International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2025)
J. A. Jorge and N. Sakata (Editors)

Facial Expression Reconstruction with Photo-Reflective Sensors Embedded in a Head-Mounted Display

Yuki Nakabayashi[1], Fumihiko Nakamura[2], Katsutoshi Masai[3] and Maki Sugimoto[1]

[1]Keio University, Japan [2]Ritsumeikan University, Japan [3] Kyushu University, Japan

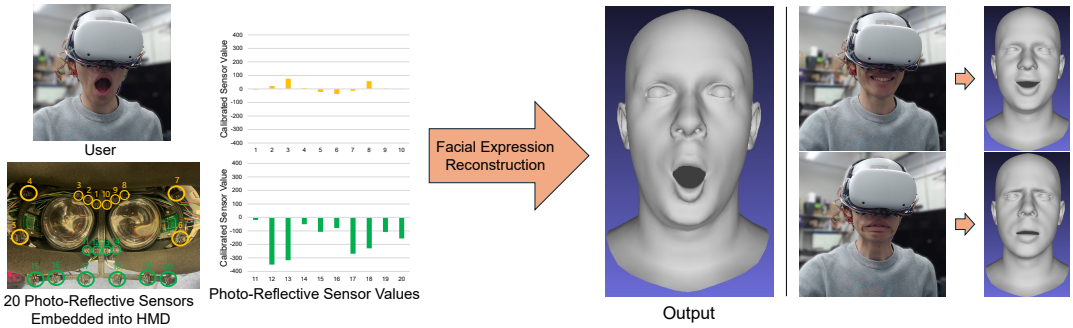

Figure 1: *Our system reconstructs facial expressions of an HMD wearer with multiple photo-reflective sensors integrated into the device.*

Abstract

Reconstructing the 3D facial expressions of head-mounted display (HMD) wearers is essential for natural avatar communication in virtual reality (VR). Camera-based methods achieve high fidelity but involve heavy processing and privacy risks, whereas non-imaging sensors are lightweight and privacy-preserving but provide only sparse features. We propose a reconstruction system that learns high-dimensional 3D facial representations from camera images during training, but performs inference using only compact photo-reflective sensors embedded in the HMD. This design integrates the expressiveness of camera-based supervision with the efficiency and privacy of sensor-based operation. Experimental results show that our method accurately reconstructs 3D facial expressions from the sensor data, training with diverse wearing conditions is more effective than collecting more data under a single condition, and accuracy further improves with a dedicated mouth-shape predictor and lightweight personalization using small wearer-specific datasets.

CCS Concepts

• ***Human-centered computing*** → ***Virtual reality;*** *Interaction devices;*

1. Introduction

Communication in VR has been growing rapidly, and nonverbal information is as important in these environments as it is in the real world. In particular, facial expressions are crucial for nonverbal communication, as they can convey emotions and intentions. VR platforms currently provide simplified features such as preset facial expressions and basic avatar gestures, which limit the richness of interaction. To overcome these limitations, ongoing research aims to reproduce nonverbal communication in ways closer to real-world interaction.

Existing approaches to capturing facial expressions in VR can be broadly divided into image-based and non-imaging sensor–based methods. Image-based approaches embed cameras inside the HMD to record facial images and reconstruct facial structures [WSS*19, TZS*18, OLSL16, CLX*22]. Although these methods can achieve high-fidelity representations, they require heavy image processing during both training and inference, and they raise privacy concerns. By contrast, non-imaging sensor approaches use electromyography (EMG) sensors [LWN*20] and photo-reflective sensors [SNO*17, NMS*22, NS23]. This sensor-based strategy reduces processing cost and alleviates privacy risks. However, reconstruct-

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.2312/egve.20251347

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

ing high-dimensional data (e.g., 3D facial geometry) remains difficult, and many studies therefore simplify the task by classifying sensor signals into a few predefined expressions [PHS17, RLM*19, MWUN21, MSO*16, MKS*20, SLG*22], often based on Ekman's six basic emotions [Ekm89]. This simplification further limits the rich nonverbal communication in VR.

To address these challenges, we propose a system that bridges the gap between camera-based and sensor-based approaches. Our method leverages camera data during training to learn rich, high-dimensional 3D facial representations, while relying solely on compact and inexpensive photo-reflective sensors during inference. In this way, the system leverages the expressive accuracy of camera-based methods, yet maintains the low processing cost, small device footprint, and privacy preservation of sensor-based approaches. We explore whether photo-reflective sensor values can support accurate 3D facial reconstruction when images are used only during training, and examine how factors such as training data scale, data diversity, and model design affect this accuracy.

2. Related Work

2.1. 3D Face Reconstruction

3D face reconstruction has traditionally relied on computer vision techniques applied to full-face images, converting features into 3D model parameters. Approaches include marker tracking [Wil06, HCTW11], landmark detection [BV03, CWLZ13], pixel- or edge-based analysis [RV05], and deep learning with Convolutional Neural Networks (CNNs) [FFBB21, DBB22]. Depth cameras have also been employed for capturing fine-grained geometry [BWP13, LYYB13, WBLP11]. However, such methods are less effective for HMD users because the device occludes the upper face.

To address this limitation, several works have explored capturing partial facial images with embedded or external cameras. Examples include cheek contour images from ear-mounted cameras [CSA*20], infrared cameras placed below the head [CLT*21], and eye-region images from gaze-tracking cameras inside HMDs [HDS*19]. Similarly, cameras integrated inside HMDs capture partial facial views for reconstructing 3D models [WSS*19, TZS*18, OLSL16, CLX*22, JDlTS*22], while RGB-D cameras are used for reconstructing exposed regions of the lower face [CLX*22, LTO*15]. These methods achieve high-precision reconstructions but require intensive computation, add hardware cost, and raise privacy concerns due to image capture.

2.2. Facial Expression Recognition Using Non-Imaging Sensors in Head-Worn Devices

To reduce computational costs and enhance privacy protection, many studies have explored facial expression recognition in wearable devices without relying on cameras. Approaches employ a wide variety of sensors, including EMG [PHS17], EOG [RLM*19], microphones and speakers [LZC*24], electrodes [MWUN21], photo-reflective sensors [MSO*16, MKS*20, AMSS17], and even piezoelectric or accelerometer sensors for detecting jaw motion [SLG*22]. Other explorations include air pressure sensors [AKST17] and electric field sensors [MSU17], which infer expressions from physiological signal changes associated with facial movements. These methods primarily adopt classification models due to the low-dimensional nature of sensor data.

For HMD users, additional modalities have been investigated, including strain gauges [LTO*15], EMG [LWN*20], EEG and EDA [BYJM18], audio and gaze [RLM*21], ultrasonic transducers [IZB*19], and photo-reflective sensors [SNO*17, NMS*22, NS23]. These sensors function within the low-light environment of an HMD and provide enhanced privacy protection by avoiding direct image capture.

Among these modalities, photo-reflective sensors have emerged as particularly promising due to their compact size, low cost, non-contact nature, and ability to capture subtle skin deformations and even subtle smiles [MPHS*22]. They have been integrated into wristbands [OSM*13], earpieces [BBP*15], glasses [MSO*16, MKS*20], masks [TTU*20], and HMDs [SNO*17], demonstrating versatility in detecting skin deformations and facial movements for gesture and expression recognition.

Despite these advances, the absence of cameras makes it difficult to capture a user's facial appearance directly, and most sensor-based systems remain limited to classification. Reconstructing high-dimensional facial geometry from low-dimensional sensor data has been explored using ground truth from facial images or motion capture, but such approaches typically rely on external recording setups and are not always practical for HMD scenarios. In contrast, our approach bridges these methods by leveraging camera images only during training, relying on sensor data for efficient and privacy-preserving inference.

3. Estimating 3D Facial Geometry of HMD Wearers with Photo-Reflective Sensors

Our design principle is to combine the strengths of camera- and sensor-based approaches: cameras provide high-dimensional supervision during training, while sensors enable lightweight, privacy-preserving inference, where privacy is ensured by avoiding any use of identifiable facial imagery during system operation. Since directly mapping low-dimensional sensor data to high-dimensional facial geometry is difficult due to the large dimensionality gap and nonlinear, ambiguous relationships, we introduce a parametric 3D face model as an intermediate representation. This intermediate is expressed as a compact set of model parameters, derived from optimized high-dimensional facial geometry. A core subset captures the primary structure of facial expressions and jaw posture, and once these are estimated, other dependent parameters can be inferred more reliably. In this framework, low-dimensional sensor values are mapped to the core parameters, which then drive the reconstruction of full 3D facial geometry.

To realize this design, we employ photo-reflective sensors as the non-imaging modality. These sensors emit infrared light and measure its reflected intensity, which allows them to estimate the distance to the skin and thus capture subtle skin deformations associated with facial movements [SNO*17, MSO*16, AMSS17]. We selected photo-reflective sensors because they are compact, inexpensive, non-contact, and enable high-speed (real-time) process-

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

ing. These characteristics make them highly practical for integration into everyday HMDs, fulfilling our design goal of a lightweight and efficient inference system.

3.1. System Overview

An overview of our system is shown in Fig. 2. We developed a training HMD equipped with three RGB cameras and 20 photo-reflective sensors. The cameras capture partial facial images for generating ground truth parameters for constructing a 3D facial shape, while the sensors measure skin deformations on the upper face and around the mouth. After learning a mapping between sensor values and parameters derived from the camera images, sensor values are first calibrated using a neutral baseline and then input into two predictors: one for expression parameters and one for jaw pose parameters. The expression predictor uses all 20 sensors, while the jaw pose predictor focuses on sensors around the mouth (11–20). Once trained, these predictors estimate parameters that are applied to a default 3D face model to reconstruct the wearer's expressions.

Since directly predicting a dense 3D mesh from low-dimensional sensor values is impractical, we adopt a parametric face model. Specifically, we use FLAME (Faces Learned with an Articulated Model and Expressions) [LBB*17], which was built from over 33,000 aligned scans. FLAME represents shape, expression, and pose in independent low-dimensional spaces, making it compact yet expressive enough to capture facial variations. In this study, we estimate expression parameters and the jaw pose parameter, as head orientation and eye movements are difficult to infer reliably from photo-reflective sensor data.

3.2. Custom HMD for Data Collection and Tracking

We developed a custom HMD equipped with three RGB cameras and 20 photo-reflective sensors for data collection (Fig. 2, left bottom). The cameras provide partial facial images for generating ground truth parameters, while the sensors capture skin deformations from the upper face and mouth regions. All components were mounted on a 3D-printed bracket and synchronized for simultaneous acquisition. Because images are required only during training, we built two versions of the device: a training HMD with cameras and sensors, and a lightweight tracking HMD with sensors only (Fig. 3). The training HMD is used to establish the mapping from sensor values to 3D facial parameters, whereas the tracking HMD performs efficient real-time reconstruction using sensor data alone. 20 photo-reflective sensors (SG-105F) weigh approximately 0.8 g in total. This corresponds to less than a 1 % increase over a typical commercial HMD. Therefore, the additional hardware imposes negligible physical or psychological load, allowing users to communicate naturally in virtual environments without discomfort when using tracking HMD.

To capture both the eyes and eyebrows, we mounted three Raspberry Pi cameras (160° FOV) on the training HMD: two eye cameras connected to a Raspberry Pi 5 and one mouth camera connected to a Raspberry Pi 3B (Fig. 3). The eye cameras were fixed at symmetric angles on a custom 3D-printed bracket to include both the eyes and eyebrows, while the mouth camera was attached above the Quest 2 headset to cover the entire mouth region. Two Raspberry Pi units were secured to the top of the HMD using a custom mount.

The system also incorporated 20 photo-reflective sensors (SG-105F) arranged as shown in Fig. 2. Sensors 1–10 (yellow) measured the upper face, and sensors 11–20 (green) measured the lower face. The layout was based on Nakamura et al. [NMS*22], with four additional sensors added around the mouth to improve detection accuracy. Sixteen sensors were connected to an Arduino Nano via a multiplexer (CD74HC4067), and the remaining four were directly wired to the Arduino's signal pins. The circuit board holding the Arduino, multiplexer, and connectors was 3D-printed and mounted at the back of the headset (Fig. 3).

3.3. Generating 3D Face Model for Expression Reconstruction

To obtain 2D facial landmarks from HMD images, we combined landmarks from three embedded cameras and aligned them to a frontal view (Fig. 4). Each eye image was annotated with 11 landmarks and the mouth image with 25 landmarks. The annotated landmarks $\boldsymbol{P}_{\text{image,h}}$ were transformed into the camera coordinate system using intrinsic calibration: $\boldsymbol{P}_{\text{camera,h}} = \boldsymbol{P}_{\text{image,h}} \cdot \boldsymbol{K}^{-1}$, $\boldsymbol{P}_{\text{camera}} = z_{\text{value}} \cdot \boldsymbol{P}_{\text{camera,h}}$.

Because all landmarks from the same image shared the same z-coordinate, the eye landmarks appeared horizontally compressed when viewed from the front. We corrected this by expanding their z-coordinates with a fixed scaling factor. The transformed eye and mouth landmarks were then unified into the coordinate system of the mouth camera using precomputed extrinsic parameters: $\boldsymbol{P}_{\text{e,m}} = \boldsymbol{R}^{\top}(\boldsymbol{P}_{\text{e,e}} - \boldsymbol{t})$, where $\boldsymbol{R}$ and $\boldsymbol{t}$ are the rotation and translation from the eye to the mouth camera. Since the nasal dorsum and nasion were not visible, we interpolated nose landmarks along the line between the midpoint of the eyes and the central points below the nose. Finally, we refined the landmarks using expressionless faces captured without the HMD. The difference vectors between these FullFace landmarks and the aligned landmarks were stored and applied to all other expressions, yielding natural-looking 2D landmarks consistent with a frontal view.

To lift the 2D landmarks to 3D, we trained a depth predictor to estimate z-coordinates. Using FLAME [LBB*17], we randomly sampled parameters to generate synthetic 3D face models and extracted corresponding 2D–3D landmark pairs. A dataset of 200,000 pairs was created, with 2D landmarks as input and 3D z-coordinates as output. The predictor was implemented as a CNN with two 1D convolutional layers, which effectively captured spatial relations among landmarks.

Finally, the estimated 3D landmarks were fitted to the FLAME model to reconstruct the 3D facial shape. Optimization was performed in two stages: rigid fitting to align global position and orientation, followed by non-rigid fitting to adjust expressions and jaw posture. We adopted the dog-leg method [LA05, Pow70] for nonlinear optimization. The resulting FLAME expression parameters $\boldsymbol{\psi}_{GT}$ and jaw pose parameters $\boldsymbol{\theta}_{GT}$ served as Ground Truth for training the sensor-based predictors. To learn this mapping, we trained supervised predictors to estimate FLAME parameters directly from photo-reflective sensor values. Sensor readings $\boldsymbol{s}_{\text{expr}}$

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

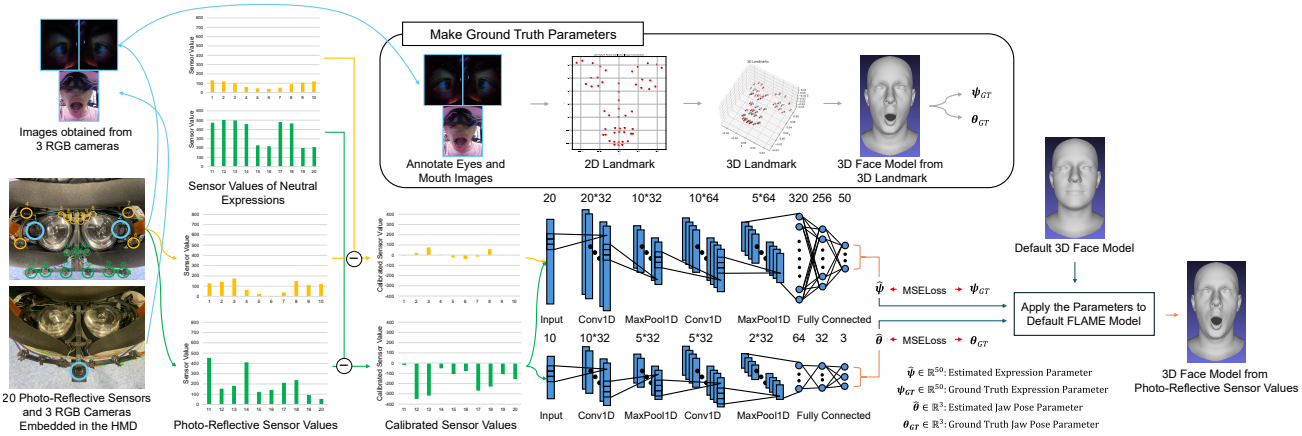

Figure 2: *Our HMD collects three partial facial images (indicated by blue frames), sensor values from the upper face (yellow), and sensor values from the area below the nose and around the mouth (green). Our system transforms landmarks into a planar coordinate system, estimates their 3D positions with a depth predictor, and fits the FLAME model. The resulting parameters $\boldsymbol{\psi}_{GT}$ and $\boldsymbol{\theta}_{GT}$ serve as Ground Truth for training a predictor, which uses photo-reflective sensor values as input. The predictor estimates the parameters $\hat{\psi}$ and $\hat{\theta}$, compares them with $\boldsymbol{\psi}_{GT}$ and $\boldsymbol{\theta}_{GT}$ applying MSELoss, and learns from sensor values. After training, the predictor outputs $\hat{\psi}$ and $\hat{\theta}$, which reconstruct expressions in the FLAME model.*

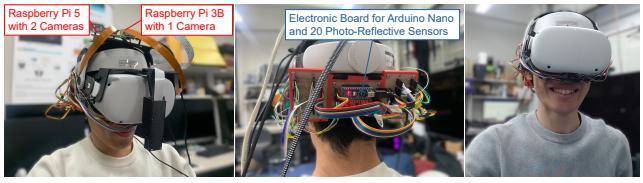

Figure 3: *(Left) Front View of the Training HMD Wearer. (Middle) Rear View of the Training HMD Wearer. (Right) User Wearing the Tracking HMD. During training, two Raspberry Pis with cameras are attached to the HMD to capture images. However, during tracking, only photo-reflective sensor values are used, and images are not needed, so the Raspberry Pis with cameras are removed.*

were baseline-normalized using neutral-expression values $\boldsymbol{s}_{\text{neutral}}$: $\boldsymbol{s}_{\text{norm}} = \boldsymbol{s}_{\text{expr}} - \boldsymbol{s}_{\text{neutral}}$.

Two CNN-based predictors were used: one for expression parameters with all 20 sensors as input, and one for jaw pose using only the 10 sensors around the mouth. This separation allowed efficient modeling of fine-grained mouth movements. Considering spatial relationships among sensors, both predictors were implemented with convolutional layers. At inference, the predictors output $\hat{\psi}$ and $\hat{\theta}$, which were applied to the FLAME model to reconstruct the wearer's facial expressions in real time.

4. Experiment

4.1. Target Facial Expressions

Participants wore the training HMD and were asked to reproduce 30 facial expressions (Fig. 5). Expression 1 was neutral, Expressions 2–15 corresponded to Ekman's six basic emotions (Happy, Surprised, Sad, Angry, Disgust, Fear) [Ekm89] plus Contempt [Mat92], and were used as the main targets for evaluation. Expressions 16–30 were combinations covering additional facial muscle movements not included in the basic emotions, ensuring a broader range of motions for training. All images in Fig. 5 were generated using DALL·E 3 [†], and their correspondence to the intended expressions was manually verified by the authors.

4.2. Data Collection Procedure

Twenty participants (16 male, 4 female, all in their twenties) took part in the study. The study protocol was approved by the institutional review board (IRB), and all participants provided informed consent prior to data collection. Participants sat on a chair, viewed a virtual environment through the HMD, and reproduced 30 facial expressions presented inside the display (see Fig. 6 for the experimental setup). Once each participant indicated readiness by tapping the desk, the experimenter initiated the data logging while the expression was maintained. For each expression, the system recorded three images from the two eye cameras and the mouth camera, along with 20 photo-reflective sensor values, at 0.5-second intervals repeated three times. One full cycle of 30 expressions was defined as a set, and each participant completed 12 sets, yielding 1080 samples per participant ($3 \times 30 \times 12$). To examine the effect of reattaching the HMD, participants removed and re-wore the device every four sets, creating three attachment conditions (sets 1–4, 5–8, and 9–12).

[†] https://openai.com/index/dall-e-3/

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

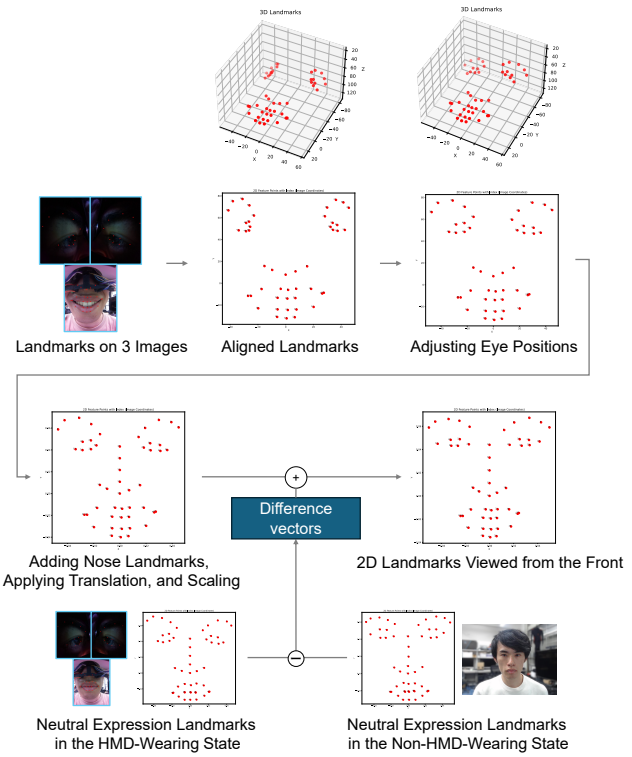

Figure 4: *Flow for Generating 2D Landmarks (used as ground truth for the estimator), Viewed from the Front of the Wearer from Landmarks Annotated on 3 Images*

Figure 5: *Generated Facial Expression Images using DALL·E 3. for Data Collection.*

4.3. Data Collection System

Data acquisition was synchronized by sending trigger signals from Unity to the devices in sequence, starting with the slowest (Raspberry Pi 3B for the mouth camera), followed by the Raspberry Pi 5 for the eye cameras, and finally the Arduino Nano for the photo-reflective sensors, so that uncontrolled processing delays (jitter) were minimized. The average delay across participants was 0.250 seconds, with 0.186 seconds between sensors and eye cameras, and 6.37×10^{-2} seconds between mouth and eye cameras.

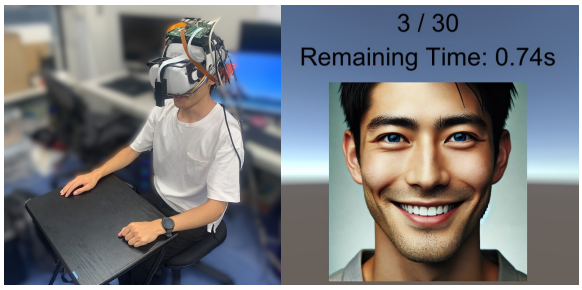

Figure 6: *(Left) Setup for Data Collection. (Right) Scene Displayed Inside the HMD During Data Collection.*

4.4. Experimental Conditions

Full-Model Training We evaluated facial expression reconstruction under seven training conditions (Table 1). Here, **set** indicates the total number of training cycles (each cycle containing 30 expressions) used, and **wear** indicates the number of distinct HMD re-attachment conditions from which that data was collected. The within-user conditions (3wear,11sets; 3wear,3sets; 2wear,8sets; 1wear,4sets; 1wear,1set) used only the participant's data, while the cross-user condition (0set) used 228 sets from 19 other participants. The final condition (3wear,11sets,Single Predictor) examined the effect of using a single predictor for both expression and jaw pose parameters, while the other conditions employed two separate predictors.

Table 1: *Experimental Conditions.*

Condition	Training Data	Test Data
3wear, 11sets	11sets	1set
3wear, 3sets	3sets (different attachment states)	9sets
2wear, 8sets	8sets (2 attachment states)	4sets
1wear, 4sets	4sets (same attachment state)	8sets
1wear, 1set	1set	11sets
0set (from other users)	228 sets from 19 participants	12sets
3wear, 11sets, Single Predictor	11sets	1set

Fine-Tuning To address scenarios with limited data from the HMD wearer, we applied a fine-tuning approach, which was also explored for photo-reflective sensing methods [NS23, MSI24]. A pre-trained model was first trained on data from 19 participants (0set condition in Table 1). Fine-tuning then updated only the final fully connected layers of the expression and jaw pose predictors, while earlier layers were frozen. We evaluated three settings using the wearer's data: 1set, 2sets, and 3sets. To examine the effect of reattachment, the 2set condition compared one versus two attachment states, and the 3set condition compared one versus three.

4.5. Evaluation Method

We evaluated reconstruction accuracy by comparing corresponding vertices $\boldsymbol{v}_i$ between the ground-truth 3D facial shape (from images) and the reconstructed shape (from sensor values). Two error metrics were used: root mean squared error (RMSE) and mean Euclidean distance. Before error calculation, the reconstructed mesh

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

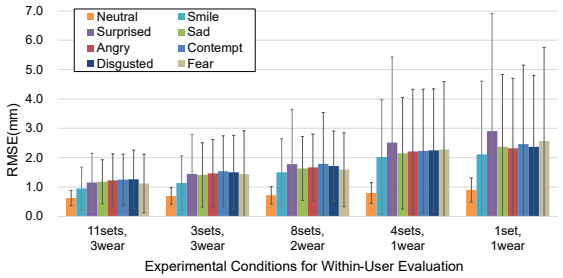

Figure 7: *Reconstruction errors (RMSE) under Within-user Conditions. Error bars indicate standard deviation across participants.*

was aligned to the ground-truth mesh over the entire face using the Iterative Closest Point (ICP) algorithm [BM92]. ICP iteratively optimizes rotation and translation to minimize vertex distances, ensuring that the reported errors reflect differences in facial expression rather than head position or orientation.

Evaluation targeted expressions 1–15 in Fig. 5. These expressions were selected for quantitative analysis to facilitate comparison with prior work. Expressions 16–30, consisting of combined or less common configurations, were included in training but excluded from evaluation.

5. Results

All machine learning and error calculations were performed on a computer with an NVIDIA GeForce RTX 3080 GPU and an AMD Ryzen 9 5900X CPU, implemented in Python 3.8.8 and PyTorch 2.2.0+cu121 on Ubuntu 22.04.3 LTS. The average processing time from sensor input to 3D facial shape generation was 23.9 ms across all conditions.

5.1. Within-User Evaluation

The within-user evaluation results are shown in Fig. 7. Accuracy improved as the number of reattachments in the training data increased, and for a fixed number of reattachments, more training data further reduced errors. A comparison of 3wear,3sets and 2wear,8sets revealed that reattachment count had a greater effect on accuracy than total data volume. Variability also decreased with more reattachments. Among individual expressions, Surprised (expression 5 in Fig. 5) showed the largest error. Qualitative results are presented in Fig. 8 (left), using Surprised as an example. Errors were most evident around the mouth, especially at the lower lip and jaw where sensors provide limited coverage, while the eye and eyebrow regions were generally more accurate. In some cases, however, eyebrow errors were more pronounced (e.g., User 2). Training times were short: for 11 sets, the expression predictor averaged 3.39 s and the jaw predictor 3.83 s.

5.2. Cross-User Evaluation

The cross-user evaluation results are shown in Fig. 9. Excluding the participant's own data from training led to higher errors than within-user conditions, despite the 19 times larger training set. However, variance across participants was smaller in the cross-user case. Among individual expressions, Surprised again showed the largest error due to pronounced eye and mouth movements. Qualitative results are presented in Fig. 8 (middle). Compared to within-user conditions, errors were larger in the mouth and jaw regions, often resulting in smaller reconstructed mouth openings relative to the ground truth. Training times were longer than in the within-user case, averaging 56.76 s for the expression predictor and 67.51 s for the jaw predictor.

5.3. Single vs. Multi-Predictor Evaluation

The results are shown in Fig. 10. Separate predictors yielded slightly lower errors overall and smaller variance across participants compared to a single predictor, though both conditions showed similar error trends across expressions. Qualitative results are presented in Fig. 8 (right). Error locations were similar between the two conditions, but in some cases, the single-predictor approach produced smaller errors around the mouth and cheeks. The average training time for the single predictor was 8.58 s.

5.4. Fine-Tuning Evaluation

Results are shown in Fig. 11. Fine-tuning reduced errors relative to the cross-user baseline (0set), with the largest improvement observed when as few as a single set of wearer-specific data (1set) was included. This brought performance closer to the within-user level. Additional sets provided smaller gains, and error reduction was driven more by reattachment diversity than by data volume. Qualitative results (Fig. 12) confirm that most of the improvement occurred from 0set to 1set, with little change beyond. Errors in the lower lip and jaw persisted even after fine-tuning. Training times were modest: 4.30 s (expression) and 2.32 s (jaw) for 1set, increasing to 13.08 s and 5.80 s, respectively, for 3set.

5.5. Comparison with State-of-the-Art

We compared our method with prior studies on 3D face reconstruction for HMD wearers using vertex-to-vertex error metrics. For a fair comparison, we report results under the 0set condition (228 sets from other users), where no data from the target participant was included in training.

Our method achieved an RMSE of 2.38 ± 0.88 mm, outperforming FaceVR [TZS*18], which reported 3.38 mm. In terms of mean Euclidean distance, our method yielded 1.67 ± 0.57 mm, better than Mask-off [ZXC*19] at 2.93 mm, though slightly higher than MIA (Multi-Identity Architecture) [JDlTS*22], which achieved 1.51 ± 0.23 mm with training data from 120 HMD wearers. These results indicate that photo-reflective sensor–based reconstruction achieves performance comparable to camera-based methods in accuracy, while retaining the advantages of efficiency and privacy during inference. These advantages make our sensor-based design highly practical and robust for everyday VR use.

6. Discussion

Our experiments showed that reconstruction accuracy depends more on the diversity of HMD attachment patterns than on the total volume of training data. This aligns with prior work reporting

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

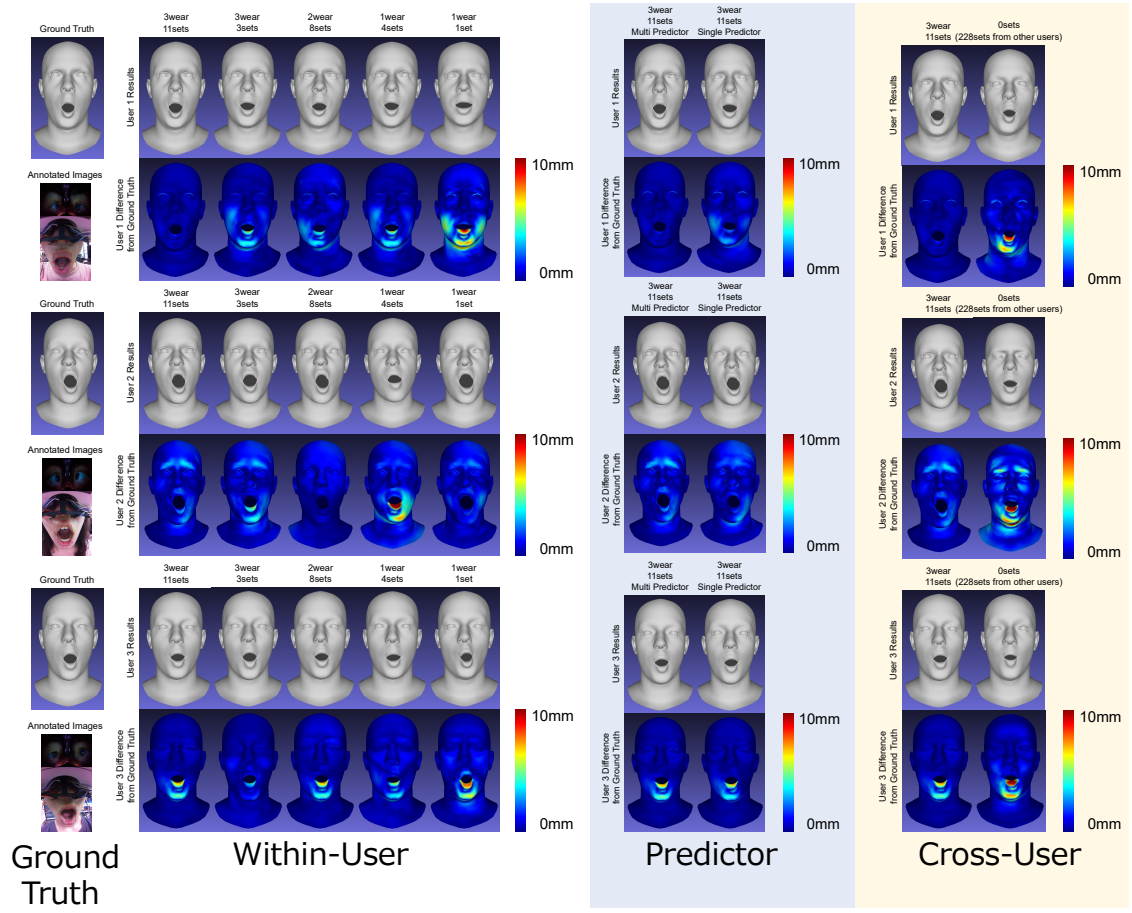

Figure 8: *Qualitative Comparison of Reconstructed 3D Faces under Within-User/Cross-User/Different Predictor Conditions.*

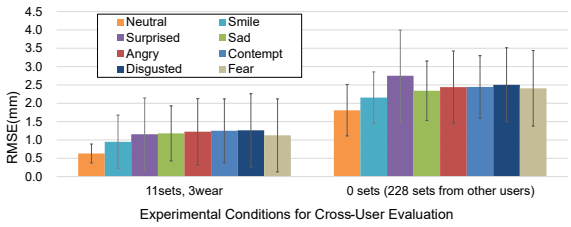

Figure 9: *Reconstruction Errors under Cross-User Conditions.*

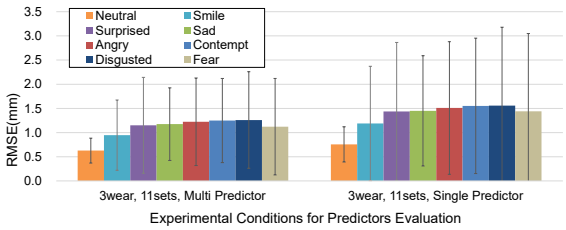

Figure 10: *Reconstruction Errors under two Predictor Conditions.*

that reattachment strongly affects recognition accuracy [SNO*17, NS23, MSO*16]. Reattachment shifts the sensor-to-face position, so even identical expressions can produce different sensor readings. Including varied attachment conditions in training therefore improves robustness and accuracy. By contrast, simply increasing data volume had little effect, likely because participants reproduced the same expressions in response to identical prompts. Future datasets that incorporate varied prompts or allow more natural expression changes may make additional training data more beneficial.

Errors were larger around the mouth than in other regions, mainly due to two factors. First, mouth movements are larger and more complex than those of the eyes or eyebrows, making them harder to estimate. Consistent with this, using a dedicated jaw-pose predictor improved accuracy (Fig. 10). Second, our device only measured the upper part of the mouth, leaving the sides and lower lip under-sampled, which likely contributed to higher errors.

Fine-tuning with limited wearer-specific data improved accuracy, but further data contributed little additional improvement, indicating that minimal personalization suffices. However, expressions involving large mouth movements (e.g., Surprised, Angry,

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

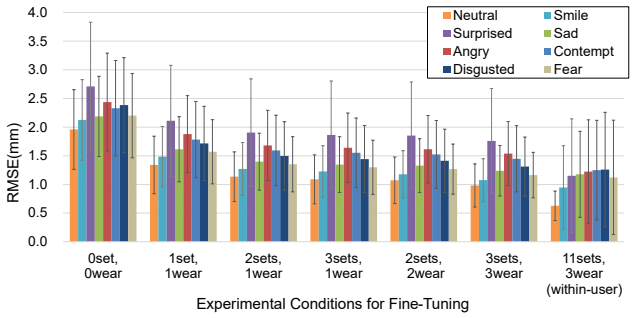

Figure 11: *Reconstruction Errors under Fine-Tuning Approaches. The leftmost condition, "0set,0wear", corresponds to the cross-user condition "0set (228sets from other users)" in Fig. 9, while the rightmost condition, "11sets,3wear (within-user)", matches the within-user condition "11sets,3wear" in Fig. 7.*

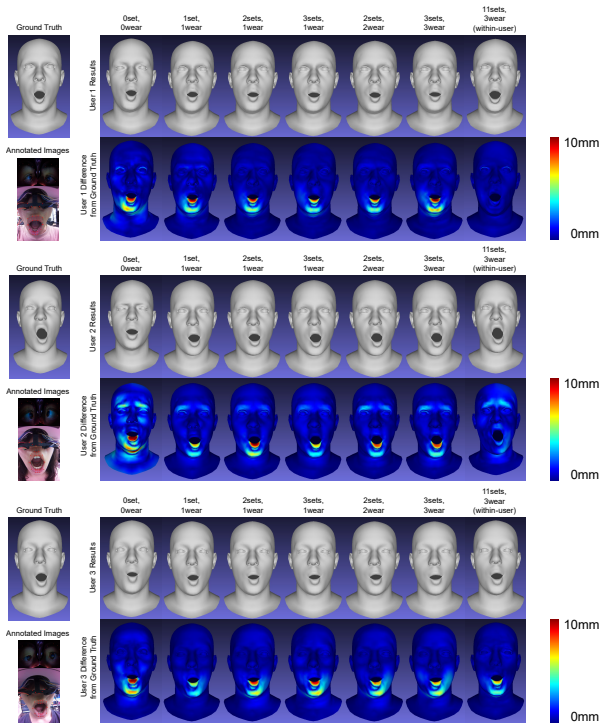

Figure 12: *Qualitative Comparison of Reconstructed 3D Faces under Different Fine-Tuning Conditions. The leftmost model (0set,0wear) for each user was trained without any data from the wearer, while the rightmost model (11sets,3wear) was trained exclusively on the wearer's data.*

Contempt, and Disgusted) still resulted in larger errors and higher standard deviations, even after fine-tuning, likely because subtle individual differences in mouth opening strongly affect sensor values.

7. Limitations and Future Work

In this study, ground truth 3D facial shapes were generated from images captured by three cameras, but we did not assess how closely the generated ground truth matched the user's actual 3D facial shape. While the ground truth generation pipeline provides sufficient consistency and fidelity for learning by combining established models, future work should employ high-resolution 3D scans to further improve training accuracy and reduce potential errors. Another challenge is HMD reattachment: when the device is worn again, the relative sensor positions may shift, degrading accuracy. Possible solutions include designing more stable mounts or using additional sensors to estimate shifts for compensation.

This study also has broader limitations. The dataset covered only 30 static expressions from 20 young participants, and sensor placement insufficiently captured the lower jaw, causing larger errors for wide mouth movements. Future work should include more varied expressions, participants, and sensor coverage. To further improve the reconstruction of the lower facial region, we plan to extend sensor placement toward the jawline and sides of the face, following approaches such as Asano et al. [AMSS17], which captured mandibular motion by placing reflective sensors near the temporomandibular joint on the sides of the face. Another promising direction is to incorporate biomechanical dependencies between the upper and lower facial regions. Since jaw motion mechanically influences the surrounding cheek and chin areas, modeling these physical constraints could enhance the plausibility and stability of the reconstructed expressions with limited sensor coverage.

For practical applications, it is desirable to develop a predictor independent of user-specific data. Our method still relies on camera images during training, limiting practicality. A promising direction is to collect larger and more diverse data so that fine-tuning and generalization can be performed within a semi-supervised learning framework, leveraging a small amount of labeled camera data and a larger set of unlabeled sensor data. Such approaches, relying solely on photo-reflective sensor input, could eliminate the need for camera calibration for new users. Even a small amount of wearer-specific sensor data improved accuracy, indicating that strengthening sensor-only or semi-supervised approaches could further enhance privacy and reduce hardware complexity, making the system more practical for everyday VR.

8. Conclusion

This study demonstrated that photo-reflective sensors can enable expressive, real-time facial reconstruction in HMDs through lightweight, privacy-preserving sensing. By combining camera-based supervision during training with sensor-only inference, our system lowers the hardware and computational load of facial tracking during inference, facilitating wider adoption and more natural nonverbal communication in VR. While promising, this work remains an initial step toward practical deployment, with challenges to be addressed for robust real-world deployment.

Acknowledgment

This work is supported by JSPS KAKENHI Grant Number 24K20823 and 25K03158.

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

References

[AKST17] ANDO T., KUBO Y., SHIZUKI B., TAKAHASHI S.: Canalsense: Face-related movement recognition system based on sensing air pressure in ear canals. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (2017), pp. 679–689. doi:10.1145/3126594.3126649. 2

[AMSS17] ASANO N., MASAI K., SUGIURA Y., SUGIMOTO M.: Facial performance capture by embedded photo reflective sensors on a smart eyewear. In *Proceedings of the 27th International Conference on Artificial Reality and Telexistence and 22nd Eurographics Symposium on Virtual Environments* (Goslar, DEU, 2017), ICAT-EGVE '17, Eurographics Association, p. 21–28. doi:10.2312/egve.20171334. 2, 8

[BBP*15] BEDRI A., BYRD D., PRESTI P., SAHNI H., GUE Z., STARNER T.: Stick it in your ear: Building an in-ear jaw movement sensor. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers* (2015), pp. 1333–1338. doi:10.1145/2800835.2807933. 2

[BM92] BESL P. J., MCKAY N. D.: Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures* (1992), vol. 1611, Spie, pp. 586–606. doi:10.1117/12.57955. 6

[BV03] BLANZ V., VETTER T.: Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence 25*, 9 (2003), 1063–1074. doi:10.1109/TPAMI.2003.1227983. 2

[BWP13] BOUAZIZ S., WANG Y., PAULY M.: Online modeling for realtime facial animation. *ACM Transactions on Graphics (ToG) 32*, 4 (2013), 1–10. doi:10.1145/2461912.2461976. 2

[BYJM18] BERNAL G., YANG T., JAIN A., MAES P.: Physiohmd: a conformable, modular toolkit for collecting physiological data from head-mounted displays. In *Proceedings of the 2018 ACM international symposium on wearable computers* (2018), pp. 160–167. doi:10.1145/3267242.3267268. 2

[CLT*21] CHEN T., LI Y., TAO S., LIM H., SAKASHITA M., ZHANG R., GUIMBRETIERE F., ZHANG C.: Neckface: Continuously tracking full facial expressions on neck-mounted wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 5*, 2 (2021), 1–31. doi:10.1145/3463511. 2

[CLX*22] CHEN S.-Y., LAI Y.-K., XIA S., ROSIN P. L., GAO L.: 3d face reconstruction and gaze tracking in the hmd for virtual interaction. *IEEE Transactions on Multimedia 25* (2022), 3166–3179. doi:10.1109/TMM.2022.3156820. 1, 2

[CSA*20] CHEN T., STEEPER B., ALSHEIKH K., TAO S., GUIMBRETIÈRE F., ZHANG C.: *C-Face: Continuously Reconstructing Facial Expressions by Deep Learning Contours of the Face with Ear-Mounted Miniature Cameras*. Association for Computing Machinery, New York, NY, USA, 2020, p. 112–125. 2

[CWLZ13] CAO C., WENG Y., LIN S., ZHOU K.: 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG) 32*, 4 (2013), 1–10. doi:10.1145/2461912.2462012. 2

[DBB22] DANĚČEK R., BLACK M. J., BOLKART T.: Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 20311–20322. doi:10.1109/CVPR52688.2022.01967. 2

[Ekm89] EKMAN P.: The argument and evidence about universals in facial expressions. *Handbook of social psychophysiology 143* (1989), 164. 2, 4

[FFBB21] FENG Y., FENG H., BLACK M. J., BOLKART T.: Learning an animatable detailed 3D face model from in-the-wild images. In *ACM Transactions on Graphics, (Proc. SIGGRAPH)* (2021), vol. 40. doi:10.1145/3450626.3459936. 2

[HCTW11] HUANG H., CHAI J., TONG X., WU H.-T.: Leveraging motion capture and 3d scanning for high-fidelity facial performance acquisition. In *ACM SIGGRAPH 2011 papers*. 2011, pp. 1–10. doi:10.1145/1964921.1964969. 2

[HDS*19] HICKSON S., DUFOUR N., SUD A., KWATRA V., ESSA I.: Eyemotion: Classifying facial expressions in vr using eye-tracking cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2019), pp. 1626–1635. doi:10.1109/WACV.2019.00178. 2

[IZB*19] IRAVANTCHI Y., ZHANG Y., BERNITSAS E., GOEL M., HARRISON C.: Interferi: Gesture sensing using on-body acoustic interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–13. doi:10.1145/3290605.3300506. 2

[JDITS*22] JOURABLOO A., DE LA TORRE F., SARAGIH J., WEI S.-E., LOMBARDI S., WANG T.-L., BELKO D., TRIMBLE A., BADINO H.: Robust egocentric photo-realistic facial expression transfer for virtual reality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 20323–20332. doi:10.1109/CVPR52688.2022.01968. 2, 6

[LA05] LOURAKIS M. L., ARGYROS A. A.: Is levenberg-marquardt the most efficient optimization algorithm for implementing bundle adjustment? In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* (2005), vol. 2, IEEE, pp. 1526–1531. doi:10.1109/ICCV.2005.128. 3

[LBB*17] LI T., BOLKART T., BLACK M. J., LI H., ROMERO J.: Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) 36*, 6 (2017). doi:10.1145/3130800.3130813. 3

[LTO*15] LI H., TRUTOIU L., OLSZEWSKI K., WEI L., TRUTNA T., HSIEH P.-L., NICHOLLS A., MA C.: Facial performance sensing head-mounted display. *ACM Transactions on Graphics (ToG) 34*, 4 (2015), 1–9. doi:10.1145/2766939. 2

[LWN*20] LOU J., WANG Y., NDUKA C., HAMEDI M., MAVRIDOU I., WANG F., YU H.: Realistic facial expression reconstruction for VR HMD users. *IEEE Transactions on Multimedia 22*, 3 (2020), 730–743. doi:10.1109/TMM.2019.2933338. 1, 2

[LYYB13] LI H., YU J., YE Y., BREGLER C.: Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph. 32*, 4 (2013), 42–1. doi:10.1145/2461912.2462019. 2

[LZC*24] LI K., ZHANG R., CHEN S., CHEN B., SAKASHITA M., GUIMBRETIÈRE F., ZHANG C.: Eyeecho: Continuous and low-power facial expression tracking on glasses. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–24. doi:10.1145/3613904.3642613. 2

[Mat92] MATSUMOTO D.: More evidence for the universality of a contempt expression. *Motivation and Emotion 16*, 4 (1992), 363–368. doi:10.1007/BF00992972. 4

[MKS*20] MASAI K., KUNZE K., SAKAMOTO D., SUGIURA Y., SUGIMOTO M.: Face commands - user-defined facial gestures for smart glasses. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (2020), pp. 374–386. doi:10.1109/ISMAR50242.2020.00064. 2

[MPHS*22] MASAI K., PERUSQUÍA-HERNÁNDEZ M., SUGIMOTO M., KUMANO S., KIMURA T.: Consistent smile intensity estimation from wearable optical sensors. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)* (2022), pp. 1–8. doi:10.1109/ACII55700.2022.9953867. 2

[MSI24] MASAI K., SUGIMOTO M., IWANA B.: Facial gesture classification with few-shot learning using limited calibration data from photo-reflective sensors on smart eyewear. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia* (New York, NY, USA, 2024), MUM '24, Association for Computing Machinery, p. 432–438. doi:10.1145/3701571.3701595. 5

[MSO*16] MASAI K., SUGIURA Y., OGATA M., KUNZE K., INAMI

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

M., SUGIMOTO M.: Facial expression recognition in daily life by embedded photo reflective sensors on smart eyewear. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (New York, NY, USA, 2016), IUI '16, Association for Computing Machinery, p. 317–326. doi:10.1145/2856767.2856770. 2, 7

[MSU17] MATTHIES D. J., STRECKER B. A., URBAN B.: Earfield-sensing: 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* (2017), pp. 1911–1922. doi:10.1145/3025453.3025692. 2

[MWUN21] MATTHIES D. J., WEERASINGHE C., URBAN B., NANAYAKKARA S.: Capglasses: Untethered capacitive sensing with smart glasses. In *Proceedings of the Augmented Humans International Conference 2021* (2021), pp. 121–130. doi:10.1145/3458709.3458945. 2

[NMS*22] NAKAMURA F., MURAKAMI M., SUZUKI K., FUKUOKA M., MASAI K., SUGIMOTO M.: Analyzing the effect of diverse gaze and head direction on facial expression recognition with photo-reflective sensors embedded in a head-mounted display. *IEEE Transactions on Visualization and Computer Graphics 29*, 10 (2022), 4124–4139. doi:10.1109/TVCG.2022.3179766. 1, 2, 3

[NS23] NAKAMURA F., SUGIMOTO M.: Exploring the effect of transfer learning on facial expression recognition using photo-reflective sensors embedded into a head-mounted display. In *Proceedings of the Augmented Humans International Conference 2023* (2023), pp. 317–319. doi:10.1145/3582700.3583705. 1, 2, 5, 7

[OLSL16] OLSZEWSKI K., LIM J. J., SAITO S., LI H.: High-fidelity facial and speech animation for VR HMDs. *ACM Trans. Graph. 35*, 6 (Nov. 2016). doi:10.1145/2980179.2980252. 1, 2

[OSM*13] OGATA M., SUGIURA Y., MAKINO Y., INAMI M., IMAI M.: SenSkin: Adapting skin as a soft interface. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (New York, NY, USA, 2013), UIST '13, Association for Computing Machinery, p. 539–544. doi:10.1145/2501988.2502039. 2

[PHS17] PERUSQUÍA-HERNÁNDEZ M., HIROKAWA M., SUZUKI K.: A wearable device for fast and subtle spontaneous smile recognition. *IEEE Transactions on Affective Computing 8*, 4 (2017), 522–533. doi:10.1109/TAFFC.2017.2755040. 2

[Pow70] POWELL M. J.: A new algorithm for unconstrained optimization. In *Nonlinear programming*. Elsevier, 1970, pp. 31–65. doi:10.1016/B978-0-12-597050-1.50006-3. 3

[RLM*19] ROSTAMINIA S., LAMSON A., MAJI S., RAHMAN T., GANESAN D.: W!NCE: Unobtrusive sensing of upper facial action units with eog-based eyewear. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3*, 1 (Mar. 2019). doi:10.1145/3314410. 2

[RLM*21] RICHARD A., LEA C., MA S., GALL J., DE LA TORRE F., SHEIKH Y.: Audio-and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (2021), pp. 41–50. doi:10.1109/WACV48630.2021.00009. 2

[RV05] ROMDHANI S., VETTER T.: Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (2005), vol. 2, IEEE, pp. 986–993. doi:10.1109/CVPR.2005.145. 2

[SLG*22] SHIN J., LEE S., GONG T., YOON H., ROH H., BIANCHI A., LEE S.-J.: Mydj: Sensing food intakes with an attachable on your eyeglass frame. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–17. doi:10.1145/3491102.3502041. 2

[SNO*17] SUZUKI K., NAKAMURA F., OTSUKA J., MASAI K., ITOH Y., SUGIURA Y., SUGIMOTO M.: Recognition and mapping of facial expressions to avatar by embedded photo reflective sensors in head mounted display. In *2017 IEEE Virtual Reality (VR)* (2017), pp. 177–185. doi:10.1109/VR.2017.7892245. 1, 2, 7

[TTU*20] TAKEGAWA Y., TOKUDA Y., UMEZAWA A., SUZUKI K., MASAI K., SUGIURA Y., SUGIMOTO M., PLASENCIA D. M., SUBRAMANIAN S., HIRATA K.: Digital full-face mask display with expression recognition using embedded photo reflective sensor arrays. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (2020), pp. 101–108. doi:10.1109/ISMAR50242.2020.00030. 2

[TZS*18] THIES J., ZOLLHÖFER M., STAMMINGER M., THEOBALT C., NIESSNER M.: FaceVR: Real-time gaze-aware facial reenactment in virtual reality. *ACM Trans. Graph. 37*, 2 (jun 2018). doi:10.1145/3182644. 1, 2, 6

[WBLP11] WEISE T., BOUAZIZ S., LI H., PAULY M.: Realtime performance-based facial animation. *ACM transactions on graphics (TOG) 30*, 4 (2011), 1–10. doi:10.1145/2010324.1964972. 2

[Wil06] WILLIAMS L.: Performance-driven facial animation. In *Acm SIGGRAPH 2006 Courses*. 2006, pp. 16–es. doi:10.1145/1185657.1185856. 2

[WSS*19] WEI S.-E., SARAGIH J., SIMON T., HARLEY A. W., LOMBARDI S., PERDOCH M., HYPES A., WANG D., BADINO H., SHEIKH Y.: Vr facial animation via multiview image translation. *ACM Transactions on Graphics (ToG) 38*, 4 (2019), 1–16. doi:10.1145/3306346.3323030. 1, 2

[ZXC*19] ZHAO Y., XU Q., CHEN W., DU C., XING J., HUANG X., YANG R.: Mask-off: Synthesizing face images in the presence of head-mounted displays. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (2019), IEEE, pp. 267–276. doi:10.1109/VR.2019.8797925. 6

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.