ExGSense: Toward Facial Gesture Sensing with a Sparse Near-Eye Sensor Array

Chen Chen
Computer Science and Engineering
University of California San Diego
chenchen@ucsd.edu

Ke Sun
Computer Science and Engineering
University of California San Diego
kesun@ucsd.edu

Xinyu Zhang
Electrical and Computer Engineering
University of California San Diego
xyzhang@ucsd.edu

ABSTRACT

Immersive face-to-virtual-face telecommunication is one unique use case for virtual reality (VR) technologies. Existing camera-based telephony systems cannot be used for such immersive VR video chat, due to the physical occlusions of head-mounted displays (HMDs) and/or unwieldy positioning of cameras. To address these, we present ExGSense, a new VR input modality that can sense and reconstruct both upper and lower face gestures, by only using lightweight biopotential sensors embedded within the HMDs. We optimize the sensor arrangement based on facial anatomy and employ a multiview classification pipeline to exploit the multiple dimensions of signal features. We thus enable ExGSense to sense whole facial gestures by using a sparse set of biopotential transducers. We prototyped ExGSense and evaluated its performance with 42 facial gestures and across different users. We showed a 93% accuracy for user-specific evaluation, and 77% accuracy for user-independent evaluation with low calibration overhead. We believe ExGSense constitutes a promising input modality for immersive VR interactions.

CCS CONCEPTS

• **Human-centered Computing** → **Human Computer Interaction (HCI)**; • **Interaction Techniques** → **General Inputs**.

KEYWORDS

Biosensing, Face Gesture Sensing, Face Synthesis, Interactive Virtual Reality, Wearables

ACM Reference Format:
Chen Chen, Ke Sun, and Xinyu Zhang. 2021. ExGSense: Toward Facial Gesture Sensing with a Sparse Near-Eye Sensor Array. In *Information Processing in Sensor Networks (IPSN' 21), May 18–21, 2021, Nashville, TN, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3412382.3458268

1 INTRODUCTION

Emerging virtual reality (VR) technologies are enabling a new form of telepresence applications, where one wearing head-mounted display (HMD) can perform face-to-virtual-face interactions in an immersive 3D environment [27, 45**?**]. Webcam capturing has been widely adopted by traditional video telephony systems such as Skype and Zoom. However, these methods fail in the interactive VR contexts because they cannot capture the upper face which is occluded by the HMD [3, 4, 17, 63]. Existing research explored "see-through" sensors, *e.g.*, electromyography (EMG) transducers, for emotion recognition. But emotion sensing alone can hardly satisfy the needs of interactive VR, where realistic facial images need to be delivered to the remote peers (see Sec. 6.2).

In this work, we present ExGSense, as a compact solution to fill the technology gap. ExGSense acts as a lightweight companion sensor kit that augments existing HMDs. It can sense eye and mouth gestures, and reconstruct a whole face, by leveraging sparse near-eye biopotential signal measurements. ExGSense uses a few low cost commercially available dry electrodes resting around the eyes, to extrapolate the bio-signal features. These features provide a high level abstraction of the sophisticated facial anatomical patterns. To make the feature extraction effective, ExGSense incorporates a novel dual-branch multi-view decision pipeline, as well as a model generalization mechanism. The solution framework enables ExGSense to reconstruct whole-face gestures with much less amount of training compared to prior arts. ExGSense strikes a balance between sensing granularity, cost and form factors. Its facial expression sensing and reconstruction mechanisms constitute a sensing primitive for future interactive VR applications, *e.g.*, creating 3D avatar faces and automatically exchanging emoji streams or facial snapshots between two users wearing HMDs for remote immersive interactions.

We prototyped ExGSense using consumer grade electronics and eye mask, which were incorporated into a VR HMD as a lightweight add-on (see Figure **??**b-e). We adopted widely used eye gestures in existing Human-Computer Interaction (HCI) literature and mouth gestures inherited from lip-syncing applications. Our experiments show that ExGSense achieves an overall accuracy of 93% in sensing the eye/mouth gesture of an individual. With a simple classification pipelines, it achieves a competitive model transferability across different users with overall 63% accuracy. Benefited from its multi-branch classification design, the cross-user classification accuracy is improved to 77% with only an extra $\sim$ 2 min of mouth gesture training examples for model calibration. While near-eye transducers have been widely used to sense eye movement, ExGSense represents the *first* to explore an *indirect* sensing modality with respect to the mouth gestures by leveraging the underlying facial muscle anatomy patterns and biopotential signal propagation. The workflow of ExGSense is summarized in Figure **??**a.

ExGSensemarks an important step towards the vision of immersive VR interaction [64], through the following technical contributions: (1) We propose a new sensing modality, using near-eye biopotential sensors to detect full face expressions. To achieve this,

IPSN' 21, May 18–21, 2021, Nashville, TN, USA

ACM ISBN 978-1-4503-8098-0/21/05.
https://doi.org/10.1145/3412382.3458268

we explore the paradigm of *indirect* sensing where the lower facial gestures can be detected by the transducers resting on the upper face. (2) We propose a dual-branch multiview representation learning pipeline, which can explicitly exploit the sensor diversities across *time-frequency-spatial* domains. We further propose a simple re-calibration approach for adapting the pretrained model for different users. (3) We build a proof-of-concept prototype of ExGSense and conduct user studies to verify its ability to *concurrently* track the fine-grained upper face eye and lower face mouth gestures by fully leveraging the facial anatomy patterns.

2 RELATED WORK

2.1 Camera Based Approach

The most intuitive facial gesture sensing modality is camera based approach. Today's webcams are widely used by Instant Message (IM) and video conferencing applications for remote face-to-face communications and collaborations [**?**]. Computational graphics tools, *e.g.*, OKAO[1], iMotions[2], and OpenFace [3, 4, 17] are also widely used in affective computing domain [13, 15, 49] to sense and analyze facial gestures. With deep learning, these tools have achieved fine-grained face tracking with competitive accuracy. For example, OpenFace can accurately detect facial landmarks, head pose, and gaze [3]. However, under the interactive VR setup, users are not able to see the upper face of remote users due to the occlusions of HMD. Although recent work [1] tried to overcome the hindrance by leveraging the front-facing camera on a smartphone inside the HMD, and has achieved 95.3% blink detection accuracy and 10.8° gaze tracking error, the approach cannot capture the lower facial characteristics. These issues are also witnessed by a vast majority of commercially available devices, *e.g.*, HTC Vive Eye Pro *etc.* Besides, these products would also be subject to high cost. Other researchers [61, 63] achieved face synthesis via multiview head mounted cameras, but the high cost, complicated model training procedures and demanding computing resources pose significant challenges.

2.2 Proximity and Pressure Based Approach

To overcome the barriers of HMDs' occlusions, researchers have explored proximity and pressure sensors with "see-through" capabilities. The idea is to approximate the surface deformation caused by facial gestures. LiGaze [36], for example, used an IR camera to estimate gaze direction inside HMD and smart glasses, and achieved 10.1° error. Others [6] used transparent capacitive sensor arrays to estimate gaze by computing the proximity. Such strategies typically involve complicated hardware design, and can only detect partial face due to limited sensing granularity.

Alternatively, Li *et al.* adopted a pressure based approach with 8 strain gauges attached on the foam line of HMD to sense upper facial gestures and a RGB-D camera to capture the lower facial gestures [34]. MindMaze[3] developed a sensing mask that can make VR more emotional aware [44]. However, such mechanical movement enabled sensing approaches cannot concurrently detect eye activities and lower facial behaviors due to limited field-of-view. Another limitation observed by [34] is low model generalization across users due to the dominance of user-dependent features over gesture-dependent features encoded from sensing data. This implies for each new user, a non-trivial pre-calibration is required, causing low usability in practice.

2.3 Interferometry and Tomography Based Approach

Instead of relying on sensing facial shapes (through proximity) and colors (through vision), interior anatomical patterns of facial gestures may be sensed via interferometry and tomography based approaches. Interferi [25], for example, used 8 ultrasonic transducers to sense 9 face gestures with 89% accuracy. However, this approach fails to sense eye activities, *e.g.*, gaze changing. Similar ideas were also widely used in multiple hand gesture sensing works. For example, by sensing the cross sectional impedance of the wrist and arm to form an electrical impedance tomography, [69, 70] can recognize 11 hand gestures at within-participant accuracy of more than 80%. However, building a globally generalized model is challenging due to the cross-user variance of cross-session anatomical patterns.

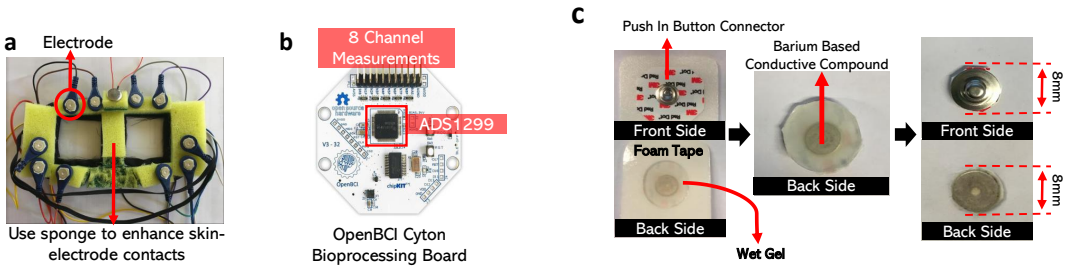
2.4 Biosignal Based Approach

Facial behaviors can also be approximated through the biosignals, which are explored in existing HCI and wearable design research.

The first kind of signal used to measure eye activities is electrooculography (EOG), which is a μV level corneo-retinal standing potentials. With this primitive, [8] used 4 dry electrodes and accelerometers to detect 8 gestures at $\sim$ 90% accuracy. J!NS glasses have achieved approximately 70% accuracy for detecting 4 daily activities with 3 electrodes [26]. However, such solutions are limited to either eye, or high level daily activities. Others [52, 53] used J!NS glasses to sense 4 upper face non-eye gestures by leveraging the signals collected by EOG sensors and accelerometers, which however cannot detect the lower facial gestures.

Electromyography (EMG) is the second approach being widely used for facial expression evaluations. Researchers have employed EMG to evaluate facial gestures and speech activities. For example, AlterEgo [28] used 7 near-mouth EMG transducers to detect a small set of words with 90% accuracy. Nonetheless, such sensors can neither detect upper face activities nor be easily incorporated into HMDs, due to the uncomfortness after adding mounting-racks near the mouth. Commercial available solutions such as EMTEQ[4] powered by FaceTeq [41] proposed a similar wearable hardware kit that can be mounted on top of HMDs with 9 DoF accelerometer, gyroscope and photo-plethysmograph (PPG) pulse rate sensors. However these tools focused on cognitive valence-arousal analysis [42], emotion detection [40] and rehabilitation of Parkinson patients [62]. In contrast, we use a novel end-to-end sensing pipeline to detect a total of 42 combined upper face eye and lower face mouth gestures to address the requirements of interactive VR. To enhance the sensing capability of such an approach, researchers have explored a wide variety of transducer configurations to maximize the information entropy [11, 22, 26, 28, 33, 39, 53, 62, 71]. While

[1]The OKAO: https://plus-sensing.omron.com/concept/
[2]The iMotions: https://imotions.com/
[3]The MindMaze: https://www.mindmaze.com/
[4]The EMTEQ: https://emteq.net/



(a) Our testbed. (b) OTS OpenBCI board. (c) Rapid fabricate low-cost electrode transducers.

Figure 1: Rapid proof-of-concept prototype.

making the final design choices for ExGSense, we have compared our transducer arrangements with several prior arts (see Sec. 4.2).

The third approach is to use Electroencephalogram (EEG), which is usually in the order of nV to μV, generated by the activation of neurons in the brain. Researchers have explored the potentialities of using EEG to detect coarse-grained emotion induced facial gestures for human-robot interactions [2]. Others [21] tried to build a silent speech interface with EEG measurements to indirectly infer lip and mouth gestures. However, the overwhelming complexities of the transducer setups hinder the practical usability in mobile VR systems.

Finally, using a mixture of aforementioned bio-signals are also applied in the existing works. PhysioHMD [7], for example, used a combination of aforementioned bio-signals to classify 12 emotion states with LeNet-5 at more than 90% accuracy. However, detecting emotion states is different to recognizing facial gestures. Discovered by Scherer *et al.* [55], facial characteristic is only one of five metrics being used for evaluating one's emotions. Although, besides those emotions investigated in [7] (*e.g.*, happy, sad, and angry *etc.*), many coarsed grained emotions can be discriminated by the upper facial characteristics, one can still hide their emotion via facial gestures [5]. Unlike this work, our focus is to transform the experience and feelings of VR based remote communication to face-to-face physical communication as close as possible. Therefore, instead of making emotions transparent to the remote user, we rely on one's innate ability to infer the internal emotion states from facial characteristics of remote collaborator [57].

3 PRELIMINARY

Although the signals being measured from the on-face biopotential transducers are heterogeneous, the primary signal sources for sensing eye and mouth gestures are EOG and surface EMG.

EOG is Involved with Eye Movements: The EOG signal signatures induced from eye movements are contributed by the positive charges at cornea side and negative charge at retina side [10]. A fine-grained measurements of such signal variations will provide informative eye activities and gaze direction.

EMG is Involved with Mouth Movements: The signals involved with lower face (mainly mouth) gesture sensing are mainly contributed by facial surface EMG. In ExGSense, we explore *indirect sensing*, *i.e.*, using transducers resting on the upper face to infer the lower face mouth gestures indirectly. Essentially, although the transducers are partially contacting the LLS (Levator Labii Superioris), ZYG (Zygomatics Majors) and RIS (Risorius) round the eyes (Figure 2b), they can still capture mouth induced signals that propagate through Nasalis. By analyzing surface EMG signal captured from near mouth muscles, Eskes *et al.* [16] demonstrated these muscle groups are involved with several mouth activities (*e.g.*, retractor of the upper lips, closure and sealing of the oral commissure).

However, besides facial gestures, the EMG signal is also subject to 2 potential unwanted noise, posing challenges to our algorithm design [43, 65]: (1) inter-person variations, caused by the diverse distributions of muscle fibres among different people; (2) and inter-session variations, caused by minor temporal variations of motor units firing pattern.

With these signal sources, ExGSense also harness the signal diversities from 3 parts. *First*, the *time-domain diversity* is majorly used for eye gesture tracking, where the eye and gaze can be inferred from the EOG patterns, *e.g.*, wave shape. *Second*, the *frequency-domain diversity* is used to demix the eye-induced EOG signals and mouth-induced fEMG signals. The key idea lies in our observation that the surface EMG energy is mainly aggregated at much higher frequency band compared to that of EOG. *Finally*, we explore *spatial diversity* resulting from a multi-transducer arrangement that creates "virtual channels" (see Sec. 4.2).

4 IMPLEMENTATION

4.1 Sensor Board and Prototype Setups

Data Acquisition: Our proof-of-concept setup (Figure 1a) is built on the OpenBCI Cyton Board [47]. The kit (Figure 1b) is developed around ADS1299 chip with 8 16-bit analog-to-digital converter (ADC) channels for μV accuracy data acquisition [24]. We configured the amplifier gain to $\times 24$ and the data communication link is supported by BLE 4.0 [51]. We mount the sensing electrodes on a customized made eye mask[5] (see Figure 1c). The head strap of the eye mask helps us enhance the skin-electrode contacts for maximizing the performance. We use 250 Hz as the sampling frequency, 1 s as the window size.

Transducers: Traditional clinical research prefers to use the sticky wet gel electrodes for recording patients' biosignals, due to their reliable contact with skin surfaces and hence more accurate data acquisition [56]. However, this is not be practical in mobile cases due to one-time disposal and uncomfortness issues while

[5]EyeMask: https://amzn.to/33C43rt

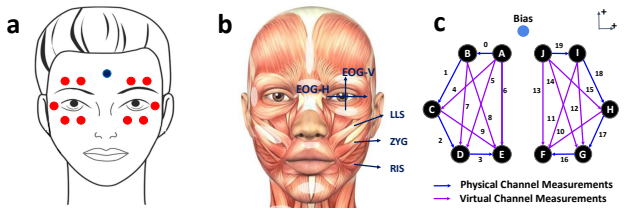


Figure 2: Transducer arrangement design. (a) Transducer placements; (b) Facial anatomy; (c) Channel connections;

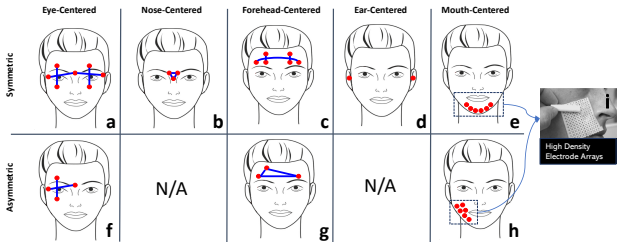


Figure 3: A taxonomy of transducer arrangements. The red dots and blue connections indicate the electrode placements and channel measurements. The clustering of red dots indicate the high density electrode array, *e.g.* [33].

pasting on human face. Ag/Cl dry electrodes are used instead considering their compact form factors and easy system integration. However, the relative high cost impedes many application usages. A typical Ag/Cl flat electrode costs around ~$20[6] whereas a wet electrode costs only ~$0.14[7]. In ExGSense, we choose to modify a low-cost wet electrode (3M 2560 Red Dot)[7]. By removing the surface conducting wet gel and sticky foam tape (Figure 1c), we enable comfortable reuse of the electrodes, at the cost of minor signal quality degradation.

4.2 Transducer Arrangements

ExGSense optimizes the transducer arrangements in order to capture the fine-grained whole facial gestures through a sparse sensor setup. To highlight the advantages, Figure 3 summarizes a wide variety of design choices that prior researchers adopted. For example, [11] detects eye and gaze activities through the *eye-centered design*. J!NS [26] and W!NCE [53] sense eye and upper-facial activities through the *nose-centered design*. Besides, [22, 39] used *forehead-centered* and *ear-centered* transducers respectively to achieve similar functionalities. Other systems [28] used *mouth-centered design* for silent speech recognition. In contrast, our goal is to concurrently sense the eye *and* mouth gestures *only* using the upper-face biopotential transducers. Thus, we design a sparse transducer array (Figure 2) based on 2 design choices.

Indirect Sensing and Symmetric Design: ExGSense aims to sense the lower facial gestures by only leveraging the transducers resting on the upper face in an *indirect* manner. We observe 3 muscle majors (Figure 2b), named LLS, ZYG and RIS engaged with upper and lower facial movements. This motivates us to put 6

[6]Dry AgCl Eelctrodes: https://bit.ly/2MMcJou
[7]Wet Electrode by 3M 2560 Red Dot: https://bit.ly/2MmPYZ8

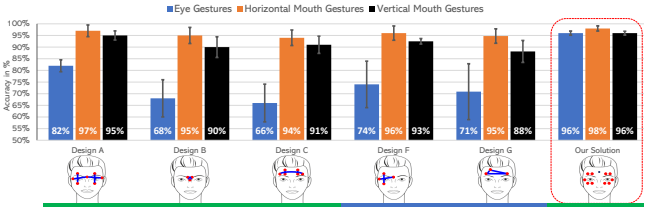


Figure 4: Performance of different transducer placement scheme. We used green and blue underscores the mark scenarios using symmetric design tenet respectively.

transducers, marked as Ⓒ – Ⓗ on top of these 3 muscle majors at the left and right of the face for maximizing the signal-to-noise ratio (SNR) (Figure 2c). While placing other transducers such as Ⓑ and Ⓘ, we also consider the auxiliary muscle majors related to mouth movements, *e.g.*, *Temporails*, to maximize the entropy of captured relevant gesture signatures. We used symmetric design for introducing *spatial redundancies*, to minimize the imperfect sensing performance in mobile interaction settings.

Virtual Channel Measurements: Inspired by the tomographic based sensing approach for hand gestures [69, 70] which maximally exploit the impedance responses of each probing path, we introduce the concept of *virtual channel measurements* shown by the pink arrow in Figure 2c. Hereby, we define the *physical channel measurements* as the pairwise potential measurements directly pulled from the ADS1299 chip configured in continuous data reading mode [24]. The *virtual channel measurements* are instead computed algebraically based on physical channel measurements. The goal is to exploit the *spatial sensing diversities* without adding additional hardware complexities. With this approach, we can approximate the EOG-V (vertical EOG) signals with virtual channel 7, 6, 13, 12 and EOG-H (horizontal EOG) with virtual channel 0, 19, 3, 16 (see Figure 2b for the EOG-V and EOG-H measurements). Such additional dimensions of features would make the facial gesture sensing more reliable and finer grained.

Benchmark: To validate this, we conduct a small scale study with 4 participants (mean age = 23.75, mean face dimension (height × width) = 9.53" × 5.10"), 6 eye gestures, marked as *neutral*, *blink*, *gaze up*, *gaze down*, *gaze left*, *gaze right*, and 3 mouth gestures in horizontal as well as vertical movements directions, marked as *small*, *medium* and *large*.

Evaluations: We focus on the design of Figure 3{a-c, f, g}. For asymmetric design, we averaged out the results when transducers are placed at left and right part of the face. We use an AlexaNet [31] to evaluate the classification performance. Figure 4 shows the results with our solution reaching the highest overall accuracy across eye and 2 different mouth gesture directions. This verifies the feasibility of *indirect sensing* with respect to mouth gestures. We observed a slightly higher sensitivity of horizontal mouth gestures than the vertical one. This is consistent with our initial analysis of the facial anatomy pattern where the horizontal mouth movement involved more with the mid and upper facial muscles, *e.g.*, LLS, ZYG and RIS. In contrast, the gestures involved with vertical mouth movements can be more easily inferred from the lower facial muscles, *e.g.*, Depressor Labii Inferioris. The results also verify the benefits of the spatial redundancies from the symmetric design. On

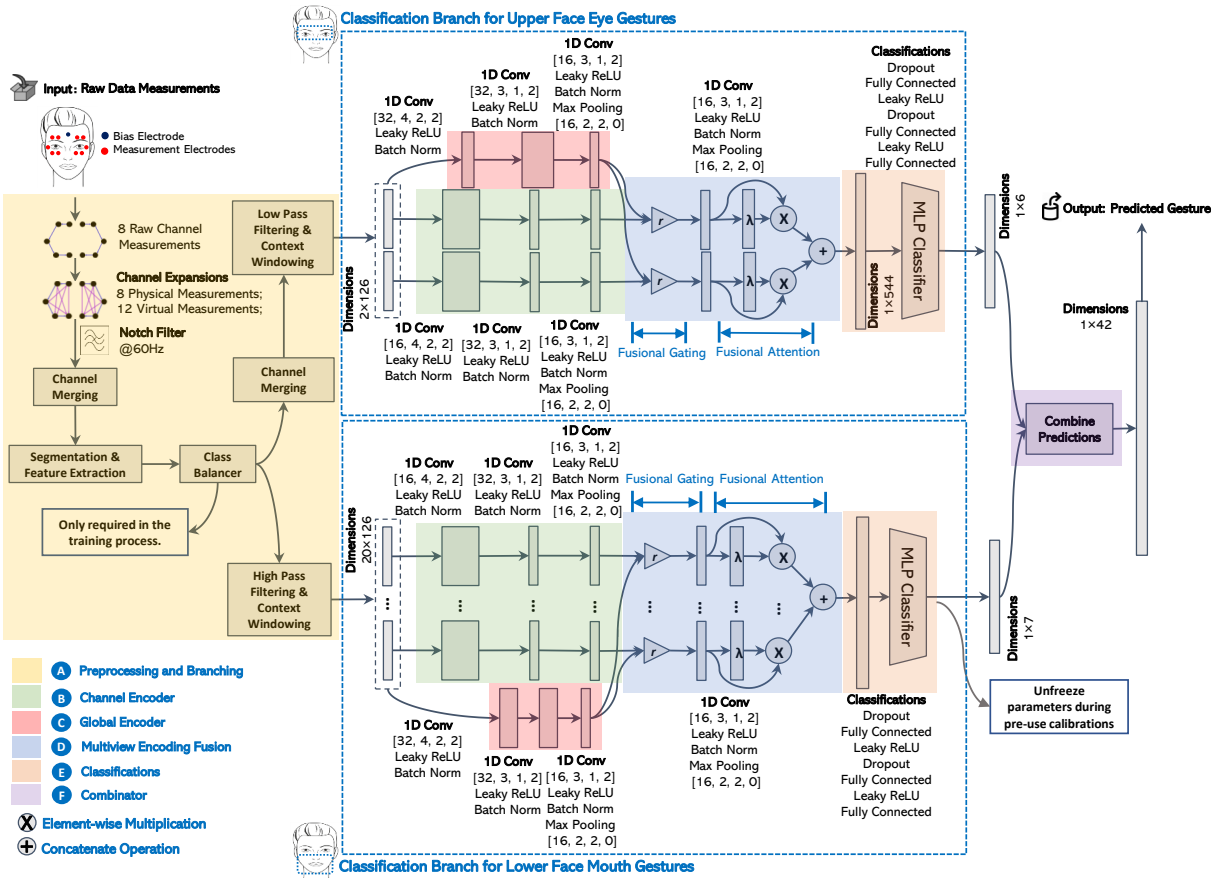


Figure 5: Our 6-stage dual-branch multi-view pipeline, capable of using a 13-non-mixed-gesture trained model to predict 42 gestures by leveraging the signal demixability over frequency domain.

Procedure for estimating EOG-H signal

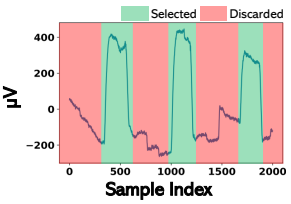
```
Function EOG_H (D, λ):
    assert 0.5 <= λ <= 1
    left = −D[0, :] + D[3, :]
    right = −D[16, :] + D[19, :]
    C_left = RMS(left)
    C_right = RMS(right)
    if C_left > C_right then
        return λ left + (1 − λ) right
    else
        return (1 − λ) left + λ right
    end
return
```

(a) EOG-H estimation.

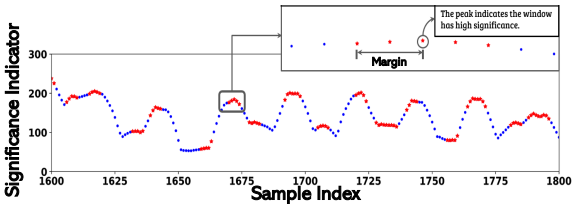
Procedure for estimating EOG-V signal

```
Function EOG_V (D, λ):
    assert 0.5 <= λ <= 1
    left = −D[7, :] - D[6, :]
    right = −D[12, :] - D[13, :]
    C_left = RMS(left)
    C_right = RMS(right)
    if C_left > C_right then
        return λ left + (1 − λ) right
    else
        return (1 − λ) left + λ right
    end
return
```

(b) EOG-V estimation.



(c) Signal example.



(d) Segmentation results.

Figure 6: (a – b) Cross channel merging for distilling of EOG signal. Note that D indicates the array of virtual channel measurements. The index of each channel can be referred to Figure 2c. (c – d) Segmentation process for EOG signal.

the other side, the asymmetric design can result in higher standard deviations of accuracies across participants, leading to less stable accuracies. Our solution greatly boosts the SNR of collected data and implies that with symmetric design, we would have higher possibilities of collecting more accurate data in the contexts of high mobility and variations of wearing styles.

5 SENSING ALGORITHMS

We design a deep learning based decision making pipeline to identify participants' facial expressions by explicitly exploiting temporal, frequency and spatial features from the ExGSense data. Figure 5 demonstrates the workflow comprised of 6 stages. Note that the raw biopotential signals are weak and easily contaminated by noise and interference. Therefore, during prepossessing, we use a cascaded high pass filter, cutoff at 0.2 Hz and a notch filter at 60 Hz to minimize the baseline drift and electromagnetic interference (EMI) [32].

Our informal measurement shows a very weak interference caused by the harmonics of EMI, thus we only remove the fundamental components at 60 Hz explicitly.

5.1 Signal De-mixing

In this stage, we take advantage of frequency-domain diversity to separate the EOG signals for tracking eye gestures and the EMG signal for detecting mouth shapes. This separation simplifies the model training process and reduces the number of training classes from KL to $(K + L)$, where K and L represents the number of eye and mouth gestures, respectively.

Estimations of Cutoff Frequencies: Our initial goal is to approximate the optimum frequency position that can well separate the EOG and EMG signals. We pilot a small-scale benchmark experiment by asking 4 participants to perform the 6 eye gestures and 6 mouth gestures used in Sec. 4.2. We then compute the FFT of each sampled window at each channel, and adopt the Kendall's tau [29] to perform a feature significance test of each FFT features [29]. The results of these p-value tests corresponding to mouth and eye gestures are plotted in Figure 7a and 7b. Note that the smaller p-value corresponds to the higher feature significance. We then plot the average p-value computed across each channels at each frequency position, shown in Figure 7c and 7d. With this observation, we approximate 25 Hz as the separating frequency to de-mix the eye and mouth induced signals.

Dual-Branch Classifications: With signal de-mixability over frequency domain, we propose a dual-branch classification method, which separates the classification processes, and hence the weight parameters, for tracking eye and mouth gestures, as illustrated in Figure 5. Our strategy differs from the prior work, which tends to use a coherent bulky decision making framework, such as SVM (support vector machine) [69, 70], random forest [25], or standardized deep learning model [7]. Such frameworks need to be trained by feeding all the possible combinations of mouth and gesture classes, which involves huge training overhead. In contrast, our approach can train the model with non-mixture gestures, but predict both mixture and non-mixture gesture sets. With the boundary for separating EOG and EMG over frequency domain in Figure 7, we apply a low-pass and high-pass filter for eye and mouth classification branch, respectively, before further processing.

5.2 Cross-Channel Feature Extractions and Segmentation

ExGSense uses EOG signals for tracking eye gestures. Unlike the clinical use case, user mobility and variations of wearing styles incur significant noise. We mitigate this issue by harnessing the redundancies of EOG-H and EOG-V channel measurements from the left and right part of the face. We also introduce a temperature parameter λ to synthesize the channel measurements collected from the two parts. Based on our pilot study, we empirically set λ to 0.9. Figure 6a and 6b illustrates the algorithms used for estimating the synthesized EOG channel measurements. Although prior work [66] used the product of RMS and sample entropy as the significance indicator, to minimize computing latency, we only used the RMS value for this purpose, indicated by C_{left} and C_{right}.

A typical eye movement comprises two forms, namely saccades and fixations. Saccades refer to the case when the eyes are moving around constantly to locate the interesting objects, and fixations occurs when gazes are held upon a specific location [8]. Our work targets to the saccade and blinking, which typically last for 80 ms and is more useful to analyze interests shifts of VR users [12]. Thus, perfectly labelling the eye gestures within a time series of ExGSense samples would be challenging (Figure 6c). To address this, we implement a segmentation block (see Figure 5) for picking sanitized eye gesture window automatically only during training process. Specifically, we compute the RMS value of each coarse segmented window, and use the window corresponding to the peak of RMS as well as its neighbours within a pre-tunned margin as relevant data segment. We consider this RMS value as the window significance indicator. With a small scale pilot benchmark, we set this margin parameter to 2. This process can be illustrated in Figure 6d.

5.3 Multiview Channel Fusion

To exploit the spatial diversity from the ExGSense transducer arrangement (Figure 2), we use a variant of the multiview framework proposed in [67, 68]. As illustrated in stage Ⓑ to Ⓔ in Figure 5, the augmented features of each channel, noted as $< X_i, y >$, will be fed into separate multiview convolution channel encoder with the output of h_i, where i refers to the virtual channel index. Simultaneously, these feature maps will also be fed into a globally shared encoder to extract the global views, noted as h_g. After that, we concatenate these outputs to compute the gate allowance rate that is similar to the *forget gate* design in the LSTM (Long Short Term Memory) cell [23]. This gate allowance rate can be computed by logistic function (Equation (1)), where we use $\oplus$ to denote feature concatenations and $F(\cdot)$ to represent the function approximator for the fully connected layers.

$$r_i = \frac{1}{1 + exp(-F(h_g \oplus h_i))}, r_i \in \mathbb{R} \tag{1}$$

Next, we use the computed allowance rate r_i to fuse the output from i-th channel encoder h_i and global encoder h_g heuristically, as shown in Equation (2), where we use $\odot$ to represent the operation of element-wise multiplications.

$$\bar{h_i} = (1 - r_i) \odot h_g + r_i \odot h_i \tag{2}$$

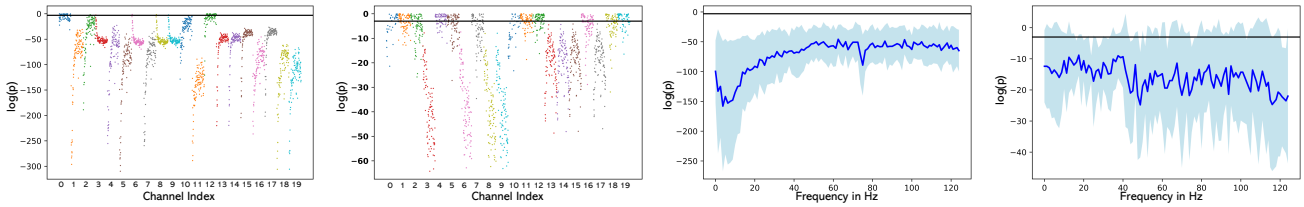
Therefore, the global attention energy vector α_g can be derived in Equation (3), where C indicates the total number of channels. Based on the strategy proposed in [67], we use a temperature value $\lambda = 0.9$ to control the aggressiveness of the exponential normalization.

$$\alpha_g = [\alpha_{g1}, \alpha_{g2}, ..., \alpha_{gC}], \alpha_{gi} = \frac{exp(\lambda F(\bar{h_i}))}{\sum_k exp(\lambda F(\bar{h_k}))} \tag{3}$$

By merging the globally shared attention energy vector α_g and the fused feature map $\bar{h_g}$ using Equation (4), we are able to compute the globally contextual vector c_g.

$$c_g = \sum_{i=1}^{C} \alpha_g i \odot \bar{h_i} \tag{4}$$

Finally, we used a stacked fully connected layer, as the last-step classification stage based on contextual vectors. Notably, during the training phase, we used a class balancer to remove samples from the majority class. Further, only the classification stage for mouth gesture branch will be unfreezed when the model is re-calibrated



(a) Raw eye gesture result (b) Raw mouth gesture result (c) Avg. eye gesture result (d) Avg. mouth gesture result

Figure 7: The p-value test results. The raw results can be referred to Figure 7a and 7b. The averaged results across each channel can be referred to Figure 7c and 7d. The lightblue region in subplot 7c and 7d indicate the error computed by ± standard deviation.

for a new user, whereas the eye branch needs no calibration (see Sec. 7.3). For each branch, we uses the cross-entropy losses [20] as the optimization object. We summarize each block of our deep learning model in Figure 5.

6 DEPLOYMENT AND IMPLEMENTATION

6.1 Data Collection

To collect data, we invite 10 participants (mean age $\mu = 24$, $\sigma = 1.41$; face height $\mu = 9.38$", $\sigma = 0.50$"; face width $\mu = 5.36$", $\sigma = 0.51$") to wear the ExGSense HMD and perform the instructed facial gestures. For the event based facial gesture, we ask participants to repeat 20 times in each session. For the state based facial gesture, around 20 seconds samples are required in each session. To avoid artifacts caused by participants' learning experience and physical muscle fatigue [38], participants were allowed to have $\sim$ 10 seconds breaks in between sub-sessions. Finally, the same gesture may vary when performed by different participants. We faithfully incorporate such practical effects, by requiring participants to perform the gestures at the levels they found comfortable and repeatable. Each session takes approximately 40 min excluding the break time. Our study was approved by the Institutional Review Board (IRB).

6.2 Sensed Events

Although emotion sensing may help in social interaction context [39], practically it is hard to use this approach to realize the goal of immersive VR communication. *First*, the granularity of easily observable emotional expressions is relatively low. For instance, there are only 6 basic emotion catalogues as proposed in [14]. *Second*, emotional expressions represent a higher level semantic abstractions of low level facial gestures, *i.e.*, the former can be easily derived from the latter. In the context of VR face-to-virtual-face communication, conveying the remote users' emotions is only one important component to achieve our goal. We envision the remote users to perform the face-gesture-to-emotion translations cognitively so as to exploit the richer information besides basic emotions. For example, the users can figure out whether the remote users are distracted based on the visually perceived gaze directions. To compose our sensed events, we adopt the commonly used eye gesture sets in [1, 8, 25] and mouth gesture sets inherited from lip-syncing applications [58] to form the fine-grained gesture set in our study. Traditionally, as a critical post-production phase in film industry, lip-syncing is used as a mapping technique between predefined lip shape and utterance, and thus being widely used for dubbing and creations of visually vivid animation avatars [9, 58]. This includes 6 eye gestures and 7 mouth gestures (Figure 8). Together there are 42 gesture combinations, which cover a wide range of practical facial expressions.

6.3 Sensing Model Implementation

The ExGSense model is developed using Pytorch [48]. The collected data are split to 70%, 10% and 20% for training, validation and testing purposes. The batch gradient descent [35, 54] with Adam Optimizer [30] is used for training all models, with batch size 64. To ensure convergence at global optimum and prevent overfitting, we set both learning rate and weight decay [37] to 10^{-5} and run 100 epochs for all networks during the training phase.

7 EVALUATION

7.1 Single User Evaluation

Recall that, by using our decision making pipeline in Figure 5, the model can classify all facial gesture combinations by solely training over the individual gestures. For the standard facial expression set comprised of 6 eye gestures and 7 mouth gestures, ExGSense only needs to collect training data for $7 + 6 = 13$ gestures, instead of $7 \times 6 = 42$ gestures. We now compare the sensing performance with these two levels of training overhead.

With the collected data from Sec. 6, we evaluate the F1 score, recall and precision with only 13 non-mixture gestures and 42 gestures being used for training purpose respectively. We used F1 score as the overall metric in order to balance between precision (low false positive) and recall (low false negative) [59]. The confusion matrix with only 13 non-mixture training gestures for 10 participants can be found in Figure 10.

Signal Demixability: We show the signal demixability with an overall accuracy for eye and mouth gesture classification being 90% and 97% respectively. This implies our lightweight model training can effectively recognize concurrent gestures.

Competitiveness of Dual Branch Classification Pipeline: Our evaluation also shows a competitive performance of our sensing approach with the dual-branch model being trained by only 13 non-mixture gestures. For comparison purpose, we also evaluated the single branch model used to classify lower facial gestures (see Figure 5). Notably, to train the single branch model, dataset with

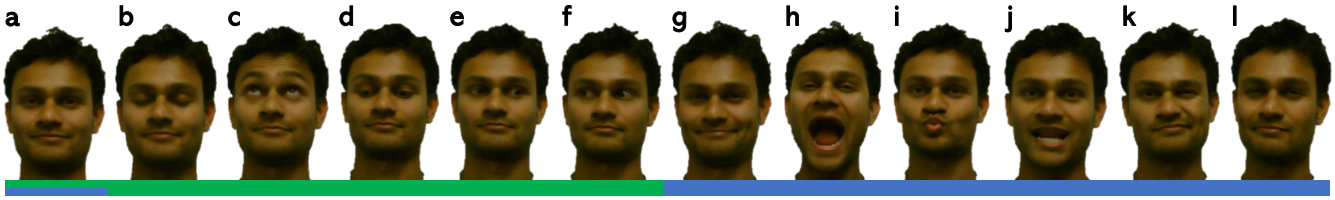


Figure 8: Final non-mixed gesture set: (a) neutral, (b) blinks, (c) gaze looking up, (d) gaze looking down, (e) gaze looking left, (f) gaze looking right, (g) mouth gesture A - smile, (h) mouth gesture B - mouth open, (i) mouth gesture C - kissy mouth, (j) mouth gesture D - tongue touch upper teeth, *e.g.* mouth shape when pronouncing "L" sound, (k) mouth gesture E - raising left cheek, (l) mouth gesture F - raising right cheek. The total number of gesture set including mixed and non-mixed are $6 \times 7 = 42$. We used green and blue underscores to indicate eye and mouth gestures respectively.

Index	Eye	Mouth	Index	Eye	Mouth	Index	Eye	Mouth	Index	Eye	Mouth	Index	Eye	Mouth
0	None	None	9	None	D	18	Gaze Up	B	27	Blink	D	36	Gaze Right	E
1	Blink	None	10	None	E	19	Gaze Down	B	28	Gaze Up	D	37	Blink	F
2	Gaze Up	None	11	None	F	20	Gaze Left	B	29	Gaze Down	D	38	Gaze Up	F
3	Gaze Down	None	12	Blink	A	21	Gaze Right	B	30	Gaze Left	D	39	Gaze Down	F
4	Gaze Left	None	13	Gaze Up	A	22	Blink	C	31	Gaze Right	D	40	Gaze Left	F
5	Gaze Right	None	14	Gaze Down	A	23	Gaze Up	C	32	Blink	E	41	Gaze Right	F
6	None	A	15	Gaze Left	A	24	Gaze Down	C	33	Gaze Up	E			
7	None	B	16	Gaze Right	A	25	Gaze Left	C	34	Gaze Down	E			
8	None	C	17	Blink	B	26	Gaze Right	C	35	Gaze Left	E			

Figure 9: Class index and corresponding sensing event.

42 classes is required. Our approach shows an overall 93% accuracy that is approximately same as the traditional way of training over all gesture combinations. However, it substantially reduces ~ 69% overhead in collecting training data from users, leading to higher usability and scalability.

Competitiveness of Multiview Encoding Pipeline: We also show the merit of the multiview pipelines over the traditional SVM approach. As part of baseline, we use the SVM with default RBF kernel ($\gamma = \frac{1}{D}$, where D is the number of features) to replace multiview sub-pipeline at each branch. Compared to our approach, the low recall (21 classes are below 70%) and accuracy (an overall 73%) show the setbacks while using traditional SVM strategies.

Sensitivity of Mouth Gestures When Mixed with Eye Gestures: The results also show a high sensitivity of mouth gesture recognition during mixture events, with the overall accuracy reaching 98%. This verifies the effectiveness of the indirect sensing approach, and potential for tracking whole face with a sparse transducer setup.

7.2 User Independent Evaluation Without Calibrations

Maintaining consistent performance across new participants and usage instances (*i.e.*, sessions) is critical but notoriously challenging for most wearable sensing devices, due to the inter-session variations and inter-person variations. We evaluate the effectiveness of ExGSense in meeting such challenges. Since each session consists of 42 subsessions of non-mixture and mixture gestures, it is impractical to collect the multi-session gestures for each participant in a large scale. Therefore, we only evaluate the inter-personal variations, *i.e.*, how ExGSense adapts to new participants using the model trained on a given set of participants, and without any calibration efforts on the target participants.

We use leave-one-user-out cross-validation to estimate the system performance. Specially, we train the model over 9 participants and tested on the remaining participant. We repeat this for 10 times where each participant's data was used once as the target domain.

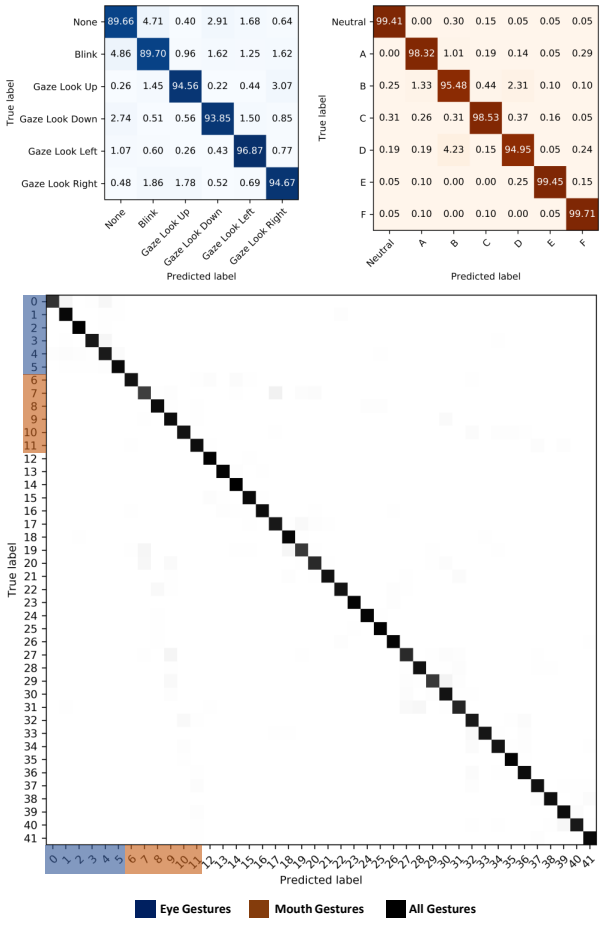


Figure 10: Confusion matrices for dual-branch multiview pipeline. The indices can be referred to Figure 9. All results are in the units of %.

From results shown in Figure 11, it is worth noting that our model has fairly reasonable transferability across a majority of gestures (an overall accuracy of 80% for eye and 78% for mouth). However, a few mouth gestures show relatively poor performance, such as Gesture B (Mouth Open) with only 46.15% recall (see Figure 11).

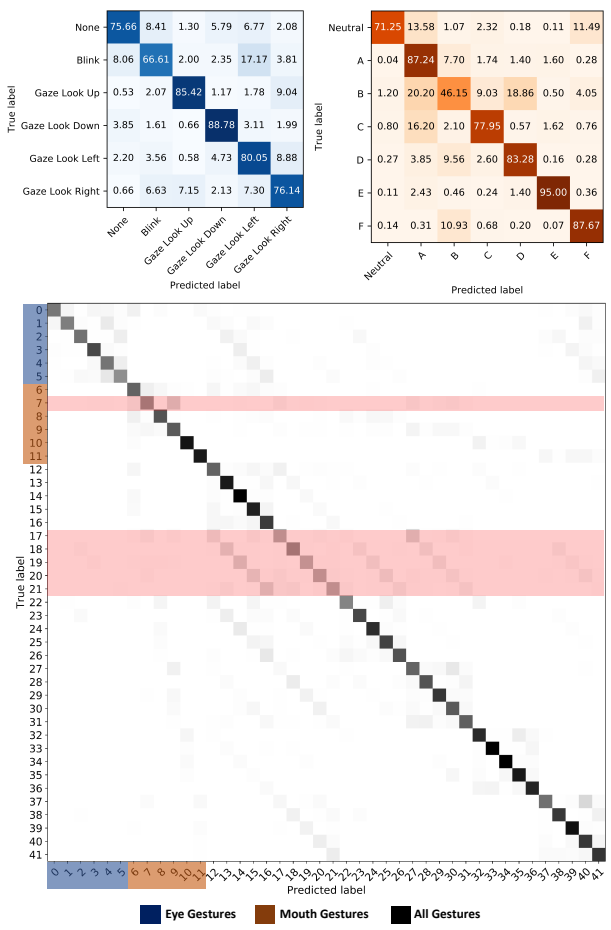


Figure 11: Confusion matrices of user independent evaluations without calibrations. The low classification performance highlighted by red region is caused by the poor performance of gesture B (Mouth Open). The indices can be referred to Figure 9. All results are in the units of %.

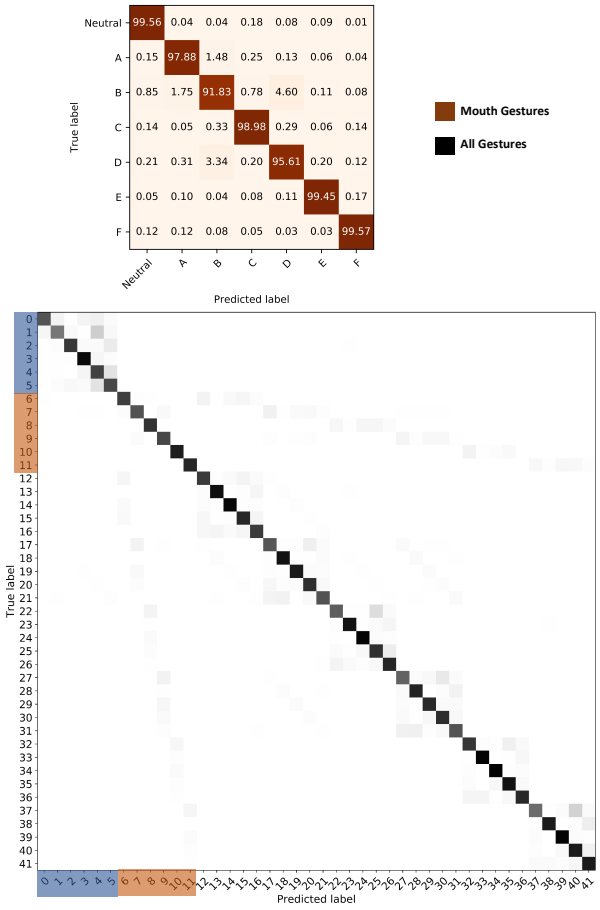


Figure 12: Confusion matrices after fine tuning the model. As we did not fine tune the classification branch for eye gestures, where therefore are not shown here. The indices can be referred to Figure 9. All results are in the units of %.

One possible reason is that such mouth gestures involve less with the muscle majors located at mid and upper face. This leads to the poor system performance for partial gesture combinations (see red region in Figure 11). Motivated by this, we ask users to recalibrate the system with solely mouth gestures (see Sec. 7.3). This ensures high usability since users are not required to recalibrate for each gestures.

7.3 User Independent Evaluations With Partial Calibration

Motivated by the observations in Sec. 7.2, we found designing a zero-shot transfer learning network for all predefined mouth gestures is challenging, while the eye gestures show reasonable results. Therefore, during system re-calibration process, we ask new participants to provide only ~ 2 min mouth gestures to fine tune the classifier layer of stage Ⓔ shown in Figure 5, so as to address this challenge. Our goal here is to evaluate the cross-participant system performance after slight fine tuning of mouth gestures.

As is shown in Figure 12 after fine tuning of mouth gesture, we demonstrate a sensing performance of our decision making pipeline with model tuning for mouth gestures at stage Ⓔ with accuracy being 77%, compared to the previous single participant evaluations at accuracy of 93%. Figure 13 also shows a performance comparisons across different gestures using F1 scores with aforementioned evaluations. With partial model calibration with *only* mouth gestures, we observe an average 15% F1 score enhancement, as to that without any calibrations. Compared to the single-branch classification pipeline, which requires new participants to provide 42-gesture training examples, we greatly reduce the calibration labour work where target users are only asked for 6 mouth gestures for model tuning purpose.

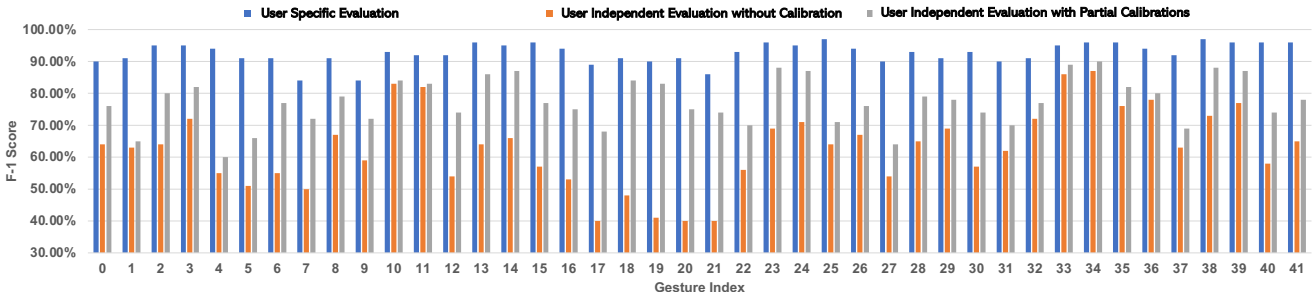


Figure 13: Comparisons of F1 score of different model training approaches. The indices can be referred to Figure 9.

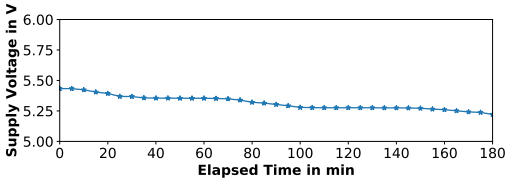


Figure 14: Measured battery voltage in the first 3 hours after batteries are fully charged while keeping data acquisition and streaming enabled.

7.4 Usability

Power Consumption. While collecting and streaming data to the processing machine, we measured the current drawn from power supply is around 40 mA. With 5 V operating voltage, the power consumption is estimated around 0.2 W. With 4 fully charged 2300 mAh battery, we profiled the battery voltage in the first 3 hours (see Figure 14). This shows the potentials of integrating our prototype with OTS VR/AR headset with additional engineering efforts on system optimizations, *e.g.*, compress the data packets in a more efficient way while being streamed to the data processing server.

Comfortness and Form Factor. The total weight of our prototype is around 105.3 g excluding the data acquisition development kit, where the Google Cardboard and additional add-on contribute 100.5 g and 4.8 g respectively. This shows that adding our prototype would increase only 4.8% of the physical weight, which is negligible in terms of end user experience. In general, all participants were satisfied with the prototype design. While discussing with participants regarding the setbacks of current prototype design, 3 participants pointed out the uncomfortable experience when their forehead skin was covered by the sponge frames and 2 participants felt troublesome due to the placements of connecting cables. Since ExGSense, so far, is only a proof-of-concept prototype, we believe such setbacks should be avoidable when ExGSense is manufactured with more carefully selected materials.

8 FACE-TO-VIRTUAL-FACE INTERACTIONS

Combining with the state-of-the-art generative model, we now show the potentiality that use ExGSense to synthesize a user's face.

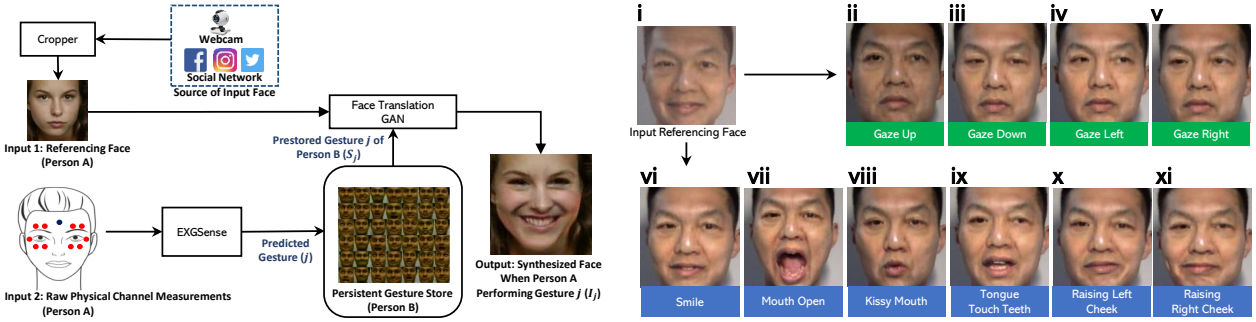
Architectures: At high level, we use ExGSense as the *gesture predictor* and face translation GAN[8] as the *face synthesizer*. The workflow is shown in Figure 15a. *First*, we required the raw channel measurements and a reference face image as the input, which can be prerecorded by webcam, or crowdsourced from social network profile photos. With the Face Recognition Toolkit[9], we *then* extracted the head portraits from the background. *Finally*, the architecture will output a synthesised participants' face which will effectively address the occlusion issues of HMDs (see Sec. 2.1).

Results and Limitations: To quantify the effectiveness for all gesture synthesis, we computed the $L2$ distance metric computed by DLib[10] between participant real image and synthesized image (I_j) (*i.e.*, is the person in the synthesized image looked same as that behind the HMD occlusions doing the same gestures?). Empirically, the Face Recognition Toolkit consider the distance below 0.6 as the criteria for justifying 2 head portraits are same person. *First*, Some gestures highly involved with mouth shows an unexpected high distance between real and generated images (mean distance = 0.53, *std.* = 0.01). One reason behind this is the infrequent occurrence of such gestures in the training dataset, leading to the poor performance for extracting the hidden gesture features. A *second* limitation is the poor performance on generating peripheral facial hairs. This is an inherent issues existing in most of state-of-arts DeepFake solutions [46, 60, 72]. *Finally*, we only focus on discrete state facial synthesis. In the future, we will investigate the potentialities of using advanced generative model, *e.g.*, Ganimation [50], to "guess" the synthesized faces between discrete state gestures. We envision the advance of generative model research in deep learning and pattern recognition domain would be able to foster our work.

9 LIMITATIONS AND FUTURE WORK

We believe ExGSense opens up several directions for future research. *First*, we showed the possibility of reconstructing a fine-grained discrete set of facial expressions. However, to realize full-fledged interactive VR, we need to track and reconstruct user's face in real-time and continuous manner. It would be also critical to acknowledge the importance of designing and integrating ExGSense into commercial grade VR headset, which is considered as part of our future work. *Second*, our current sensing model can adapt to different users through fine tuning, with a small set of user-specific training samples (~ 2 min for mouth gesture). However, practically,

[8]Face Translation GAN: https://github.com/shaoanlu/fewshot-face-translation-GAN
[9]Face Recognition Toolkit: https://github.com/ageitgey/face_recognition
[10]The DLib toolkit: http://dlib.net



(a) Pipeline for face synthesis applications.

(b) Example outputs from the face synthesizer.

Figure 15: The architecture pipeline and the example outputs for reconstructing participants' face through ExGSense. Note that gestures of looking left and right are with respect to the participants.

Mouth Gestures \ Eye Gestures	Neutral	Gaze Up	Gaze Down	Gaze Left	Gaze Right
Neutral	0.41	0.40	0.48	0.47	0.50
Gesture A	0.46	0.46	0.51	0.46	0.46
Gesture B	0.48	0.46	0.51	0.49	0.49
Gesture C	0.53	0.41	0.49	0.46	0.45
Gesture D	0.55	0.52	0.54	0.53	0.53
Gesture E	0.49	0.42	0.45	0.40	0.43
Gesture F	0.50	0.41	0.42	0.40	0.39

Figure 16: Distance between real and synthesized image (I_j). A small distance indicates two head portraits have high probability coming from same person.

our ultimate goal is to eliminate such efforts on the new users, with zero-shot transfer learning. With support of sufficient amount of training data, this can potentially be realized through designing a separator network, which splits the latent participant dependent features from gesture dependent features. *Third*, we expect continuous efforts on optimizing metrics with respect to multiple usability aspects, *e.g.*, adding a richer set of VR interaction modalities [18, 19], reducing power consumption and latency, as well as enhancements of comfortness and design (see Sec. 7.4). *Finally*, we only demonstrate the feasibility of ExG based facial reconstruction through a rapid prototype and a small number of participants. A larger scale and more diverse user study is part of our future work.

10 CONCLUSIONS

We explored ExGSense, an input modality that leverages the underlying facial anatomy patterns and the muscle activity propagation to concurrently sense eye and mouth gestures, with a small set of upper face transducers only. We showed this can be done through a well designed dual-branch multi-view classification pipeline. We verified that the channel measurements of upper face transducers carry not only the EOG signals activated by eye movements, but the propagated EMG information involved with mouth gestures. With prototype, we showed an overall sensing accuracy of 93% for individual user and 77% for user independent evaluation, corresponding to 42 facial gestures composed of 6 eye gestures and 7 mouth gestures. By integrating ExGSensewith the state-of-the-art generative model and facial recognition toolkit, we demonstrate the promising of using ExGSense to reconstruct a user's whole face, to realize face-to-virtual-face telephony. Our results are encouraging and presenting a new competitive tenet for designing future generation interactive VR.

11 ACKNOWLEDGMENT

We appreciate feedback from the anonymous reviewers and fellow colleagues from Jacobs School at UC San Diego. We also appreciate the constructive feedback offerred by Prof. Edward Wang, Prof. Scott Klemmer and Prof. Nadir Weibel from UC San Diego and Prof. Yang Zhang from UCLA. We particularly thank Ish Kumar Jain for voluntarily providing us with the demonstrative photos. This work was partially supported by the Google Faculty Research Award, and by the NSF under grant CNS-1901048, CNS-192767 and CNS-1952942.

REFERENCES

[1] Karan Ahuja, Rahul Islam, Varun Parashar, Kuntal Dey, Chris Harrison, and Mayank Goel. 2018. EyeSpyVR: Interactive Eye Sensing Using Off-the-Shelf, Smartphone-Based VR Headsets. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2, Article 57 (July 2018), 10 pages. https://doi.org/10.1145/3214260

[2] Rocio Alba-Flores, Fernando Rios, Stephanie Triplett, and Antonio Casas. 2019. Gesture Recognition Using an EEG Sensor and an ANN Classifier for Control of a Robotic Manipulator. In *Intelligent Computing*, Kohei Arai, Rahul Bhatia, and Supriya Kapoor (Eds.). Springer International Publishing, Cham, 1181–1186.

[3] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1–10.

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*.

[5] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest* 20, 1 (2019), 1–68.

[6] Janne Bergman, Jari Saukko, and Jussi Severi Uusitalo. 2015. US20150015847A1 - Capacitive eye tracking sensor - Google Patents. https://patents.google.com/patent/US20150015847A1/en. (Accessed on 04/09/2019).

[7] Guillermo Bernal, Tao Yang, Abhinandan Jain, and Pattie Maes. 2018. PhysioHMD: A Conformable, Modular Toolkit for Collecting Physiological Data from Head-mounted Displays. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers (ISWC '18)*. 160–167.

[8] Andreas Bulling, Daniel Roggen, and Gerhard Tröster. 2008. It's in Your Eyes: Towards Context-awareness and Mobile HCI Using Wearable EOG Goggles. In *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp '08)*. ACM, New York, NY, USA, 84–93.

[9] Jungsung Cho. 2013. Research on Animation Lip synchronization technology: A study on application and development of domestic animation Lip synchronization. *International Journal of Asia Digital Art and Design* 17, 3 (2013), 87–92.

[10] CleveLabs. 2006. Electro-Oculography Laboratory. https://glneurotech.com/docrepo/teaching-labs/Electro-Oculography_I_Student.pdf (Accessed on 07/27/2019).

[11] Aniana Cruz, Diogo Garcia, Gabriel Pires, and Urbano Nunes. 2015. Facial Expression Recognition Based on EOG Toward Emotion Detection for Human-Robot Interaction. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 4 (BIOSTEC 2015)*. SCITEPRESS - Science and Technology Publications, Lda, Portugal, 31–37. https://doi.org/10.5220/0005187200310037

[12] Andrew Duchowski. 2007. *Eye Tracking Methodology*. Springer-Verlag.

[13] Ekman and Friesen. 2019. FACS (Facial Action Coding System). https://www.cs.cmu.edu/~face/facs.htm. (Accessed on 04/16/2019).

[14] Paul Ekman. 2003. *Emotions Revealed*. Owl Books, New York, New York, USA.

[15] Paul Ekman. 2019. Facial Action Coding System - Paul Ekman Group. https://www.paulekman.com/facial-action-coding-system/. (Accessed on 04/16/2019).

[16] Merijn Eskes, Maarten J. A. van Alphen, Alfons J. M. Balm, Ludi E. Smeele, Dieta Brandsma, and Ferdinand van der Heijden. 2017. Predicting 3D lip shapes using facial surface EMG. *PLOS ONE* 12, 4 (04 2017), 1–16. https://doi.org/10.1371/journal.pone.0175025

[17] Angeliki Fydanaki and Zeno Geradts. 2018. Evaluating OpenFace: an open-source automatic facial comparison algorithm for forensics. *Forensic Sciences Research* 3, 3 (2018), 202–209. https://doi.org/10.1080/20961790.2018.1523703 arXiv:https://doi.org/10.1080/20961790.2018.1523703

[18] Chuhan Gao, Yilong Li, and Xinyu Zhang. 2018. Livetag: Sensing Human-Object Interaction through Passive Chipless WiFi Tags. In *Proceedings of the 15th USENIX Conference on Networked Systems Design and Implementation (NSDI'18)*. USA, 533–546.

[19] Chuhan Gao, Xinyu Zhang, and Suman Banerjee. 2018. Conductive Inkjet Printed Passive 2D TrackPad for VR Interaction. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom '18)*. New York, NY, USA, 83–98. https://doi.org/10.1145/3241539.3241546

[20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org

[21] Jose L. Herrero Ashesh D. Mehta Nima Mesgarani Hassan Akbari, Bahar Khalighinejad. 2019. Towards reconstructing intelligible speech from the human auditory cortex | Scientific Reports. *Nature* (01 2019). https://doi.org/10.1038/s41598-018-37359-z

[22] Lubos Hladek, Bernd Porr, and W. Owen Brimijoin. 2018. Real-time estimation of horizontal gaze angle by saccade integration using in-ear electrooculography. *PLOS ONE* 13, 1, 1–24. https://doi.org/10.1371/journal.pone.0190420

[23] Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9 (12 1997), 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

[24] Texas Instruements. 2017. Low-Noise, 4-, 6-, 8-Channel, 24-Bit, Analog-to-Digital Converter for EEG and datasheet (Rev. C). http://www.ti.com/lit/ds/symlink/ads1299.pdf (Accessed on 04/10/2019).

[25] 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 (CHI '19)*. ACM, New York, NY, USA, Article 276, 13 pages. https://doi.org/10.1145/3290605.3300506

[26] Shoya Ishimaru, Kai Kunze, Yuji Uema, Koichi Kise, Masahiko Inami, and Katsuma Tanaka. 2014. Smarter Eyewear: Using Commercial EOG Glasses for Activity Recognition. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. ACM, New York, NY, USA, 239–242. https://doi.org/10.1145/2638728.2638795

[27] Robert Kanaat. 2016. Five Reasons Why Virtual Reality Is A Game-Changer. https://www.forbes.com/sites/robertadams/2016/03/21/5-reasons-why-virtual-reality-is-a-game-changer/#6a8a52bc41be (Accessed on 07/20/2019).

[28] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. AlterEgo: A Personalized Wearable Silent Speech Interface. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 43–53. https://doi.org/10.1145/3172944.3172977

[29] M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1-2 (06 1938), 81–93. https://doi.org/10.1093/biomet/30.1-2.81 arXiv:http://oup.prod.sis.lan/biomet/article-pdf/30/1-2/81/423380/30-1-2-81.pdf

[30] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. arXiv:cs.LG/1412.6980

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12)*. Curran Associates Inc., Red Hook, NY, USA, 1097–1105.

[32] Ron Kurtus. 2019. List of Worldwide AC Voltages and Frequencies by Ron Kurtus - Physics Lessons: School for Champions. https://www.school-for-champions.com/science/ac_world_volt_freq_list.htm#.XTekIuhKiUk. (Accessed on 07/23/2019).

[33] B. G. Lapatki, J. P. van Dijk, I. E. Jonas, M. J. Zwarts, and D. F. Stegeman. 2004. A thin, flexible multielectrode grid for high-density surface EMG. *Journal of Applied Physiology* 96, 1 (2004), 327–336. https://doi.org/10.1152/japplphysiol.00521.2003 arXiv:https://doi.org/10.1152/japplphysiol.00521.2003 PMID: 12972436.

[34] 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 Trans. Graph.* 34, 4, Article 47 (July 2015), 9 pages.

[35] Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J. Smola. 2014. Efficient Mini-batch Training for Stochastic Optimization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 661–670. https://doi.org/10.1145/2623330.2623612

[36] Tianxing Li, Qiang Liu, and Xia Zhou. 2017. Ultra-Low Power Gaze Tracking for Virtual Reality. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems (SenSys '17)*. ACM, New York, NY, USA, Article 25, 14 pages. https://doi.org/10.1145/3131672.3131682

[37] Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. arXiv:cs.LG/1711.05101

[38] Gazzoni Marco, Botter Alberto, and Vieira Taian. 2017. Surface EMG and muscle fatigue: multi-channel approaches to the study of myoelectric manifestations of muscle fatigue. *Physiological Measurement* 38, 5 (2017), R27–R60.

[39] Katsutoshi Masai, Yuta Sugiura, Masa Ogata, Kai Kunze, Masahiko Inami, and Maki Sugimoto. 2016. 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 (IUI '16)*.

[40] I. Mavridou, M. Hamedi, M. Fatoorechi, J. Archer, A. Cleal, E. Balaguer-Ballester, E. Seiss, and C. Nduka. 2017. Using Facial Gestures to Drive Narrative in VR. In *Proceedings of the 5th Symposium on Spatial User Interaction (SUI '17)*. ACM, New York, NY, USA, 152–152. https://doi.org/10.1145/3131277.3134366

[41] Ifigeneia Mavridou, James T. McGhee, Mahyar Hamedi, Mohsen Fatoorechi, Andrew Cleal, Emili Ballaguer-Balester, Ellen Seiss, Graeme Cox, and Charles Nduka. 2017. FACETEQ: A Novel Platform for Measuring Emotion in VR. In *Proceedings of the Virtual Reality International Conference - Laval Virtual 2017 (VRIC '17)*. Article 9, 3 pages.

[42] Ifigeneia Mavridou, Ellen Seiss, Theodoros Kostoulas, Emili Balaguer-Ballester, and Charles Nduka. 2018. A System Architecture for Emotion Detection in Virtual Reality *(EuroVR '18)*. EuroVR Association.

[43] Roberto Merletti. 2004. *Electromyography : Physiology, Engineering, and Non-Invasive Applications*. Vol. 11. https://doi.org/10.1002/0471678384

[44] MindMaze. 2019. MindMaze reveals Mask to capture your facial expression in virtual reality | VentureBeat. https://venturebeat.com/2017/04/12/mindmaze-reveals-mask-to-capture-your-facial-expression-in-virtual-reality/.

[45] Hidenobu Nagata, Dan Mikami, Hiromu Miyashita, Keigo Wakayama, and Hideaki Takada. 2017. Virtual Reality Technologies in Telecommunication Services. *Journal of Information Processing* 25 (2017), 142–152.

[46] Asher Flynn Nicola Henry, Anastasia Powell. 2018. AI can now create fake porn, making revenge porn even more complicated. http://theconversation.com/ai-can-now-create-fake-porn-making-revenge-porn-even-more-complicated-92267 (Accessed on 08/13/2019).

[47] OpenBCI. 2019. Cyton Biosensing Board (8-channels) - OpenBCI Online Store. https://shop.openbci.com/collections/frontpage/products/cyton-biosensing-board-8-channel?variant=38958638542 (Accessed on 07/22/2019).

[48] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch.

[49] Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA, USA.

[50] Albert Pumarola, Antonio Agudo, Aleix M. Martínez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. GANimation: Anatomically-aware Facial Animation from a Single Image. *CoRR* abs/1807.09251 (2018).

[51] RFDigital. 2013. RFD22301 Data Sheet. https://www.mouser.com/ds/2/470/rfd22301.data.sheet.11.24.13_11.38pm-272240.pdf (Accessed on 07/22/2019).

[52] Soha Rostaminia, Alexander Lamson, Subhransu Maji, Tauhidur Rahman, and Deepak Ganesan. 2019. W!NCE: Eyewear Solution for Upper Face Action Units Monitoring. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications (ETRA '19)*. ACM, New York, NY, USA, Article 63, 3 pages. https://doi.org/10.1145/3314111.3322501

[53] Soha Rostaminia, Alexander Lamson, Subhransu Maji, Tauhidur Rahman, and Deepak Ganesan. 2019. W!NCE: Unobtrusive Sensing of Upper Facial Action Units with EOG-based Eyewear. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 23 (March 2019), 26 pages. https://doi.org/10.1145/3314410

[54] Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. arXiv:cs.LG/1609.04747

[55] Klaus R. Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information* 44, 4 (2005), 695–729. https://doi.org/10.1177/0539018405058216 arXiv:https://doi.org/10.1177/0539018405058216

[56] A Searle and Les Kirkup. 2000. A direct comparison of wet, dry and insulating bioelectric recording electrodes. *Physiological measurement* 21 (06 2000), 271–83. https://doi.org/10.1088/0967-3334/21/2/307

[57] Susan Sullivan, Ted Ruffman, and Sam B. Hutton. 2007. Age Differences in Emotion Recognition Skills and the Visual Scanning of Emotion Faces. *The Journals of Gerontology: Series B* 62, 1 (2007), P53–P60. https://doi.org/10.1093/geronb/62.1.P53

[58] Daniel Swolf. 2019. DanielSWolf/rhubarb-lip-sync: Rhubarb Lip Sync is a command-line tool that automatically creates 2D mouth animation from voice recordings. You can use it for characters in computer games, in animated cartoons, or in any other project that requires animating mouths based on existing recordings. https://github.com/DanielSWolf/rhubarb-lip-sync Accessed on 07/21/2019.

[59] Alaa Tharwat. 2018. Classification Assessment Methods. *Applied Computing and Informatics* (2018). https://doi.org/10.1016/j.aci.2018.08.003

[60] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nie. 2018. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. *Commun. ACM* 62, 1 (Dec. 2018), 96–104. https://doi.org/10.1145/3292039

[61] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nie. 2018. FaceVR: Real-Time Gaze-Aware Facial Reenactment in Virtual Reality. *ACM Trans. Graph.* 37, 2, Article 25 (June 2018), 15 pages. https://doi.org/10.1145/3182644

[62] David Vintiner. 2018. These face-reading glasses track physical and mental health | WIRED UK. https://www.wired.co.uk/article/emteq-vr-digital-phenotyping-charles-nduka (Accessed on 10/17/2019).

[63] Shih-En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. 2019. VR Facial Animation via Multiview Image Translation. *ACM Trans. Graph.* 38, 4, Article 67 (July 2019), 16 pages.

[64] Xiufeng Xie and Xinyu Zhang. 2017. POI360: Panoramic Mobile Video Telephony over LTE Cellular Networks. In *Proceedings of the 13th International Conference on Emerging Networking EXperiments and Technologies (CoNEXT '17)*.

[65] Lin Yang, Wei Wang, and Qian Zhang. 2016. Secret from Muscle: Enabling Secure Pairing with Electromyography. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM (SenSys '16)*. ACM, New York, NY, USA, 28–41. https://doi.org/10.1145/2994551.2994556

[66] Qiang Yang, Yongpan Zou, Meng Zhao, Jiawei Lin, and Kaishun Wu. 2018. ArmIn: Explore the Feasibility of Designing a Text-entry Application Using EMG Signals. In *Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous '18)*. ACM, New York, NY, USA, 117–126. https://doi.org/10.1145/3286978.3287030

[67] Ye Yuan and Kebin Jia. 2019. FusionAtt: Deep Fusional Attention Networks for Multi-Channel Biomedical Signals. *Sensors* 19, 11 (2019).

[68] Ye Yuan, Guangxu Xun, Kebin Jia, and Aidong Zhang. 2019. A Multi-View Deep Learning Framework for EEG Seizure Detection. *IEEE Journal of Biomedical and Health Informatics* 23, 1 (2019).

[69] Yang Zhang and Chris Harrison. 2015. Tomo: Wearable, Low-Cost Electrical Impedance Tomography for Hand Gesture Recognition. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software (UIST '15)*. ACM, New York, NY, USA, 167–173. https://doi.org/10.1145/2807442.2807480

[70] Yang Zhang, Robert Xiao, and Chris Harrison. 2016. Advancing Hand Gesture Recognition with High Resolution Electrical Impedance Tomography. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA, 843–850. https://doi.org/10.1145/2984511.2984574

[71] Wei-Long Zheng, Kunpeng Gao, Wei Liu, Jing-Quan Liu, Guoxing Wang, and Bao-Liang Lu. 2019. Vigilance Estimation Using a Wearable EOG Device in Real Driving Environment. *IEEE Transactions on Intelligent Transportation Systems* (2019), 1–15. https://doi.org/10.1109/TITS.2018.2889962

[72] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *2017 IEEE International Conference on Computer Vision (ICCV)* (Oct 2017). https://doi.org/10.1109/iccv.2017.244