and Yulia Hicks. 2004. Speech driven facial animation using a hidden markov coarticulation model. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., Vol. 1. IEEE, 128–131.

[15] David Cristinacce and Timothy F Cootes. 2006. Feature detection and tracking with constrained local models.. In Bmvc, Vol. 1. Citeseer, 3.

[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] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. 2018. Style aggregated network for facial landmark detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 379–388.

[18] Paul Ekman and Dacher Keltner. 1997. Universal facial expressions of emotion. Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture (1997), 27–46.

[19] Rosenberg Ekman. 1997. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA.

[20] Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan. 2018. Generating talking face landmarks from speech. In International Conference on Latent Variable Analysis and Signal Separation. Springer, 372–381.

[21] Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan. 2019. Noise-resilient training method for face landmark generation from speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2019), 27–38.

[22] Avard Tennyson Fairbanks and Eugene F Fairbanks. 2005. Human proportions for artists. Fairbanks Art and Books.

[23] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. 2018. Wing loss for robust facial landmark localisation with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2235–2245.

[24] Taesik Gong, Yeonsu Kim, Jinwoo Shin, and Sung-Ju Lee. 2019. Metasense: few-shot adaptation to untrained conditions in deep mobile sensing. In Proceedings of the 17th Conference on Embedded Networked Sensor Systems. 110–123.

[25] Mahyar Hamedi, Iman Mohammad Rezazadeh, and Mohammad Firoozabadi. 2011. Facial gesture recognition using two-channel bio-sensors configuration and fuzzy classifier: A pilot study. In International Conference on Electrical, Control and Computer Engineering 2011 (InECCE). IEEE, 338–343.

[26] Mahyar Hamedi, Sh-Hussain Salleh, Mehdi Astaraki, and Alias Mohd Noor. 2013. EMG-based facial gesture recognition through versatile elliptic basis function neural network. Biomedical engineering online 12, 1 (2013), 73.

[27] M Hamedi, Sh-Hussain Salleh, TS Tan, K Ismail, J Ali, C Dee-Uam, C Pavaganun, and PP Yupapin. 2011. Human facial neural activities and gesture recognition for machine-interfacing applications. International Journal of Nanomedicine 6 (2011), 3461.

[28] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3d avatar creation from hand-held video input. ACM Transactions on Graphics (ToG) 34, 4 (2015), 1–14.

[29] Texas Instruments. 2020. ADS1299-x Low-Noise, 4-, 6-, 8-Channel, 24-Bit, Analog-to-Digital Converter for EEG and Biopotential Measurements. https://www.ti.com/lit/ds/symlink/ads1299.pdf?ts=1615154540121.

[30] 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. 1–13.

[31] Sasan Karamizadeh, Shahidan M Abdullah, Azizah A Manaf, Mazdak Zamani, and Alireza Hooman. 2013. An overview of principal component analysis. Journal of Signal and Information Processing 4, 3B (2013), 173.

[32] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

[33] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 2021. 1D convolutional neural networks and applications: A survey. Mechanical Systems and Signal Processing 151 (2021), 107398.

[34] Jyoti Kumari, R Rajesh, and KM Pooja. 2015. Facial expression recognition: A survey. Procedia Computer Science 58 (2015), 486–491.

[35] Lumen Learning. 2021. Muscle Contraction and Locomotion. https://courses.lumenlearning.com/ivytech-bio1-1/chapter/muscle-contraction-and-locomotion/

[36] 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.

[37] Richard Li, Jason Wu, and Thad Starner. 2019. TongueBoard: An Oral Interface for Subtle Input. In Proceedings of the 10th Augmented Human International Conference 2019. 1–9.

[38] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. ACM Trans. Graph. 36, 6 (2017), 194–1.

[39] Katsutoshi Masai, Kai Kunze, Daisuke Sakamoto, Yuta Sugiura, and Maki Sugimoto. 2020. Face Commands-User-Defined Facial Gestures for Smart Glasses. In 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 374–386.

[40] Denys JC 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. 1911–1922.

[41] Yunjun Nam, Bonkon Koo, Andrzej Cichocki, and Seungjin Choi. 2013. GOM-Face: GKP, EOG, and EMG-based multimodal interface with application to humanoid robot control. IEEE Transactions on Biomedical Engineering 61, 2 (2013), 453–462.

[42] Phuc Nguyen, Nam Bui, Anh Nguyen, Hoang Truong, Abhijit Suresh, Matt Whitlock, Duy Pham, Thang Dinh, and Tam Vu. 2018. Tyth-typing on your teeth: Tongue-teeth localization for human-computer interface. In Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services. 269–282.

[43] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. 2019. Few-shot adaptive gaze estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 9368–9377.

[44] Xi Peng, Rogerio S Feris, Xiaoyu Wang, and Dimitris N Metaxas. 2016. A recurrent encoder-decoder network for sequential face alignment. In European conference on computer vision. Springer, 38–56.

[45] Patrick Perkins and Steffen Heber. 2018. Identification of ribosome pause sites using a Z-Score based peak detection algorithm. In 2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS). IEEE, 1–6.

[46] Hai X Pham, Yuting Wang, and Vladimir Pavlovic. 2017. End-to-end learning for 3d facial animation from raw waveforms of speech. arXiv preprint arXiv:1710.00920 (2017).

[47] Utsav Prabhu, Keshav Seshadri, and Marios Savvides. 2010. Automatic facial landmark tracking in video sequences using kalman filter assisted active shape models. In European Conference on Computer Vision. Springer, 86–99.

[48] Marcos Quintana, Sezer Karaoglu, Federico Alvarez, Jose Manuel Menendez, and Theo Gevers. 2019. Three-d wide faces (3dwf): Facial landmark detection and 3d reconstruction over a new rgb–d multi-camera dataset. Sensors 19, 5 (2019), 1103.

[49] Iman Mohammad Rezazadeh, S Mohammad Firoozabadi, Huosheng Hu, and S Mohammad Reza Hashemi Golpayegani. 2011. A novel human–machine interface based on recognition of multi-channel facial bioelectric signals. Australasian physical & engineering sciences in medicine 34, 4 (2011), 497–513.

[50] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2013. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In Proceedings of the IEEE International Conference on Computer Vision Workshops. 397–403.

[51] 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.

[52] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. 2011. Deformable model fitting by regularized landmark mean-shift. International journal of computer vision 91, 2 (2011), 200–215.

[53] 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.

[54] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 1145–1153.

[55] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (ToG) 36, 4 (2017), 1–13.

[56] M Emin Tagluk, Necmettin Sezgin, and Mehmet Akin. 2010. Estimation of sleep stages by an artificial neural network employing EEG, EMG and EOG. Journal of medical systems 34, 4 (2010), 717–725.

[57] Lucas D Terissi and Juan Carlos Gomez. 2008. Audio-to-visual conversion via HMM inversion for speech-driven facial animation. In Brazilian Symposium on Artificial Intelligence. Springer, 33–42.

[58] Chun Sing Louis Tsui, Pei Jia, John Q Gan, Huosheng Hu, and Kui Yuan. 2007. EMG-based hands-free wheelchair control with EOG attention shift detection. In 2007 IEEE International Conference on Robotics and Biomimetics (ROBIO). IEEE, 1266–1271.

[59] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, Vol. 1. IEEE, I–I.

[60] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. 2020. Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence (2020).

[61] Delsys: wearable sensors for movement sciences. 2021. How to improve EMG signal quality. https://delsys.com/emgworks/signal-quality-monitor/improve/

[62] John G Webster. 1984. Reducing motion artifacts and interference in biopotential recording. IEEE transactions on biomedical engineering 12 (1984), 823–826.

[63] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime performance-based facial animation. ACM transactions on graphics (TOG) 30, 4 (2011), 1–10.

[64] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. 2018. Look at boundary: A boundary-aware face alignment algorithm. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2129–2138.

[65] Yue Wu, Tal Hassner, KangGeon Kim, Gerard Medioni, and Prem Natarajan. 2017. Facial landmark detection with tweaked convolutional neural networks. IEEE transactions on pattern analysis and machine intelligence 40, 12 (2017), 3067–3074.

[66] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. 2016. Robust facial landmark detection via recurrent attentive-refinement networks. In European conference on computer vision. Springer, 57–72.

[67] Xuehan Xiong and Fernando De la Torre. 2013. Supervised descent method and its applications to face alignment. In Proceedings of the IEEE conference on computer vision and pattern recognition. 532–539.

[68] Uldis Zarins. 2018. Anatomy of Facial Expressions. Exonicus, Incorporated.

[69] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. In European conference on computer vision. Springer, 94–108.

[70] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. 2015. Face alignment by coarse-to-fine shape searching. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4998–5006.