Realistic Facial Expression Reconstruction for VR HMD Users

Jianwen Lou, Yiming Wang, Charles Nduka, Mahyar Hamedi, Ifigeneia Mavridou, Fei-Yue Wang, Fellow, *IEEE*, Hui Yu, *Senior Member, IEEE*

***Abstract*—We present a system for sensing and reconstructing facial expressions of the virtual reality (VR) head-mounted display (HMD) user. The HMD occludes a large portion of the user's face, which makes most existing facial performance capturing techniques intractable. To tackle this problem, a novel hardware solution with electromyography (EMG) sensors being attached to the headset frame is applied to track facial muscle movements. For realistic facial expression recovery, we first reconstruct the user's 3D face from a single image and generate the personalized blendshapes associated with seven facial action units (AUs) on the most emotionally salient facial parts (ESFPs). We then utilize pre-processed EMG signals for measuring activations of AU-coded facial expressions to drive pre-built personalized blendshapes. Since facial expressions appear as important nonverbal cues of the subject's internal emotional states, we further investigate the relationship between six basic emotions - anger, disgust, fear, happiness, sadness and surprise, and detected AUs using a fern classifier. Experiments show the proposed system can accurately sense and reconstruct high-fidelity common facial expressions while providing useful information regarding the emotional state of the HMD user.**

***Index Terms*— facial expression reconstruction, head-mounted display, electromyogram, 3D face reconstruction, facial action unit**

I. Introduction

Recent progress in virtual reality (VR), augmented reality (AR) and mixed reality (MR) has introduced immersive user experience in virtual worlds. Existing mainstream head-mounted displays (HMDs), such as Oculus Rift [1] and HTC Vive [2] enable users to perceive the virtual world, but they only allow limited interactions between the user and the virtual environment. These interactions are mainly based on human body motion capture and hand tracking technologies but ignore the importance of facial expressions for communication.

Facial expressions serve as the primary nonverbal means of communication among human beings [3], [4]. A truly interactive and immersive experience cannot be envisioned without the technologies for sensing and recovering the user's facial expressions in VR. However, VR HMDs usually occlude a large part of the user's face, which rules out most existing facial performance sensing methods, such as ordinary camera-based technologies. A few recent works [5], [6], [7], [8], [9] have explored solutions to this problem. Li et al. [6] and Olszewski et al. [7] made preliminary trials of equipping HMDs with the facial performance capture ability. However, their solutions require a RGB-D or RGB camera mounted on the HMD, which are not ergonomically comfortable and cause an extra head burden. Some works resorted to other advanced facial sensing technologies, such as infrared (IR) sensors [5] and [8] electromyography (EMG) sensors [10], [11]. These lightweight optical or contact-based sensors can be easily embedded into the headset in an unobtrusive manner, thus open a new era of HMD-based wearable facial performance sensing systems.

However, existing solutions [6], [7], [8] build the HMD user's face embodiment with non-realistic facial shape or texture. Moreover, a few systems [8] can only detect the facial expression category which is subsequently represented with pre-defined facial movements on a pre-made virtual avatar. This prohibits natural interactions between participants in the virtual world. To address these problems, we develop a framework (see Fig. 1) that captures facial activities coded in facial action units (AUs, see Fig. 2) [13] upon an advanced facial sensing hardware and can exhibit realistic expressions via

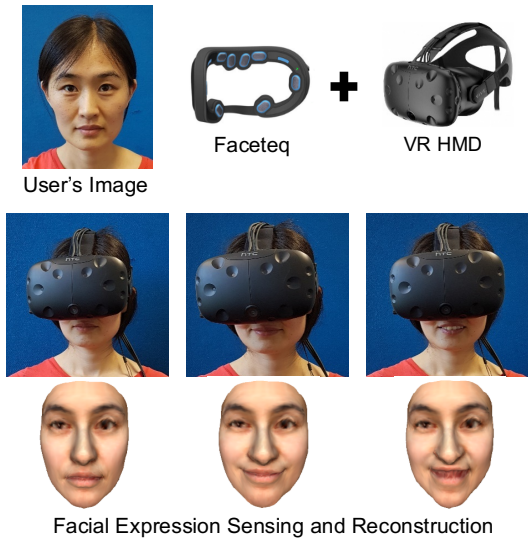

Fig. 1. A demonstration of our system. With a single face image, we are able to sense and reconstruct realistic facial expressions of the head-mounted display (HMD) user. Faceteq is a facial sensing wearable device that can be attached on mainstream HMDs. By utilizing eight integrated electromyography (EMG) sensors, Faceteq enables the detection of the facial muscle contractions of the HMD wearer.

This work was supported by EPSRC Grant (EP/N025849/1) and the Royal Academy of Engineering Grant (IFS1819\9) and Emteq Ltd.

Jianwen Lou, Yiming Wang and Hui Yu are with the School of Creative Technologies, University of Portsmouth, Portsmouth, PO1 2DJ, UK.

Charles Nduka, Mahyar Hamedi and Ifigeneia Mavridou are with Emteq Ltd, Sussex Innovation Center, Brighton, BN1 9SB, UK.

Fei-Yue Wang is with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, 100190, China.

Corresponding author: Hui Yu, hui.yu@port.ac.uk.

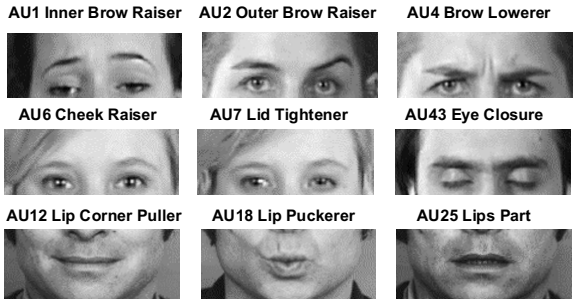

Fig. 2. Examples of facial action units [12].

a compelling digital embodiment of the user's face in virtual scenarios.

We embed a pioneering facial sensing hardware – Faceteq™ [11] into the HMD to detect facial muscle activities through integrated EMG sensors placed on the most emotionally salient facial part (ESFP) – the eye region. Relevant AU-coded facial expressions are then identified with a machine learning method from pre-processed Root Mean Square (RMS) levels of recorded EMG signals. With a single image of the user, we first reconstruct the user's 3D face and generate AU-based blendshapes using a popular analysis-through-synthesis approach [14] and a robust non-rigid registration algorithm [15] Classic psychological studies predict basic emotions from AUs following a few heuristic rules [13]. However, as each category of emotions can have slightly different muscle group contraction, the real relationship between AUs and emotions turns to be complicated, which can be hardly explained with limited rules. The approach becomes infeasible when the observed AUs are not covered by the rules, such as in our case where seven AUs (AU1, AU2, AU4, AU6, AU12, AU25 and AU43) focusing on the ESFPs – the eye and mouth region are investigated. To this end, we use the fern classifier to model the probability of emotions given activated AU information.

The developed system in this paper has been validated through appropriate experiments. Here is a brief summary of the contributions:

- Proposed the first automatic system of its kind that senses and reconstructs the HMD user's facial expressions with a realistic face embodiment.
- Developed an innovative correlation from facial biometric (EMG) signals to facial expressions through individual AUs, which explicitly captures the detailed facial movements performed by the HMD user.
- Proposed a novel probabilistic model that builds relationships between AUs and the six basic emotions.

II. Related Work

A. HMD-based Facial Sensing Systems

The VR HMD occludes a significant portion of the user's face, preventing most existing facial performance capture approaches from being applied. Some recent works embedded small advanced optical sensors inside the headset, such as IR sensors [5], [8] to recognize facial expressions of basic emotions by detecting facial movements. Meanwhile, the contact-based sensing technology is drawing great attention in facial wearable device application because of its superiority in scenarios with highly constrained visibility. The electroencephalography (EEG) has been used in [16] to record brain activities to detect basic emotions, but extensive training and user concentration are required. A commercial device from Looxid Labs [17] incorporates two-channel EEG into a HMD, but whilst this may provide some information of user focus, this will not provide information that directly relates to user valence and facial expressions. An alternative is to measure the electrical signals of muscle activities. EMG is more sensitive at detecting micro-contractions of muscles and indeed was used to calibrate initial computer vision facial tracking algorithms [18]. It has been successfully combined with facial wearable devices for recognizing facial expressions and emotional states [10], [11]. All these technologies offer a wide range of pathways to a reliable HMD-based facial sensing system. However, the studies above only predict from biometric signals facial expression categories (e.g. happiness, anger) whereby what facial movements were involved is still unknown.

Li et al. [6] first equipped the HMD with the ability to sense almost the whole face region. They integrated the HMD with eight strain gauges and a RGB-D camera to capture facial activities of the occluded upper face region and the mouth. The captured facial performance data were then mapped to blendshape coefficients through a linear regression to realize real-time facial animation. However, their solution requires tedious calibrations for each user and the mounted RGB-D camera introduces an extra head burden. Olszewski et al. [7] subsequently improved Li's solution with a lightweight RGB camera and two IR cameras that see direct views of the HMD user's mouth and eyes, and the convolutional neural network that builds a robust mapping between facial images and animation parameters of a pre-built virtual character. Their approach is free of user-specific calibrations and can generate relatively comprehensive facial animation.

Although existing techniques and approaches have pushed the boundary of HMD-based facial sensing systems forward, the following three problems remain unresolved:

1) Most previous systems [5], [8], [10], [11] concentrate on recognition of facial expression categories while ignoring the fact that facial expression has multiple appearance representations. For example, happiness can be expressed with either the AU12 (lip corner puller) or a combination of AU12 and AU6 (cheek raise), so a deeper insight into the composition (e.g. AUs) of facial expressions is needed.
2) The fidelity of reconstructed facial expressions has not attracted sufficient attentions yet. Previous works [6], [7] only consider the geometry of the facial expression while omitting the 3D shape and texture of the user's face, which would generate unrealistic facial expressions in virtual environment.
3) Previous studies have not created an integrated pipeline whereby a personalized model of the user's face can be captured with a smartphone and used to better represent how they express themselves.

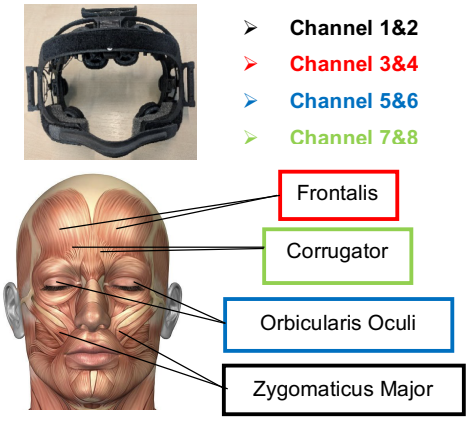

Fig. 3. A prototype of Faceteq. The device equips with eight dry EMG sensors placed on the ESFP that output eight-channel EMG signals. Each channel correlates to a specific facial muscle (highlighted with rectangles in different color) contraction (the facial muscle picture is retrieved from https://fineartamerica.com/).

B. 3D Face Reconstruction from a Single Image

The field of 3D face reconstruction from a single image has witnessed significant progresses over the past two decades [14], [19], [20], [21], [22]. Readers are referred to [23] for a comprehensive survey. Researches differ along two main dimensions, the underlying face prior and the reconstruction algorithm.

Face Priors. Face priors that model the geometry and texture of faces typically serve as the basis of 3D face reconstruction in the ill-posed monocular setting. 3D Morphable Model (3DMM) is a statistical face prior which was originally generated from a database of 200 scanned neutral human faces [19]. It depicts the facial geometry and albedo within a multi-linear Principal Component Analysis (PCA) subspace. The 3DMM can also be extended to faces with expressions [24], [25]. Although 3DMM has laid the foundation for monocular 3D face reconstruction, it shows limitations when dealing with in-the-wild data captured in uncontrolled scenarios. With the available large-scale scanned 3D/4D face data [26], [27], [28] and in-the-wild texture modelling [29], 3DMM has been pushed closer to fully solve the 3D face reconstruction from unconstrained images. Apart from 3DMM, a single 3D face reference has also been applied [30]. As this reference model should provide an initial estimation of the facial geometry and albedo, it thus needs to closely depict the desired face.

Reconstruction Algorithms. There are two main lines in this phase: a) generative approaches [14], [19], [31] and b) discriminative approaches [21], [25]. The generative approach treats the monocular 3D face reconstruction as an inverse rendering problem and formulates it as a complex optimization process. Metrics such as the color consistency, feature similarity and regularization constraints are then used to direct the optimization [14], [20]. This kind of approach can recover promising 3D facial geometry and texture information, while achieving real-time performance with the GPU solver [14]. However, the inverse rendering problem is highly ill-posed due to the incomplete input data, it is hence prone to degenerating in challenging scenarios, such as dealing with faces under severe occlusions and large poses.

Recently, the discriminative approach has emerged as an essential research branch resulting from dramatic progress in deep learning [21], [22]. Built on top of a database that contains extensive face images as well as the corresponding 3D facial data, deep learning is able to embed massive image-face relationships into a robust non-linear regression. This alleviates the problem of ill-posed images which creates incomplete or uncontrolled input data. Existing 3D face databases [32], [33] were collected with sophisticated 3D facial capture systems in controlled settings or using the aforementioned generative approaches on in-the-wild face images [25]. Such processes are time-consuming, labor-intensive or not able to provide the fine-scale 3D facial data. To tackle these problems, a few recent studies resorted to the synthetic 3D facial data and yielded impressive results [34], [35]. This is in line with an interesting theory of parallel vision [36], [37] which discusses the significance of synthetic data [38], [39], [40], [41], [42] in addressing the problems of visual perception and understanding.

The combination of the generative approach and the discriminative approach now appears as an important direction in this [43], [44]. The state-of-the-art [43] that integrates a convolutional encoder network with an expert-designed generative model does not require any 3D facial data for training, while is still able to output promising reconstructions. Recovery of facial geometry and texture details using deep neural networks is also an interesting direction [45], [46].

To eliminate the need of massive training data, we turn to a practical and reliable generative approach [14] which has been validated in several state-of-the-art works [34], [46], [47]. We move a step further to generate personalized blendshapes with a robust non-rigid registration method [15] and a direct deformation transfer from generic AU-based blendshapes [48].

C. Emotions from Facial Action Units

The Facial Action Coding System (FACS) [13] is the best-known taxonomy of human facial expressions. It uses AUs to code the visually observable actions of individual or a group of facial muscles. For example, AU12 describes the contraction of the Zygomaticus major muscle that is typically observed in the expression of happiness. Thus, FACS AUs can offer a detailed interpretation of facial expressions resulting from facial muscle contractions. The revised FACS [49] defines 32 anatomic AUs and 14 Action Descriptors (ADs) with respect to the head pose, gaze and other actions such as blow and bite. In this work, we equip an EMG-based facial sensing hardware [11] with a learning method to identify common facial expressions coded in AUs.

Psychological studies [13] suggest that emotions [50] such as happiness, sadness, fear, anger, surprise, and disgust can be predicted from AUs with a few heuristic rules (e.g. AU6, 12 normally indicates happiness). However, the rules cannot fully explain the complicated relationship between AUs and emotions, since each category of emotions can have various facial appearance representation. Existing rule-based methods [51], [52] are thus unable to provide emotional information

from AU combinations that are not included in the established rules. To address this problem, we propose to use the fern classifier [53] to build the relationship between six basic emotions and AUs with a posterior probability model. The proposed method fits well with our scenario where only AUs around the eye and mouth region are studied.

III. System Overview

A. Device

Our system integrates a wearable facial sensing hardware - Faceteq [11] (see Fig. 3) that fits with mainstream commercial VR head-mounted displays (HMDs). The hardware utilizes the EMG technology to detect facial muscle activations. It consists of eight surface dry EMG sensors placed on the ESFP that do not require skin preparation, conductive gel and adhesive pads, while having a 24-bit signal resolution, 1kHz sampling rate, 20-450Hz signal bandwidth and no inter-sensor latency. Each EMG sensor accounts for a unique muscle action on Zygomaticus major, Frontalis, Orbicularis oculi and Corrugator. The output gives eight channels of muscle activations as well as their intensity scores at 1kHz when wired to a PC based VR system such as Oculus Rift or HTC Vive, or 25Hz via Bluetooth when using a smartphone. The proposed learning method in this paper can further associate EMG channel activations to common AU-coded facial expressions.

The hardware contains two photoplethysmogram (PPG) sensors and an inertial measurement unit (IMU) including accelerometer and gyroscope which provide nine channels values of head movement, posture and state analysis at 50Hz. Meanwhile, it supports real-time signal quality monitoring, ASCII data files as well as binary files for post-acquisition data analysis.

B. Work Pipeline

As illustrated in Fig. 4, our system consists of offline 3D face reconstruction, personalized blendshapes generation, online AU-coded facial expression detection and six basic emotions estimation.

To build a realistic digital embodiment of the user's face, our system only requires one photo image of the user. Then a state-of-the-art 3D face reconstruction approach [14] is applied to recover the dense 3D geometry as well as the texture of the user's face. The algorithm solves the reconstruction within a non-convex optimization process which takes photo consistency, sparse feature similarity and statistical regularization as constraints. The reconstruction is built upon a PCA-based morphable model [54]. FACS-based blendshapes [25], [32] and an illumination model based on Spherical Harmonics [55].

A template neutral facial mesh is warped to fit the reconstructed 3D face using a robust non-rigid registration method [15]. We first use a linear rigid-alignment based on 68 facial landmarks [56] to estimate the pose between the template and the reconstruction. Then a coupled global and local deformation is applied to each point on the template to conform it to the reconstruction. Personalized blendshapes are obtained

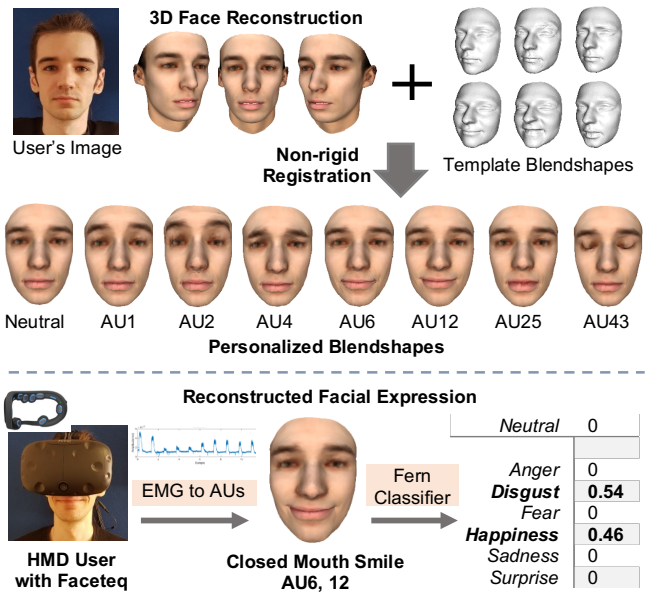

Fig. 4. An overview of our system. We first reconstruct a fully textured 3D face from a user's image. Then, template AU-based blendshapes are conformed to the reconstructed face to generate personalized blendshapes. With Faceteq and a robust learning method, our system predicts from EMG signals AU-coded facial expressions which are further reconstructed by fusing personalized blendshapes. A fern classifier is learned for estimating emotions from AUs.

by simply transferring the deformations from a series of generic AU-based blendshapes [48] to the deformed template neutral face as they share the same topology. During the non-rigid registration process, the facial texture is transferred simultaneously with the geometric deformations based on the built point correspondences.

Eight-channel EMG signals from the facial sensing hardware are then fed into a learning method to predict common AU-coded facial expressions which drive personalized blendshapes to generate realistic facial expressions during online tracking. The learning method applies the Least-square Support Vector Machine (LS-SVM) on RMS features from EMG signals. Training and testing are conducted on a database collected from 15 confirmed mentally and physically healthy participants without any signs of conditions that could affect their face and thus facial expressions.

To get a deeper insight into the VR HMD user's internal emotional states, a probabilistic model is built to map AUs to six basic emotions. We use a fern classifier to model the posterior probability of basic emotions given combinations of AUs. The fern classifier is learned from CK+ [57] and EmotioNet [58].

IV. Face Embodiment Construction

With a single image, we build a digital embodiment of the user's face that contains a series of fully-textured 3D facial meshes in neutral pose or with AUs.

A. 3D Face Reconstruction

Following Thies et al. [14], we convert the 3D face reconstruction from a single image into an inverse-rendering problem which is solved through an analysis-by-synthesis

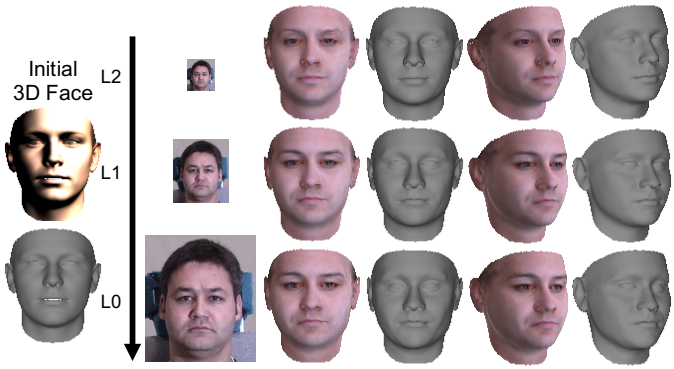

Fig. 5. Reconstruction results from each level of the image pyramid. The initial 3D face is set with mean identity, neutral pose, mean albedo and rendered with the lighting model whose 27 Spherical Harmonics (SH) coefficients are all set with 1. The facial texture here is rendered with estimated albedo and lighting. Face images have been cropped and resized for better presentation, the original face image is 640*480.

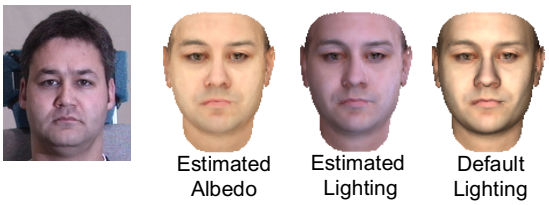

Fig. 6. Reconstructed facial texture. The right two images show faces rendered with the estimated lighting and the default lighting used in this paper.

process. We use a multi-linear PCA face model based [25], [32], [54] which has $n = 53$k vertices and 106k faces:

$$M_{geo}(\alpha_{id}, \alpha_{exp}) = \bar{S} + S_{id} \cdot \alpha_{id} + S_{exp} \cdot \alpha_{exp} \quad (1)$$

$$M_{alb}(\alpha_{alb}) = \bar{T} + T \cdot \alpha_{alb} \quad (2)$$

This parametric face model has three dimensions, where identity and expression represent the facial geometry and the third dimension represents the skin reflectance (albedo). It assumes the geometry and albedo obey a multivariate normal distribution centered at the average shape $\bar{S} = \bar{S}_{id} + \bar{S}_{exp} \in \boldsymbol{\mathcal{R}}^{3n}$ and reflectance $\bar{T} \in \boldsymbol{\mathcal{R}}^{3n}$. The corresponding bases are $S_{id} \in \boldsymbol{\mathcal{R}}^{3n\times 99}$, $S_{exp} \in \boldsymbol{\mathcal{R}}^{3n\times 29}$ and $T \in \boldsymbol{\mathcal{R}}^{3n\times 99}$. Standard deviations are $\sigma_{id} \in \boldsymbol{\mathcal{R}}^{99}$, $\sigma_{exp} \in \boldsymbol{\mathcal{R}}^{29}$ and $\sigma_{alb} \in \boldsymbol{\mathcal{R}}^{99}$.

The face is assumed to have the Lambertian surface reflectance and the illumination is modelled with a second order Spherical Harmonics (SH) denoted by $L \in \boldsymbol{\mathcal{R}}^{27}$ [55]. A face image I_{syn} is synthesized by rasterizing the parametric face model under a rigid transformation (R, t) and a perspective projection $\Pi_P(M_{geo})$ with the camera parameters K.

In an analysis-through-synthesis loop, face model and rendering parameters are optimized mainly along the direction of generating a face image as close as the input image. The objective function is formulated as:

$$E(\mathcal{P}) = w_{col}E_{col}(\mathcal{P}) + w_{lan}E_{lan}(\mathcal{P}) + w_{reg}E_{reg}(\mathcal{P}) \quad (3)$$

where $\mathcal{P} = \{\alpha_{id}, \alpha_{exp}, \alpha_{alb}, R, t, K, L\}$. $w_{col} = 1$, $w_{lan} = 10$ and $w_{reg} = 2.5\times10^{-5}$ are weights to balance three energy terms.

The photo-consistency term E_{col} measures the colour distance between the synthesized face image and the input image:

$$E_{col}(\mathcal{P}) = \frac{1}{|V_I|}\sum_{v_I\in V_I}\left\|I_{syn}(v_I) - I_{in}(v_I)\right\|_2 \quad (4)$$

$$I_{syn}(v_I) = \left[I_r^v, I_g^v, I_b^v\right]^T$$

$$I_{ch}^v = M_{alb,\,ch}^v \cdot \sum_{i=1}^{9} \gamma_{i,ch} y_i(\boldsymbol{n}(v)),\ ch \in \{r, g, b\}$$

$$v_I = \Pi_P(Rv + t)$$

where I_{in} is the input image and $v_I \in V_I$ denote all visible pixel locations in the synthesized image I_{syn}, v_I is obtained by projecting the visible 3D vertex v on the face mesh onto the image plane. Its pixel value $I_{syn}(v_I)$ is assigned with v's texture value - I_{ch}^v, where y_i is the ith SH basis function, $\gamma_{i,ch}$ is the corresponding SH coefficient of a specific color channel – ch, $\boldsymbol{n}(v)$ is the normal of v, and $M_{alb,\,ch}^v$ is the albedo value of v in channel ch.

The landmark-fitting term E_{lan} enforces a constraint to the reconstructed facial geometry according to some fiducial facial points, namely projected 3D landmarks, which should align to the corresponding landmarks on the input face image as accurate as possible:

$$E_{lan}(\mathcal{P}) = \frac{1}{|F|}\sum_{f_i\in F}\|f_i - \Pi_P(Rv_i + t)\|_2^2 \quad (5)$$

f_i is a 2D facial landmark detected on the input face image from our implementation of [59]. To ensure the plausibility of the reconstructed 3D face, a statistical regularization term is used to restrict face model parameters to a reasonable range:

$$E_{reg}(\mathcal{P}) = \sum_{i=1}^{99}\left[\left(\frac{\alpha_{id,i}}{\sigma_{id,i}}\right)^2 + \left(\frac{\alpha_{alb,i}}{\sigma_{alb,i}}\right)^2\right] + \sum_{i=1}^{29}\left(\frac{\alpha_{exp,i}}{\sigma_{exp,i}}\right)^2 \quad (6)$$

The objective function is transformed with the method of Iteratively Reweighted Least Squares (IRLS) and optimized using a Gauss-Newton (GN) solver. In our implementation, the optimization converges within 7, 5 and 3 GN steps from the coarsest to the finest level of a three-level image pyramid (see Fig. 5). For generating personalized blendshapes, the expression component will be removed from the reconstructed 3D face. Please also note that the reconstructed facial texture presented in this paper only keeps the estimated albedo and is rendered with a default lighting (see Fig. 6). The estimated lighting is discarded because it only models the lighting of the input face image, while the reconstructed 3D face should be rendered with the lighting of the virtual space for a more realistic face embodiment.

B. Personalized Blendshapes Generation

To get a digital face embodiment with AUs, we adopt a robust non-rigid registration algorithm [15] and a series of template blendshapes [48]. The template blendshapes are based on 3D scans of three female FACS certified actors using a 4D stereo imaging system [60]. Each actor performed 20 to 30 AUs, providing a total of 37 AUs (counting lateralizations) as well as a neutral face. The template mesh consists of 4,735 vertices and 8,760 faces. In this work, we use seven AUs (AU1, AU2, AU4, AU6, AU12, AU25 and AU43) for association with and prediction from the available EMG signals. We calculate the mean of all actor's meshes to get a more general template (see Fig. 4).

The non-rigid registration conforms the template neutral face (source point cloud) to the reconstructed 3D face (target point

cloud). A coupled global and local deformation is applied to the vertex v_i on the source point cloud:

$$\tilde{\mathrm{v}}_i = \Phi_{local} \circ \Phi_{global}(\mathrm{v}_i) \quad (7)$$

A rotation matrix R relative to the centre-of-mass m and a translation vector t define the global rigid deformation:

$$\Phi_{global}(\mathrm{v}_i) = R(\mathrm{v}_i - m) + m + t \quad (8)$$

The non-rigid deformation is defined by a set of deviation vector d_i:

$$\Phi_{local}(\mathrm{v}_i) = \mathrm{v}_i + \mathrm{d}_i \quad (9)$$

For each vertex v_i on the source point cloud, we associate a corresponding position $\mathrm{c}(\mathrm{v}_i)$ on the target which is initialized with a closest point computation and updated iteratively within the optimization. The non-rigid registration can hence be casted as an unconstrained energy minimization problem with unknowns $\mathcal{K} = \{R, t, \mathbf{D}, \mathrm{c}\}$, where $\mathbf{D} = \{\mathrm{d}_i\}$. The objective function is formulated as:

$$E(\mathcal{K}) = w_{fit}E_{fit}(\mathcal{K}) + w_{smooth}E_{smooth}(\mathcal{K}) \quad (10)$$

The weights w_{fit} and w_{smooth} compensate for different scales of the energy terms.

E_{fit} measures the corresponding point distance between the source point cloud and the target point cloud:

$$E_{fit}(\mathcal{K}) = \sum_{i=1}^{n} w_{conf,i}^2 \|\tilde{\mathrm{v}}_i - \mathrm{c}(\mathrm{v}_i)\|_2^2 + \sum_{i=1}^{n}\left(1 - w_{conf,i}^2\right)^2 \quad (11)$$

where n is the number of correspondences and $w_{conf,i}$ is the confidence of the correspondence. And $w_{conf,i}$ close to one indicates a reliable correspondence, while $w_{conf,i}$ close to zero indicates that no proper correspondence is found. Source vertices without valid correspondence are excluded from the optimization process. The texture of the reconstructed face is transferred to the neutral template using the correspondence as well. To enhance the surface smoothness, an energy term enforcing small changes of point neighbourhoods and triangle areas is augmented to the objective function:

$$E_{smooth}(\mathcal{K}) = E_{neigh}(\mathcal{K}) + E_{area}(\mathcal{K}), \quad (12)$$

$$E_{neigh}(\mathcal{K}) = \sum_{i=1}^{n}\sum_{j=\mathcal{N}(i)}\left(\left\|\tilde{\mathrm{v}}_i - \tilde{\mathrm{v}}_j\right\|_2 - \left\|\mathrm{v}_i - \mathrm{v}_j\right\|_2\right)^2 \quad (13)$$

$$E_{area}(\mathcal{K}) = \sum_{i=1}^{n}\left(\mathcal{A}(\tilde{\mathrm{v}}_i) - \mathcal{A}(\mathrm{v}_i)\right)^2 \quad (14)$$

where $\mathcal{N}(i)$ is the one-ring neighbourhood and $\mathcal{A}(\mathrm{v}_i)$ is the summing area of triangles attached to v_i on the source mesh.

We use the Levenberg-Marquardt algorithm to solve the non-linear least squares problem above. After obtaining the deformed neutral template, we calculate deviations between the original neutral template and template blendshapes. These deviations are then transferred to the deformed neutral template to generate personalized blendshapes. The texture of the deformed neutral template is transferred at the same time, providing fully-textured personalized blendshapes.

V. Facial Expressions from EMG Signals

Facial expressions are results of facial muscle movements. The pioneering hardware solution – Faceteq offers an efficient way to sense facial muscle contractions through eight integrated EMG sensors placed on the ESFP. However, mapping EMG signals to facial expressions is nontrivial. To this end, we recruit 15 subjects and use the Faceteq interface for data collection and analysis. We define all facial expressions according to action units (AUs), which enables a convincing facial expression recovery in subsequent steps.

Fig. 7. AU-coded facial expressions studied in this work. From left to right: forehead wrinkle, frown, eye closure, close mouth smile, and open mouth smile.

A. Data Collection

In this study, 15 volunteers are recruited (11 male and 4 female), aged from 21 to 52 years old (Mean: 31.93, Std : 12.75). Each participant is asked to perform five common facial expressions (see Fig. 7) – closed mouth smile, eye closure, forehead wrinkle, frown and open mouth smile, while wearing the prototype of the EMG-based facial sensing interface. All the facial expressions are defined with AU combinations following the work [3]: AU6, 12 for closed mouth smile, AU43 for eye closure, AU1, 2 for forehead wrinkle, AU4 for frown and the combination of AU6, 12 and 25 for open mouth smile. Eight surface dry EMG electrodes are placed symmetrically on the left and right side of the Faceteq interface, providing eight-channel EMG signals. These EMG electrodes monitor activations of specific facial muscles (see Fig. 3), including Zygomaticus major (channel 1&2), Frontalis (channel 3&4), Orbicularis oculi (channel 5&6) and Corrugator (channel 7&8).

An audio track is used to instruct the subject to make facial expressions or return to the neutral pose. Each facial expression is repeated ten times with each lasting for two seconds. There is a ten-second rest between two adjacent repetitions. EMG signals are recorded at 1kHz sampling rate, resulting in $2\times10\times1{,}000 = 20{,}000$ EMG samples for each facial expression of each subject in theory. The actual EMG sample amount may drift around the estimated value as there is latency in starting or ending the facial expression for each subject.

B. EMG Signal Pre-Processing

Facial EMG signals have small amplitude and can be easily interfered by various external or internal factors, such as motion artefacts, incorrect sensor placement and environmental noise. We therefore use multiple filters to clean the raw EMG data. A baseline correction on raw EMG signals is first adopted to remove mean values and the linear trend. To eliminate artefacts such as the line interference introduced by electrical devices, the Notch filter is then applied to remove the 50Hz component and its harmonics up to 350Hz. Signals are further passed through a band-pass filter retaining components from 30 to 450Hz. Finally, we obtain the eight-channel clean EMG signals (see Fig. 8).

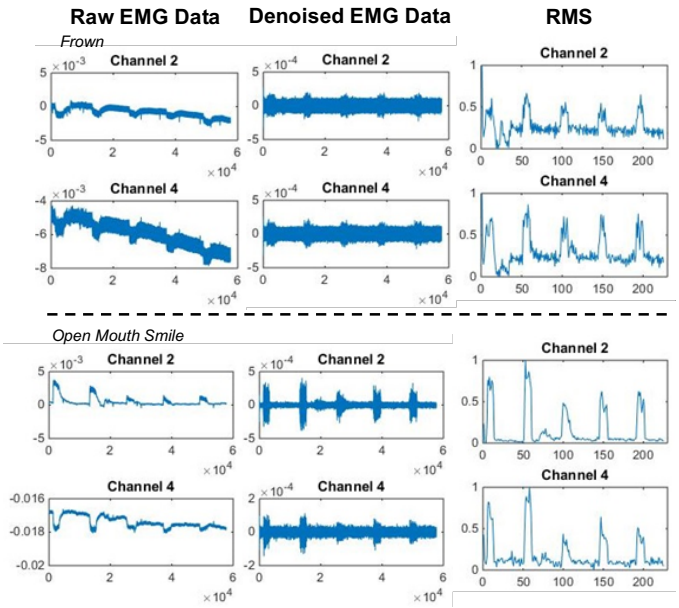

Fig. 8. EMG signals and RMS features. The left two columns compare raw and denoised EMG data. The right column shows the RMS feature extracted from the denoised EMG data within a 256-msec time window.

C. Feature Extraction

To reduce the dimensionality of data and extract the most informative segments, it is crucial to compress EMG signals along the time axis. Generally, EMG signals are partitioned into temporal segments of the same length, from where features are extracted. Long segments can supress bias and variance of the feature, however, they may fail to reach the efficiency requirement [61]. Some recent works report that using segments with 256 msecs length is a good trade-off between the feature effectiveness and the overall processing efficiency [62], [63]. We follow the setting of [62] by segmenting the pre-processed EMG signals into non-overlapping 256-msec pieces.

Root Mean Square (RMS) is one of the representative time-domain features and has been widely used for analysing the contraction of facial muscle. With the hypothesis of the Gaussian random process, RMS provides the maximum likelihood estimation of EMG amplitude when a facial muscle is under constant force and non-fatiguing contraction. According to a recent survey [62], RMS shows superiority against the other time-domain features. We hence extract RMS from each 256-msec segment of EMG signals:

$$RMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2} \qquad (15)$$

where $N = 256$, x_n is an EMG sample within the 256-msec segment.

D. Facial Expression Prediction

Multiple external or internal factors such as EMG electrode drift, individual variances or muscle fatigue, usually result in a large variation of the EMG pattern even for the same facial expression. Hence, a robust learning algorithm is required for mapping EMG features to facial expressions accurately. The influence of the classifier to the final prediction performance has been much studied. A recent study [62] compares 14 classifiers and reports that Least-square Support Vector Machine (LS-SVM), Regularized Discriminative Analysis (RDA), Normal Density Discriminant Function (NDDF) and Maximum-likelihood (ML) estimation provide a much higher classification accuracy against the other classifiers. ANOVA statistical analysis shows that there is no significant difference among the classification performance of the top four classifiers [64]. In this work, we choose LS-SVM as the classifier and adopt the libSVM [65] framework to train the multiclass LS-SVM.

VI. Basic Emotions Prediction form AUs

Existing rule-based methods [51], [52] cannot be extended to scenarios where observed AUs are not included in the established heuristic rules. Restricted by the number of EMG sensors applied and the range of facial expressions covered by the collected database, our system outputs specified AUs. It makes previous rule-based methods infeasible. To address this problem, we propose to use the fern classifier to model the relationship between AUs and six basic emotions. Specifically, our target is to learn the posterior probability of emotions given occurrences of AUs. It is a typical Bayesian classification problem that can be solved efficiently with the fern classifier.

Fern classifier has been successfully applied in image key-points recognition [53]. Each fern is a composition of a small set of features and a series of binary tests on these features. It returns the probability that a sample belongs to a class. Outputs from all ferns are then combined together in a Naïve Bayesian way. In our task, six basic emotions - anger, disgust, fear, happiness, sadness and surprise are treated as classes, while the occurrence of AU is a binary feature. Since the feature pool consists of limited AUs, there is no need to partition it into groups of features. One fern is sufficient to learn the class-conditional distribution in our case. Let $c_i, i = 1, \dots, H$ be the set of class (emotion) and $f_j, j = 1, \dots, N$ be the set of binary feature (AU occurrence), we are looking for:

$$\hat{c}_i = \underset{c_i}{\operatorname{argmax}}\, P\left(C = c_i \mid f_1, f_2, \dots, f_N\right) \qquad (16)$$

$$f_j = \begin{cases} 1 & if\ AU_j\ occurs \\ 0 & otherwise \end{cases}$$

where C represents the class. With Bayes' theorem, we have

$$P(C = c_i \mid f_1, f_2, \dots, f_N) = \frac{P(f_1, f_2, \dots, f_N \mid C = c_i) P(C = c_i)}{P(f_1, f_2, \dots, f_N)} \qquad (17)$$

$$P(f_1, f_2, \dots, f_N) = \sum_{i=1}^{H} P(f_1, f_2, \dots, f_N \mid C = c_i) P(C = c_i)$$

$P(C = c_i)$ is the prior probability of emotion.

We build the fern classifier on two benchmark facial expression databases - CK+ [57] and EmotioNet [58] that contain both AU and basic emotion labels.

CK+ involves 123 subjects who are instructed to perform 23 facial expressions forming a database of 593 image sequences. Each sequence incorporates the onset (the neutral face) to peak formation of the facial expression. The peak frame of the facial expression is coded with AU and seven basic emotion labels (anger, contempt, disgust, fear, happiness, sadness and surprise) which are further rectified according to the FACS manual [3].

Overall, CK+ offers 327 samples with both AU and basic emotion labels for our study. As contempt is beyond the scope of this study, we removed samples labelled as contempt, leaving 309 samples in total for subsequent analysis.

EmotioNet consists of a million in-the-wild images of facial expressions in which 975,000 images are made available to the public. Within the released database, there are 950,000 images automatically annotated with AUs and AU intensities. The remaining 25,000 are manually annotated with AUs by qualified coders. A small part of these images (2,479 images) are labelled with one of 16 compound emotions defined in [66] based on AU combinations. Since we only consider six basic emotions in this study, we relabelled images according to their compound emotion labels. For example, if an image has been annotated as happily surprised, we categorize it into both happiness and surprise. In total, there are 3,581 samples available with AU and basic emotion labels.

CK+ and EmotioNet, covering a wide range of AU-emotion relations, are appropriate for statistical analysis. Experimental results demonstrate the proposed method is able to give valuable emotion information when only limited AUs are available. This function is hence incorporated into our system to assist the VR HMD user to understand the other users' emotions in the virtual world.

VII. Results and Analysis

The proposed system was developed to enable realistic facial expression reconstruction for the VR user wearing a HMD. It was validated on 13 subjects. Each of the three principal system parts – face embodiment construction, facial expression prediction and basic emotions estimation, has been carefully evaluated. In the following, we will report experimental results of each part and the overall system performance afterwards.

Face Embodiment Construction. We evaluated the robustness of this part using face images from various subjects. The results are shown in Fig. 9. With a single face image of the user, our system is able to reconstruct a fully textured 3D face and generate vivid AU-based blendshapes.

Since each AU originates from 3D scans of FACS certified actors, the generated blendshapes form a solid basis for natural facial expression composition. To enable deformation transfer between AU and the neutral face, we removed facial expression components from the reconstructed 3D face.

Facial Expression Prediction. Predicting facial expressions from dry EMG signals is not easy as the raw signals are quite noisy. We thus applied aforementioned multiple processing steps to clean the raw signals. Fig. 8 compares the raw EMG signal and the denoised signal. EMG signals from channel 2 and channel 4 when collecting data for frown and open mouth smile are plotted for illustration. After obtaining clean EMG signals, we extracted the RMS feature within a 256-msec time segment using Eq. 15 (see Fig. 8 for example). As the instructional audio track was started manually, there were some variability in the time recorded and therefore the data. It also had a very short time delay to make or stop the facial expression once the subject heard the prompt tone. Therefore, it is infeasible to accurately

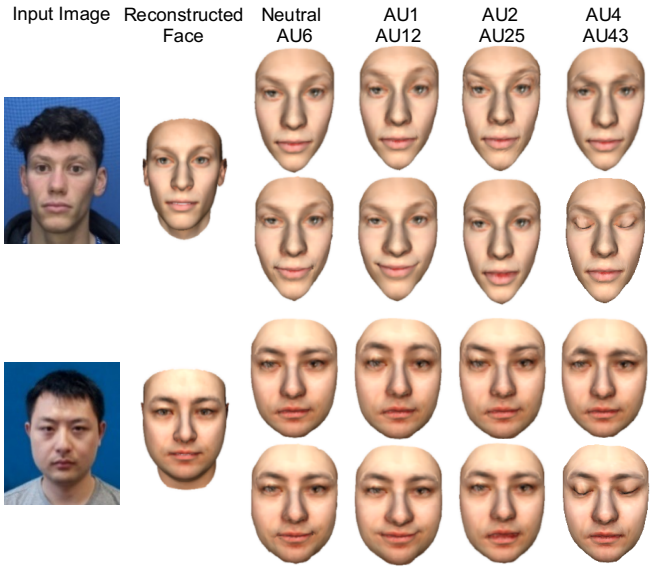

Fig. 9. Realistic face embodiment generation from a single image. The reconstructed face shown above has discarded facial expression components.

annotate the RMS samples with correct facial expression labels according to the timestamp. To separate facial expression samples from neutral samples, we calculated a RMS threshold - $mean(RMS) + n \times std(RMS)$ (shown as the red line in Fig. 10) for each EMG channel [67], where n is an empirical value set manually according to each RMS curve. In our experiment, $n = 0.25$ works well for all tests. Samples whose RMS values are above at least one threshold line were annotated with the specific facial expression label, while the others were annotated as neutral. This process extracted the most significant facial expression samples while filtering out samples with insignificant RMS features which were probably caused by the distraction of the subject or other noise-related interferences.

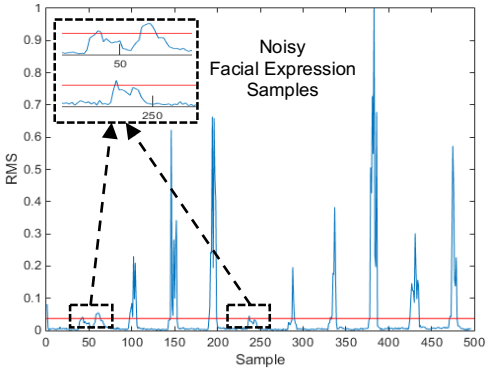

Fig. 10. Representative RMS values of channel-1 EMG signals of the closed mouth smile. The red line indicates the reference RMS level which is used to differentiate facial expression samples from neutral samples. The black dashed box indicates noisy facial expression samples that were labeled as neutral.

TABLE I
The Distribution of Labelled Facial Expression RMS Samples

	CMS	EC	FW	FR	OMS
Sample Number	2061	2416	1940	2483	1853

CMS: closed mouth smile EC: eye closure FR: frown
FW: forehead wrinkle OMS: open mouth smile

TABLE II
FACIAL EXPRESSION RECOGNITION ACCURACY FROM RMS FEATURES

Classification Accuracy							
	E2						E1
	CMS	EC	FW	FR	OMS	ALL	ALL
S1	0.16	0.24	0.96	1.00	0.01	0.60	0.73
S2	0.89	0.45	0.97	0.99	0.43	0.77	0.89
S3	0.28	0.77	0.98	1.00	0.66	0.74	0.88
S4	0.86	0.43	0.96	1.00	0.16	0.73	0.84
S5	0.42	0.62	0.95	0.82	0.86	0.68	0.94
S6	0.75	0.85	0.97	1.00	0.16	0.82	0.94
S7	0.81	0.84	1.00	1.00	0.24	0.76	0.78
S8	0.55	0.93	0.99	1.00	0.74	0.87	0.90
S9	0.38	0.96	0.99	0.80	0.65	0.82	0.89
S10	0.30	0.97	0.99	1.00	0.92	0.87	0.91
S11	0.97	0.88	1.00	1.00	0.28	0.80	0.91
S12	0.87	0.32	0.98	1.00	0.18	0.70	0.78
S13	0.65	0.47	0.73	0.59	0.54	0.51	0.79
S14	0.71	0.91	0.99	0.98	0.71	0.87	0.90
S15	0.95	0.87	0.97	1.00	0.23	0.83	0.86
E1	0.79	0.91	0.99	0.99	0.55		

Note: 1) E1-E2 represent experiments; 2) The results of E2 are in orange and the results of E1 are in purple. 3) S1-S15 represent subjects; 4) ALL is the classification accuracy on all the testing data, including neutral samples.

The overall distribution of facial expression RMS samples is listed in Table I. To balance the neutral and facial expression samples in the data set, we randomly selected the same amount of neutral samples as the facial expression samples for each facial expression of each subject. We did two experiments to validate the accuracy of facial expression prediction from RMS features. For both experiments, we did max-min normalization for all the RMS features of the data using the max and min RMS values of the training data [68]. The max and min RMS values were chosen for each EMG channel and each facial expression.

First, we used the whole dataset and randomly selected 80% data for training and the left 20% data for testing [62], [64]. The LS-SVM with the RBF kernel was adopted for the target multi-class classifier. 10-fold cross validation was applied to select the hyper-parameters C and $gamma$ which were set as 2 and 32 respectively in our experiment. The classification accuracy on the whole testing set is 86.77%. We also calculated the classification accuracy for each subject or each facial expression in the testing set (see E1 in Table II). The average classification accuracy across subjects is 86.27% with a standard deviation of 0.0644. The average classification accuracy across expressions is 84.74% with a standard deviation of 0.1867.

To further validate the generality of the facial expression classifier to EMG signals from new subjects, we applied leave-one-out validation. Specifically, we left one subject's data for testing while using the data from the other subjects for training. We repeated the same process for each subject, which resulted in 15 different classifiers. The performance of the classifier is shown in E2 in Table II. The mean classification accuracy for 15 classifiers is 75.8% with a standard deviation of 0.1033. To validate the effectiveness of the EMG signal denoising step and the RMS feature, we conducted the same experiment as in E2 on denoised EMG signals and RMS features extracted from raw EMG signals. When the denoised EMG signals were fed directly into the classifier, the classification accuracy declines significantly, with a mean of 20.74% and a standard deviation of 0.0316. The results also show that facial expressions could not be accurately predicted from RMS features extracted from raw EMG signals (mean: 25.16%, standard deviation: 0.2016).

As shown in Table II, both experiments show high classification accuracy for FW and FR, while lower accuracy for CMS and OMS. We can also find that the classification accuracy varies from one subject to another. For example, S8 and S10 achieve high accuracy in both experiments, while S1 and S13 have a much lower recognition rate, even when their data was included in training in E1. This is probably due to EMG signals from some subjects are noisier or contain patterns that are quite different from EMG signals of the other subjects. This can be remedied by collecting data from more subjects and improving the data capture process.

Basic Emotions Estimation. All facial expressions involved in this study have been coded in AUs, providing a detailed and anatomic description of facial expressions. AUs describing the physical appearance of facial display offer valuable clues for predicting six basic emotions. Our system outputs AUs – AU1, AU2, AU4, AU6, AU12, AU25 and AU43 focusing on the ESFPs, which makes previous rule-based methods intractable. To get an insight into the HMD user's internal emotional states, we built the relationship between AUs and six basic emotions with a probabilistic model – the fern classifier. Following Bayes' theorem, the model estimates the posterior probability of a basic emotion when observing a specific group of AUs.

We learned the probabilistic model from two benchmark FACS-annotated facial expression databases – CK+ [57] and EmotioNet [58]. There are only two samples observing AU43 in CK+, while AU43 is not used when defining compound emotion category in EmotioNet. We hence discarded AU43 when estimating basic emotions.

As shown in Table III, the number of samples belonging to each emotional category varies from each other. If we regard the prior probability of an emotion as the proportion of samples belonging to it, the posterior probability of emotion given an AU combination will become the proportion of samples coded in the current AU combination within the whole database (see Eq. 17). This will cause large deviations when calculating probabilities. We therefore assumed that basic emotions have identical prior probabilities.

After applying Eq. 17, we have obtained probabilities of basic emotions given the occurrence of AU for both databases (see Table III). Table III includes observed combinations of AUs that are used to define facial expressions in this work. From the results, we found the following phenomena:

1) Emotions can be expressed in various forms of facial expressions. As can be seen from the table, when none of AUs specified in our work occur, emotions such as disgust, anger, surprise can still be observed. On the other side, a combination

TABLE III
THE PROBABILITY OF EMOTION GIVEN AUS LEARNED FROM CK+ AND EMOTIONET

CK+ (309)						
AU code AU25_AU12_AU6_AU4_AU2_AU1	Anger (45)	Disgust (59)	Fear (25)	Happiness (69)	Sadness (28)	Surprise (83)
'000000'	0.0655	0.8990	0	0	0	0.0355
'000001'	0	0	0	0	1	0
'000011' - AU1, 2 (3)	0	0	0	0	1 (3)	0
'000100' - AU4 (54)	0.6839 (35)	0.2533 (17)	0	0	0.0628 (2)	0
'000101'	0	0	0.0618	0	0.9382	0
'000111'	0	0	0.2188	0	0.7183	0
'001000'	0.7387	0.1409	0	0.1204	0	0
'001100'	0.2875	0.7125	0	0	0	0
'010100'	1	0	0	0	0	0
'011000' - AU6, 12 (2)	0	0.5391 (1)	0	0.4609 (1)	0	0
'100000'	0	0.3897	0	0.3332	0	0.2770
'100011'	0	0	0.1471	0	0	0.8529
'100100'	0	0.3610	0.6390	0	0	0
'100101'	0	0	1	0	0	0
'100111'	0	0	0.9300	0	0	0.0700
'110000'	0	0.3690	0	0.6310	0	0
'111000' - AU6, 12, 25 (64)	0	0	0	1 (64)	0	0
'111101'	0	0	1	0	0	0
EmotioNet (3,581)						
	Anger (289)	Disgust (977)	Fear (150)	Happiness (1536)	Sadness (359)	Surprise (270)
'000000'	0	1	0	0	0	0
'000100' - AU4 (760)	0.3268 (173)	0.1274 (228)	0	0	0.5459 (359)	0
'100011'	0	0	0.4348	0	0	0.5652
'100100'	0.4099	0	0.3404	0	0	0.2496
'100101'	0	0	1	0	0	0
'110000'	0	0.4274	0	0.5726	0	0
'110011'	0	0	0	0.1495	0	0.8505

Note: 1) AU code: '1' indicates the AU occurs, '0' indicates the AU doesn't occur; 2) the number in parentheses denotes the amount of samples belonging to the category

of AUs can describe several basic emotions. For instance, in CK+, AU6 and AU12 indicate both disgust and happiness, while the results of EmotioNet show that we can probably observe AU4 and AU25 for anger, fear and surprise. Most combinations of AUs used in our work, namely AU6,12, AU1,2 and AU4 etc. are not discriminative for predicting an emotion category.

2) A few AUs or combinations of AUs show stronger links to emotions than the others. In CK+, AU1,4 indicates a high probability (0.9382) of sadness, while AU6,12,25 indicates a high probability (1.0) of happiness. In both CK+ and EmotioNet, AU4 normally indicates a negative emotion, such as anger, disgust and sadness. We can find that the specific probability of each emotion when AU4 occurs differs between CK+ and EmotioNet. It is mainly due to the variety of facial expressions of emotions, which leads to a single database only covering a limited range of AU-emotion relationships. This also explains the first phenomena. Some other AUs or AU combinations, e.g. AU25 can be found in all six basic emotions.

To further verify the second phenomena, we calculated the discriminative power of each AU to an emotion [52]:

$$D = P(A_j|E_i) - P(A_j|\overline{E_i}) \quad (18)$$

where $P(A_j|E_i)$ is the probability of observing AU A_j when emotion E_i occurs, and $P(A_j|\overline{E_i})$ is the probability of observing A_j when E_i doesn't occur. D measures the relationship between the AU and the emotion. D close to -1 represents a strong negative correlation, while D close to 1 represents a strong positive correlation. We generated a correlation matrix of AUs and emotions expressed with the discriminative power D from

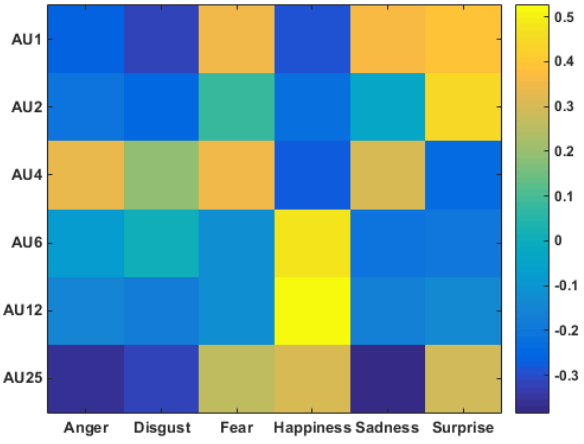

Fig. 11. Correlations between AUs and basic emotions. The value close to 1 represents a strong positive correlation, while the value close to -1 represents a strong negative correlation.

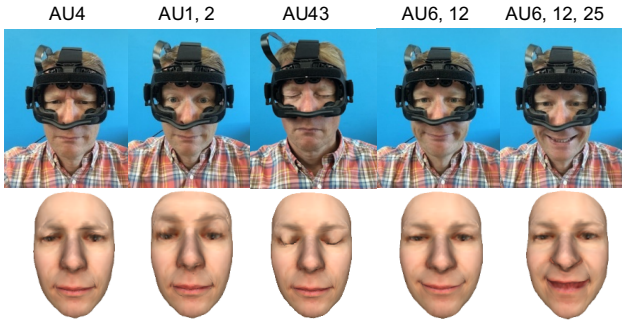

Fig. 12. Facial expressions sensed and reconstructed with the proposed system when the user is wearing the Faceteq prototype.

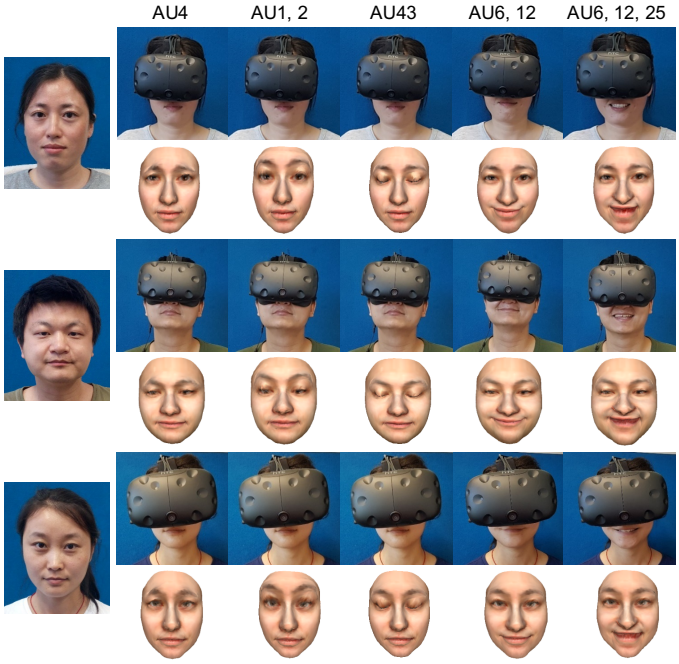

Fig. 13. Facial expressions sensed and reconstructed with the proposed system when the user was wearing the VR HMD integrated with the Faceteq.

CK+. Each *D* was normalized across all the AUs for each of the emotions. The original relation matrix contains 35 AUs and 7 emotions. We removed most matrix components while only keeping AUs and six basic emotions studied in this work. As shown in Fig. 11, AU1 associates closer to fear, sadness and surprise, while AU6 and AU12 have a distinctive connection with happiness. AU4 shows a closer relation with anger, fear and sadness. AU2 links closely to surprise. Consequently, the correlation matrix demonstrates the emotional saliency of the studied AUs and is consistent with the previous probabilistic model.

Overall, the above probabilistic analysis takes a deep insight into the relationship between AUs and six basic emotions. The built probabilistic model can provide useful emotional information when only limited AUs are available.

Full System Evaluation. We first tested the system with the prototype of Faceteq which was used to collect the EMG data of specified facial expressions. Since the prototype can be used independently from a VR HMD and hence doesn't occlude the user's principal face region, it provides direct comparisons between the reconstructed facial expression and the ground

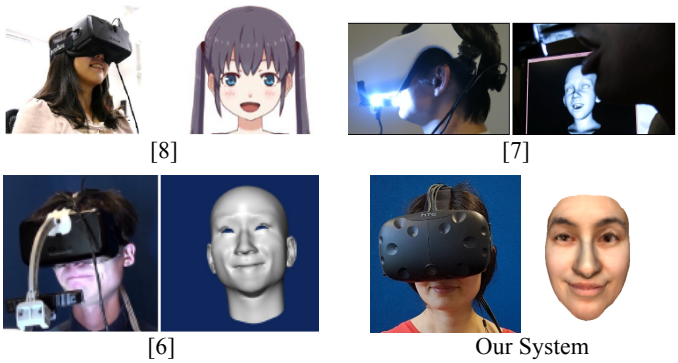

Fig. 14. Comparison with other similar systems from [6], [7] and [8]. Avatars in [6] and [7] capture the user's facial movements with a clumsy RGB/RGBD camera attached on the VR HMD, while not preserving the user's facial identity and texture. [8] simply uses a 2D cartoon image to represent the user's face.

truth. Example results are shown in Fig. 12. The facial sensing hardware can detect the user's facial expression through eight integrated EMG sensors. The facial expression is then mapped onto a realistic face embodiment of the user. As all the facial expressions used in this study are specified with AUs, we can get a detailed interpretation of the facial expression.

Then, we did validation tests with the Faceteq and the VR HMD. As the EMG-based facial sensing interface has been designed to softly enfold the wearer's face, the type of HMD will not affect the scale of the EMG data. During the test, each subject was asked to perform facial expressions specified in this study when wearing a commercial VR HMD [2] attached with the Faceteq hardware. Fig. 13 shows example results. By comparing with other comparable systems, our system can reconstruct a more realistic 3D face embodiment for the VR HMD user and doesn't require additional cameras to capture the user's facial movements (see Fig. 14). With the learned fern classifier, we can further obtain a probabilistic model between AUs and six basic emotions, which is important to VR applications.

Limitations. Since the biometric data collection step is labour-intensive and expensive, this paper focuses on specific AU combinations on the ESFP. Future work could be extended to the detection of various AUs independently from EMG signals from a larger biometric database involving 1) a wider selection of facial expression and 2) additional sensor inputs such as from an eye tracker. Then, any combination of these AUs obtained would be able to cover a wider range of facial expressions of emotion. Furthermore, the current system does not encode the intensity of facial expression [69]. Both problems can be alleviated by collecting EMG data for single AUs with different scales of intensity. The current system mainly focuses on facial expressions displayed on the upper-half face, with the exception of AU12 and AU25 which are sensed from the cheek. As such certain fundamental expressions occurring around the mouth area are ignored, in the future, we will also look to infer AUs associated with visemes during speech [70], such as AU24, AU25, AU26 and AU27.

Restricted by the required EMG signal processing, the system described above cannot attain a real-time performance. The original EMG data should be partitioned into a series of time sequences for feature extraction. Short sample windows

lead to bias and variance in feature estimation, while long sample windows reduce the system efficiency. We applied a 256-msec time window, which means that only about four (3.9) frames can be output from the system in a second. This issue could potentially be alleviated by using shorter or overlapping time windows, or applying EMG sensors with a higher sampling rate.

The digital embodiment of the VR HMD user output from our system is restricted to the frontal face region and still has significant room for improvement. A more compelling full head embodiment could be constructed by modelling hair [71], texture [46] and shape details [34], [45].

VIII. Conclusions

We have proposed a method and developed a prototype system that can sense and reconstruct the VR HMD user's facial expression. Our hardware component is portable and compatible with mainstream VR HMDs. It can detect facial muscle movements accurately with eight integrated EMG sensors placed on the ESFP. With a single face image, the system can reconstruct the user's fully textured 3D face and generate personalized AU-based blendshapes. Specifically, the system can capture AU-coded facial movements with integrated EMG sensors and a robust classifier learned from the data collected from 15 subjects. It can also provide useful emotional information for participants in the virtual world with a novel probability model of AUs and six basic emotions built with the fern classifier. We believe the developed system can facilitate a wide range of VR applications, such as game, physical therapy and rehabilitation.

In future, we plan to equip the system with the ability to detect independent AUs and its intensity. A significant sized database consisting of AU and corresponding EMG signals will be created. We will extend to AUs associated with visemes during speech to cover more facial expressions. The system could potentially be improved to achieve real-time performance with additional biometric sensors or more efficient signal processing methods. Supplementary improvements could involve features such as hairs, texture and geometric details for a compelling full-head digital embodiment for VR applications.

References

[1] OCULUS. Oculus Rift. 2018. https://www.oculus.com/.

[2] HTC. HTC Vive. 2018. https://www.vive.com/.

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

[4] L. Zheng, Y. Li, J-H Tao, J. Huang and M-Y Niu, "Expression Analysis Based on Face Regions in Read-world Conditions", *International Journal of Automation and Computin*g, p. 1-12, 2019.

[5] J. Cha, J. Kim and S. Kim, "An IR-based facial expression tracking sensor for head-mounted displays*", In Proceedings of the 15th IEEE SENSORS Conference*, pp. 1-3, 2016.

[6] H. Li, L. Trutoiu, K. Olszewski, L.-Y. Wei, T. Trutna, P.-L. Hsieh, A. Nicholls and C.-Y. Ma, "Facial performance sensing head-mounted display", *ACM Transactions on Graphics(ToG)*, vol. 34, no. 4, p. 47, 2015.

[7] K. Olszewski, J.J. Lim, S. Saito and H. Li, "High-fidelity facial and speech animation for vr hmds", *ACM Transactions on Graphic (ToG)*, vol. 35, no. 6, p. 221, 2016.

[8] K. Suzuki, F. Nakamura, J. Otsuka, K. Masai, Y. Itoh, Y. Sugiura and M. Sugimoto, "Recognition and mapping of facial expressions to avatar by embedded photo reflective sensors in head mounted display", *In Proceedings of the IEEE Virtual Reality Conference (VR)*, 2017.

[9] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt and M. Nießner, 2016b, "Facevr: Real-time facial reenactment and eye gaze control in virtual reality", *arXiv preprint arXiv:161003151,* 2016b.

[10] A. Gruebler and K. Suzuki, "Design of a wearable device for reading positive expressions from facial emg signals", *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 227-237, 2014.

[11] I. Mavridou, J.T. McGhee, M. Hamedi, M. Fatoorechi, A. Cleal, E. Ballaguer-Balester, E. Seiss, G. Cox and C. Nduka, "FACETEQ interface demo for emotion expression in VR", *in Proceedings of the IEEE Virtual Reality Conference (VR)*, pp. 441-442, 2017.

[12] B. Martinez, M. F. Valstar, B. Jiang and M. Pantic, "Automatic Analysis of Facial Actions: A Survey", *IEEE Transactions on Affective Computing*, 2017.

[13] P Ekman and W. V. Friesen, "Manual for the facial action coding system", *Consulting Psychologists Press*, 1978.

[14] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos", *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2387-2395, 2016a.

[15] S. Zhang, J. Dong and H. Yu, "Automatic 3D face recovery from a single frame of a RGB-D sensor", *in Proceedings of the 28th British Machine Vision Conference AFAHBU Workshop (BMVA)*, 2017.

[16] D. J. McFarland and J. R. Wolpaw, "Brain-computer interfaces for communication and control", *Commun. ACM*, vol. 54, no. 5, pp. 60-66, 2011.

[17] Looxid Labs. 2018. http://looxidlabs.com/.

[18] J.F. Cohn and K. Schmidt, "The timing of facial motion in posed and spontaneous smiles". *Active Media Technology*, pp. 57-69, 2003.

[19] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces", in Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, ACM, pp. 187-194, 1999.

[20] S. Romdhani and T. Vetter, "Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior", *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[21] M. Sela, E. Richardson and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation", *in IEEE International Conference on Computer Vision (ICCV)*, 2017.

[22] S. Sengupta, A. Kanazawa, C. D. Castillo and D. W. Jacobs, "SfSNet: Learning Shape, Reflectance and Illuminance of Faces "in-the-wild"", arXiv preprint arXiv:1712.01261, 2017.

[23] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner and C. Theobalt, "State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications", *Computer Graphics Forum*, vol. 37, no. 2, pp. 523-550, 2018.

[24] L. Jiang, J.-Y. Zhang, B.-L. Deng, H. Li and L.-G. Liu, "3D Face reconstruction with geometry details from a single image", *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4756-4770, 2018.

[25] X.-Y. Zhu, Z Lei, X.-M. Liu, H.-L. Shi and S. Z. Li, "Face alignment across large poses: A 3d solution", *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] J. Booth, A. Roussos, A. Ponniah, D. Dunaway and S. Zafeiriou, "Large scale 3d morphable models", *International Journal of Computer Vision,* vol. 126, no. 2-4, pp. 233-254, 2018.

[27] S.-Y. Cheng, I. Kotsia, M. Pantic and S. Zafeiriou, "4DFAB: A Large Scale 4D Database for Facial Expression Analysis and Biometric Applications", *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2018.

[28] T.-Y.Li, T Bolkart, M. J. Black, H Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans", *ACM Transactions on Graphics(ToG)*, vol. 36, no. 6, p. 194, 2017.

[29] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis and S. Zafeiriou, "3D face morphable models "in-the-wild"", *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[30] I. Kemelmacher-Shlizerman and R. Basri, "3D face reconstruction from a single image using a single reference face shape", *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 33, no. 2, pp. 394-405, 2011.

[31] O. Aldrian and W. A. P. Smith, "Inverse Rendering of Faces with a 3D Morphable Model", *IEEE transactions on pattern analysis and machine intelligence(TPAMI)* , vol.35, no. 5, pp. 1080-1093, 2013.

[32] C. Cao, Y.-L. Weng, S. Zhou, Y.-Y. Tong and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing", *IEEE TVCG*, vol. 20, no. 3, pp. 413-425, 2014.

[33] L. Yin, X.-Z. Wei, Y. Sun, J. Wang and M. J. Rosato, "A 3D facial expression database for facial behavior research", *in Proceedings of IEEE International Conference on Automatic face and gesture recognition (FG)*, pp. 211-216, 2006.

[34] Y.-D. Guo, J.-Y. Zhang, J.-F. Cai, B.-Y. Jiang and J.-M. Zheng, "CNN-based Real-time Dense Face Reconstruction with Inverse-rendered Photo-realistic Face Images". *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2018.

[35] E. Richardson, M. Sela and R. Kimmel, "3D face reconstruction by learning from synthetic data", *in Proceedings of the 4th IEEE International Conference on 3D Vision (3DV)*, pp. 460-469, 2016.

[36] K.Wang, C. Gou, N. Zheng, J. M. Rehg and F.-Y. Wang, "Parallel vision for perception and understanding of complex scenes: methods, framework, and perspectives", *Artificial Intelligence Review*, vol. 48, no. 3, pp. 299-329, 2017.

[37] L.Li, Y. Lin, N. Zheng and F.-Y. Wang, "Parallel learning: a perspective and a framework", *IEEE/CAA Journal of Automatica Sinica* , vol. 4, no. 3, pp. 389-395, 2017.

[38] S. Agarwal and D. P. Mukherjee, "Synthesis of Realistic Facial Expressions using Expression Map," *IEEE Transactions on Multimedia*, 2018.

[39] W. Xie, L. Shen and J. Jiang, "A novel transient wrinkle detection algorithm and its application for expression synthesis", *IEEE Transactions on Multimedia*, vol. 19, no. 2, pp. 279-292, 2017.

[40] F. Lu, Y. Gao and X. Chen, "Estimating 3D gaze directions using unlabeled eye images via synthetic iris appearance fitting", *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1772-1782, 2016.

[41] H. Yu, O. Garrod, R. Jack and P. Schyns, "A framework for automatic and perceptually valid facial expression generation", *Multimedia Tools and Applications*, vol. 74, no. 21, pp. 9427-9447, 2015.

[42] X. Dong, J. Dong, G. Sun, Y. Duan, L. Qi and H. Yu, "Learning-based texture synthesis and automatic inpainting using support vector machines", *IEEE Transactions on Industrial Electronics*, vol. 66, no. 6, pp. 4777-4787, 2018.

[43] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez and C. Theobalt, "Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz", *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[44] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez and Christian Theobalt, "Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction", *In IEEE International Conference on Computer Vision (ICCV)*, 2017.

[45] L. Huynh, W.-K. Chen, S. Saito, J. Xing, K. Nagano, A. Jones, P. Debevec and H. Li, "Mesoscopic Facial Geometry Inference Using Deep Neural Networks", *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[46] S. Saito, L.-Y. Wei, L.-W. Hu, K Nagano and H Li, "Photorealistic facial texture inference using deep neural networks", *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[47] L.-W. Hu, S. Saito, L.-Y. Wei, K. Nagano, J. Seo, J. Fursund, I.Sadeghi, C. Sun, Y.-C. Chen and H. Li, "Avatar digitization from a single image for real-time rendering", *ACM Transactions on Graphics (ToG)*, vol. 36, no. 6, p.195, 2017.

[48] H. Yu, O. GB. Garrod and P. G. Schyns, "Perception-driven facial expression synthesis", *Computers & Graphics*, vol. 36, no. 3, pp. 152-162, 2012.

[49] P. Ekman, "Facial action coding system (FACS)", *A Human Face*, 2002.

[50] B. A. Erol, A. Majumdar, P. Benavidez, P. Rad, K-K. R. Choo and M. Jamshidi, "Toward Artificial Emotional Intelligence for Cooperative Social Human-Machine Interaction", *IEEE Transactions on Computational Social Systems*, 2019.

[51] M. François Valstar and M. Pantic, "Biologically vs. logic inspired encoding of facial actions and emotions in video", *In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 325-328, 2006.

[52] S. Velusamy, H. Kannan, B. Anand, A. Sharma and B. Navathe, "A method to infer emotions from facial action units", *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2028-2031, 2011.

[53] M. Ozuysal, P. Fua and V. Lepetit, "Fast keypoint recognition in ten lines of code", *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[54] P. Paysan, R. Knothe, B. Amberg, S. Romdhani and T. Vetter, "A 3D face model for pose and illumination invariant face recognition", *in Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS),* pp. 296-301, 2009.

[55] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces", *in IEEE International Conference on Computer Vision (ICCV)*, 2001.

[56] R. Weng, J. Lu, Y.-P. Tan and J. Zhou, "Learning cascaded deep auto-encoder networks for face alignment", *IEEE Transactions on Multimedia*, vol. 18, no. 10, pp. 2066-2078, 2016.

[57] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression", *in IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010.

[58] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild*", in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[59] X.-H. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment", *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[60] DI4D. 2018. http://www.di4d.com/.

[61] M. A. Oskoei and H.-S. Hu, "Myoelectric control systems—A survey", *Biomedical Signal Processing and Control*, vol. 2, no. 4, pp. 275-294, 2007.

[62] M. Hamedi, S. Hussain Salleh, C.-M.Ting, M. Astaraki and A. Noor, "Robust facial expression recognition for MuCI: a comprehensive neuromuscular signal analysis", *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 102-115, 2018.

[63] I. M. Rezazadeh, S. M. Firoozabadi, H.-S. Hu, and S. M. Reza Hashemi Golpayegani, "A novel human–machine interface based on recognition of multi-channel facial bioelectric signals", *Australasian Physical & Engineering Sciences in Medicine*, vol. 34, no. 4, pp. 497-513, 2011.

[64] M. Hamedi, S. H. Salleh and A. Noor, "Facial neuromuscular signal classification by means of least square support vector machine for MuCI", *Applied Soft Computing*, 30, pp. 83-93, 2015.

[65] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines", *ACM Transactions on Intelligent Systems and Technology,* vol. 2, no. 3, p. 27, 2011.

[66] S.-C. Du, Y Tao, and A. M. Martinez, "Compound facial expressions of emotion", *in Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. 1454-1462, 2014.

[67] M. Hamedi, S-H Salleh, T. T. Swee, "Surface electromyography-based facial expression recognition in Bi-polar configuration", *Journal of Computer Science*, vol. 7, no. 9, p. 1407, 2011.

[68] M. Halaki and K. Ginn, "Normalization of EMG signals: To normalize or not to normalize and what to normalize to?", *in Computational intelligence in electromyography analysis-a perspective on current applications and future challenges*, *IntechOpen*, 2012.

[69] H. Yu and H.-H. Liu, "Regression-Based Facial Expression Optimization", *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 3, pp. 386-394, 2014.

[70] Z.-B. Meng, S.-Z. Han and Y Tong, "Listen to Your Face: Inferring Facial Action Units from Audio Channel", *IEEE Transactions on Affective Computing*, 2017.

[71] L.-W. Hu, C.-Y. Ma, L.-J. Luo and H Li, "Single-view hair modeling using a hairstyle database", *ACM Transactions on Graphics (ToG*), vol. 34, no. 4, p.125, 2015.