Yi Wu *University of Tennessee, Knoxville* **Vimal Kakaraparthi** *University of Colorado Boulder*
Zhuohang Li *University of Tennessee, Knoxville* **Tien Pham** *University of Texas at Arlington*
Jian Liu *University of Tennessee, Knoxville* **VP Nguyen** *University of Texas at Arlington*

Editors: Nicholas D. Lane and Xia Zhou

BIOFACE-3D:
3D Facial Tracking and Animation via Single-ear Wearable Biosensors

Photo, istockphoto.com

Excerpted from "BioFace-3D: continuous 3d facial reconstruction through lightweight single-ear biosensors" from MobiCom '21: *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking* with permission. https://dl.acm.org/doi/10.1145/3447993.3483252 ©ACM 2021

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, emotion recognition, etc. However, existing camera-based solutions require users to be confined to a particular location and face a camera at all times without occlusions, which largely limits their usage in practice. To overcome these limitations, 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. 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.

Facial landmark tracking and 3D reconstruction are becoming fundamental in various emerging applications that require facial analysis. For instance, facial landmark tracking can be used for driver attentiveness monitoring to detect drowsiness and abnormal behaviors [1]. 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 [2]. Moreover, recognizing facial movements can enable silent-speech interfaces for convenient human-computer interactions [3].

However, traditional camera-based solutions [4] require users to face a camera at all times without occlusions and under good lighting conditions, which largely restricts their application scenarios.

Alternatively, there exist several audio-driven approaches that rely on speech to reconstruct speaking-associated facial movements [5]. However, they neither distinguish between expressions while talking (e.g., talking in an enthusiastic or sad manner) nor can they 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 a user's facial gestures [6, 7]. However, all these studies can only distinguish a small set of pre-defined facial gestures.

To circumvent all the limitations of existing approaches, we provide a wearable biosensing system that can unobtrusively, continuously, and reliably sense entire facial movements, track 2D facial landmarks, and further render 3D facial animations by fitting a 3D head model to the 2D facial landmarks. 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 with surface electrodes attached to a very small area around one side of the user's ear to capture both EMG and EOG bioelectrical signals. These sensor positions ensure the sensing capability of BioFace-3D in providing sufficient information for the entire facial reconstruction, while still remaining at 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.

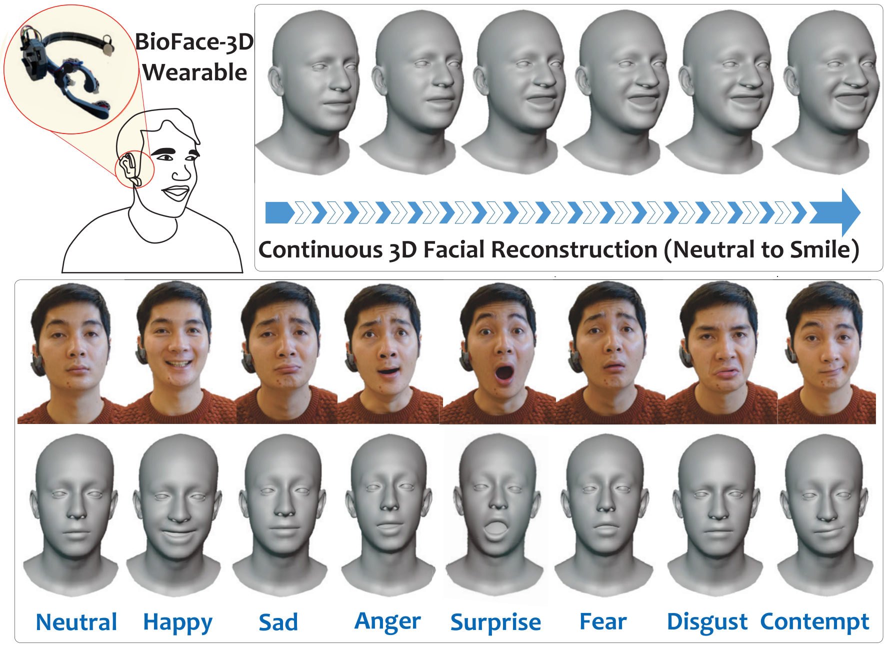

FIGURE 1. Illustration of the reconstructed 3D facial avatar with various facial expressions.

SENSING FACIAL MUSCLE CONTRACTIONS VIA SINGLE-EAR BIOSENSORS

Facial muscles 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. 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. 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. Different facial movements or expressions are produced by the contraction of a different set of facial muscles. This shows the potential of tracking entire facial movements and eye movements through sensing the contraction of corresponding facial muscles and the bioelectrical signals caused by eye movements.

SYSTEM OVERVIEW

As shown in Figure 2, 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 mapping from biosignal stream to facial landmarks, and the *testing phase*, in which the well-trained biosignal network can work independently to perform continuous 3D facial reconstruction without any visual input. Specifically, during training, we collected 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, 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 processed separately 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 groundtruth to guide the training of the biosignal network.

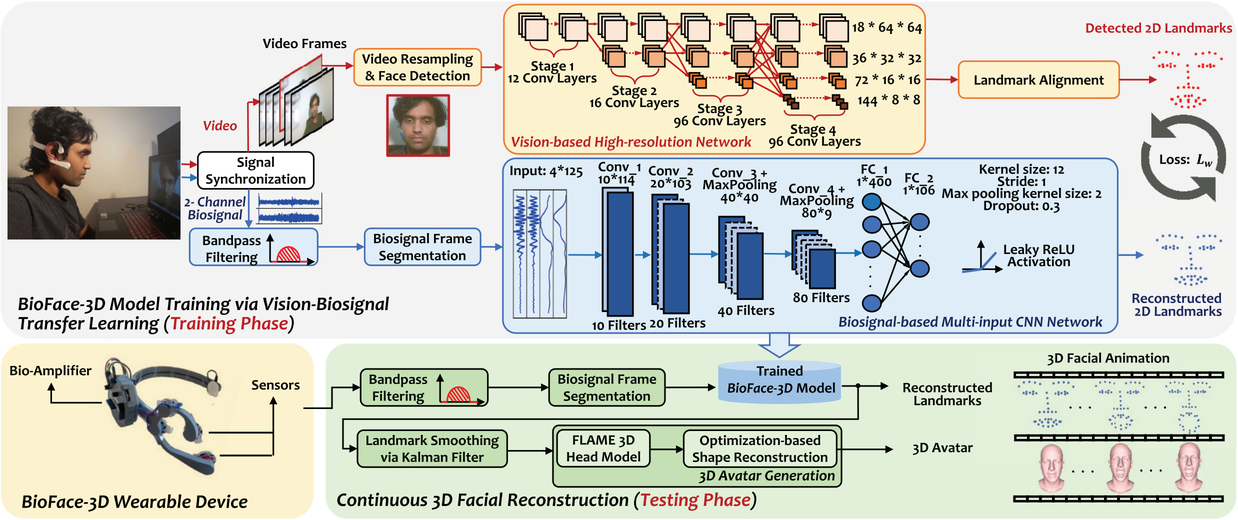

FIGURE 2. BioFace-3D system overview.

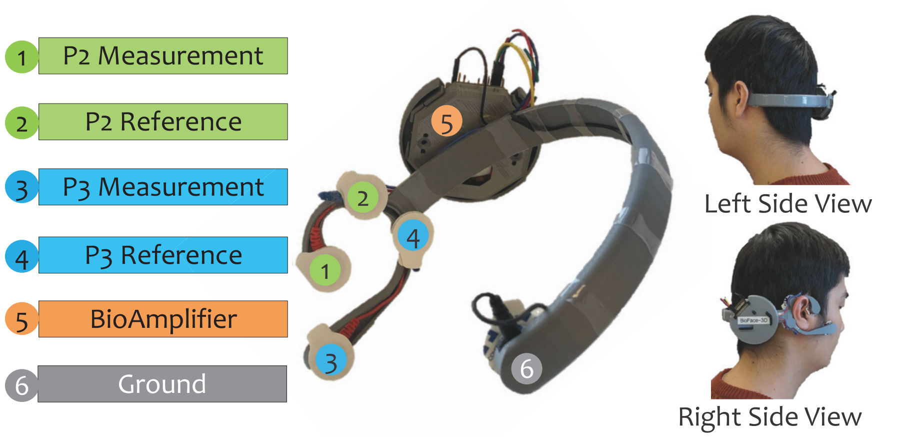

FIGURE 3. BioFace-3D prototype.

Biosignal Stream in Training. BioFace-3D collects two biosignal streams from the biosensors integrated into a single earpiece wearable. Each biosignal stream is first processed to obtain both EOG and EMG biosignal streams via *bandpass filtering*. 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 [8] to enhance the attention of the landmarks, which are important but less active (e.g., pupils) 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 FLAME (Faces Learned with an Articulated Model and Expressions) model [9]. The generated sequence of fitted head models can then be used for rendering a 3D facial animation that recovers the user's facial movements.

WEARABLE IMPLEMENTATION

The BioFace-3D wearable device design is dictated by the most suitable facial locations of measurement electrodes. 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 3. 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.

PERFORMANCE EVALUATION

We recruited 16 participants to evaluate the performance of BioFace-3D. The participants were asked to sit in front of a camera (for training and ground truth recording purposes) and repeatedly perform seven universal expressions while wearing our implemented BioFace-3D prototype. We use Mean Absolute Error (MAE), the absolute error between the reconstructed landmarks and groundtruth landmarks, which are converted from pixels to a physical unit (millimeter), and Normalized Mean Error (NME), the mean error between the groundtruth and reconstructed landmark coordinates, normalized by the interocular distance, as evaluation metrics.

Figure 4 (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., 2.54 mm MAE). Figure 4 (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. Our rendered facial animation samples can be found at [10]. We can observe that BioFace3D 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 all of the facial changes.

CONCLUSION

We propose BioFace-3D, the first single-earpiece lightweight biosensing system for continuous 2D facial landmarks tracking and 3D facial animation rendering. BioFace3D can accurately track major facial landmarks in a continuous manner with mm-level error. The rendered 3D facial animations are smooth, continuous, and highly consistent with the real human facial movements, showing the system's promising capability. ■

Yi Wu is a Ph.D. student in the Department of Electrical Engineering and Computer Science at the University of Tennessee, Knoxville. He received his B.E. and M.S. degrees from the University of Electronic Science and Technology of China and Rutgers University, respectively. His research interests include mobile sensing and cybersecurity.

Vimal Kakaraparthi is a Ph.D. student in the Department of Computer Science at University of Colorado Boulder. His research interests involve sensor systems, machine learning, and autonomous control. He has a Master's and a Bachelor's degree in Mechanical Engineering from University of Colorado Boulder and Manipal Institute of Technology, respectively. During this project, he was advised by Dr. Nguyen at UTA.

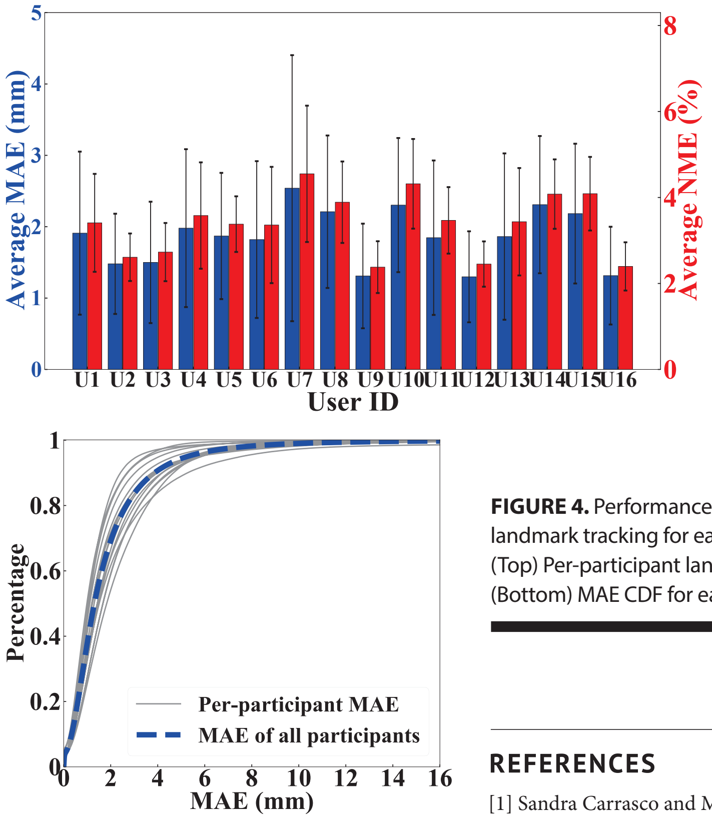

FIGURE 4. Performance of continuous facial landmark tracking for each participant. (Top) Per-participant landmark tracking error. (Bottom) MAE CDF for each participant.

Zhuohang Li is a Ph.D. student in the Department of Electrical Engineering and Computer Science at the University of Tennessee, Knoxville. He received his M.S. in Computer Engineering from Rutgers University. His research interests include cybersecurity, machine learning, and intelligent systems.

Tien Pham is a Ph.D. student in the Department of Computer Science and Engineering at University of Texas at Arlington. He received his B.E degree from Ho Chi Minh City University of Technology, Vietnam. His research focuses on wearable sensing in healthcare, robotics and human-computer interaction.

Jian Liu is an assistant professor in the EECS department at the University of Tennessee, Knoxville. He received his Ph.D. from Rutgers University in 2019. His research focuses on mobile sensing, mobile security, and trustworthy AI and machine learning. He has won multiple awards, including the best paper awards at IEEE SECON 2017 and IEEE CNS 2018.

VP Nguyen is an assistant professor of Computer Science and Engineering at University of Texas at Arlington. He received his Ph.D. from University of Colorado Boulder in 2018. His research interests are mobile and wearable computing, mobile health, and mobile system security. He is the recipient of multiple awards, including SONY Faculty Innovation Award, CACM Research Highlights 2020, and ACM SIGMOBILE Research Highlights 2017-2020.

REFERENCES

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

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

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

[4] Yue Wu, and Qiang Ji. 2019. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127.2 (2019): 115-142.

[5] 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.

[6] Chun Sing Louis Tsui, et al. 2007. EMG-based hands-free wheelchair control with EOG attention shift detection. 2007 IEEE International Conference on Robotics and Biomimetics (ROBIO). *IEEE.*

[7] 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, 453–462.

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

[9] 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, 194–1.

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