Smile Action Unit detection from distal wearable Electromyography and Computer Vision

Monica Perusquía-Hernández*,1,2,4, Felix Dollack*,1,2, Chun Kwang Tan[2], Shushi Namba[3], Saho Ayabe-Kanamura[2], Kenji Suzuki[2]
[1] NTT Communication Science Laboratories, Japan
[2] University of Tsukuba, Japan
[3] RIKEN, Japan
[4] University of Essex, United Kingdom

***Abstract*— Distal facial Electromyography (EMG) can be used to detect smiles and frowns with reasonable accuracy. It capitalises on volume conduction to detect relevant muscle activity, even when the electrodes are not placed directly on the source muscle. The main advantage of this method is to prevent occlusion and obstruction of the facial expression production, whilst allowing EMG measurements. However, measuring EMG distally entails that the exact source of the facial movement is unknown. Therefore, we investigated whether we could identify specific Facial Action Units (AUs) from distal facial EMG after an initial calibration phase with Computer Vision (CV). We compared Support Vector Machines (SVM) and Random Forest (RF) with several types of feature engineering and early fusion of the two modalities. The detection performance for AU6 (Orbicularis Oculi) and AU12 (Zygomaticus Major) was estimated by calculating the agreement with Facial Action Coding System (FACS) certified coders. The best results were achieved using Random Forest. Using a fusion of CV and EMG features resulted in F1 scores of 0.83 for AU6; and the fusion of engineered EMG plus CV returned an F1 score of 0.81 for AU12. Both these results are well above the CV baseline that shows F1 scores of 0.56 and 0.62 for AU6 and AU12 respectively. This demonstrates the potential of distal EMG to detect individual facial movements. It also enables researchers to compare the results measured with this wearable device to psychological research on facial expressions using FACS. Using a wearable enables measurements with higher ecological validity. Finally, we observed that EMG activity starts before the onset of visually perceived movement. Because of this, the agreement between EMG-based methods and FACS coders might be underestimating the ground truth.**

I. INTRODUCTION

Smiles are a facial expression characterised by the corner of the lips moving upwards. This movement is generated by the *Zygomaticus Major* muscle (ZM). Smiles are the prototypical facial expression of happiness [9]. However, smiles can also be produced and perceived with other social communication aims [23]. The so-called Duchenne marker, or movement from the *Orbicularis Oculi* muscle (OO), often co-occurs with the *ZM* activity. Whilst it has been claimed that the Duchenne marker is a signal of smile spontaneity [9], [12], [8], other studies have found this marker in posed smiles as well [33].

* The authors contributed equally to this work.
This work was supported by NTT Communication Science Laboratories.

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

The Facial Action Coding System (FACS) [11] is a method to identify facial movements. Movements are described as the configuration of Action Units (AUs) in a standardised manner without judging the underlying emotion or the communicated message. These AU configurations can be used by experts to make inferences in the frame of different theories of emotion. In the FACS, the lip corner pulling upwards is labelled as Action Unit 12 (AU12), and the movement around the eyes in the form of a cheek raiser is labelled as AU6. AU6 is also the AU associated with the Duchenne Marker. FACS labelling usually requires a trained coder to watch and assess a video on a frame-by-frame basis. This is a time-consuming and cumbersome method, therefore, there have been several attempts to automatise AU detection.

Computer Vision (CV) algorithms [1] are an alternative to humans measuring AUs by visual inspection [10]. Additionally, the underlying muscle activity can be measured with Electromyography (EMG) [42], [34]. The standard method is to place the EMG electrodes directly on top of the relevant muscle to increase Signal-to-Noise Ratio (SNR). More recently, several studies have proven the feasibility of measuring facial expressions with distal EMG [16], [15]. Distal EMG refers to measuring muscle activity from a body location that is distant from the relevant muscle. Distal EMG measurements are possible through volume conduction whereby the electrical activity generated by each muscle spreads to adjacent areas [42]. By measuring EMG distally, the unnatural obstruction that the electrodes pose to the production of facial expressions is reduced. Despite this advantage, distal measurements make it difficult to know the exact location of the EMG activity source. Hence, current technology has been used only to identify grouped muscle activity such as smiles or frowns. Detecting such facial expressions from EMG has its own merit, such as high temporal resolution and robustness against occlusion. However, to compare the knowledge drawn using this technology to the large body of facial expression research that uses AUs as the basis of analysis, we need to identify muscle movement activity at the AU level.

We propose a **Sensing-Source framework** to analyse sensed signals by estimating their sources. Since AUs are closely related to individual muscle activity, we refer to them as "sources" (Figure 1). Sources are facial movement units

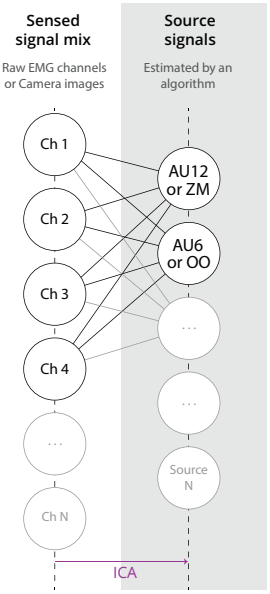

Fig. 1. Sensing-Source framework. Sensors often do not read the signals of interest but rather a mixture of those signals and other artefacts that can be considered noise. In some cases, the measured signals can be considered directly as the sources.

caused by a certain muscle. These individual sources often move together to form visible facial expressions such as smiles. We apply Independent Component Analysis (ICA) to the EMG signal to go from a sensed signal mix to the source signals underlying the observed movement. Next, we use an initial segment of the resulting ICA components to match components to AUs through cross-correlation with the continuous AU signals estimated from videos with OpenFace.

By combining CV- and EMG-based methods, it is possible to create an AU identification system that works in recording sessions where high movement or high facial occlusion are expected. In those cases, CV alone would struggle to continuously identify certain AUs. On the other hand, wearable distal EMG can deal with occlusion and movement, but it cannot disentangle AUs so easily. Therefore, we investigated the possibility to detect AUs with distal EMG, and compared the performance of single-modality models to multimodal models. We devised a method to pre-select the AU sources from EMG and compared these feature-engineered inputs to regular EMG pre-processing.

II. RELATED WORK

Accurate automatic AU detection has been one of the main challenges for affective computing scientists over the past decades. State-of-the-art methods use Computer Vision. It was only in the last decade that wearable EMG has started to be seen as a viable alternative. Here, we review the most relevant works from both perspectives. Nevertheless, it is important to note that most works use EMG to identify facial expressions as a whole, and, to the best of our knowledge, our work presents the first attempt to detect AUs using wearable distal EMG.

Fig. 2. Wearable used to measure distal EMG from four channels placed on the sides of the face, on both temples of the head. This configuration enables facial expression identification without obstructing the face. However, this makes identifying which muscle produced the measured activity challenging.

A. EMG-based identification

Compared to traditional EMG measurements, a reduced set of electrode positions has proven to yield high facial expression recognition rates of 87% accuracy for seven posed facial expressions, including sadness, anger, disgust, fear, happiness, surprise and neutral expressions. This subset includes electrodes placed on the *Corrugator* and *Frontalis* muscles on the forehead; and *Zygomaticus Major* (ZM) and *Masseter* muscles on the cheek [35]. Distal EMG has been used to identify different facial gestures by using different electrode configurations. Two EMG bipolar channels were placed on the *Temporalis* muscle on each side of the face, and one placed on the *Frontalis* muscle gave input to distinguish ten facial expressions. The achieved accuracy was 87% using a very fast versatile elliptic basis function neural network (VEBFNN) [16]. Although not all gestures were facial expressions of emotion, they did include symmetrical and asymmetrical smiling, raising eyebrows, and frowning.

Distal EMG has been implemented as a wearable designed to keep four EMG channels attached to the sides of the face at eye level (Figure 2). With this placement, it is possible to reliably measure smiles in different situations without obstructing facial movement [30], [14]. This is possible because smile-related distal activity measured from the ZM is sufficiently large to be robust against non-affective facial movements such as chewing gum and biting [27], [42], [15]. Hence, the information picked up by the four channels is used to approximate different sources of muscular activity using ICA [7]. The separated components contain muscle activity involved in generating smiles and can be used to identify these [15]. This approach can be used offline for fast and subtle spontaneous smile identification [30] and it is possible even in real time [40]. Finally, this device has also been used to analyse spatio-temporal features of a smile by fitting envelopes to the EMG's Independent Components (ICs), and later performing automatic peak detection on those envelopes [31] with performance similar to that achieved by Computer Vision [29]. Furthermore, four EMG leads placed around the eyes in a Head-Mounted Display (HMD) have been used successfully to identify facial expressions distally even when the face is covered by the HMD. Facial expressions of anger, happiness, fear, sadness, surprise, neutral,

clenching, kissing, asymmetric smiles, and frowning were identified with 85% of accuracy [3]. Another recent work proposed the use of a thin sticker-like hemifacial 16 electrode array to paste on one side of the face and identify ten distinct Facial Building Blocks (FBB) of different voluntary smiles. Their electrode approach is novel, robust against occlusion, and provides a higher density electrode array than that of the aforementioned arrangements. This enabled them to use ICA and clustering to define several FBB corresponding to a certain muscle [18]. Nevertheless, they require electrode usage proximal to each muscle. This entails that a large sticker needs to be placed on the skin, obstructing spontaneous facial movement through increased stiffness. Moreover, the physical connection of the electrode array enhances artefact cross-talk between electrodes. To eliminate such cross-talk, ICA was used and the resulting clusters were derived manually. Finally, two around-the-ear electrode arrays with 18 channels have been used to successfully identify reading, speaking, chewing, jaw clenching, and six posed emotion expressions (i.e., happy, angry, disgusted, fear, sadness, surprise) with a Random Forest classifier. These electrodes are able to measure both Electroencephalography (EEG) and distal EMG. Smiles were identified with an F1 score of 0.83 [20].

B. CV-based identification

Computer Vision (CV) is the most widely used technique for identifying facial expressions [2], even at the individual AU level. The ubiquitous presence of cameras and its ease of use make it the method of choice for scenarios where the face is still and unobstructed. There are different approaches to extract relevant features for AU identification and intensity estimation. Among these, appearance-based, geometry-based, motion-based, and hybrid approaches. Several algorithms show F1 scores in the range between 0.45 and 0.57 for occurrence detection and between 0.21 and 0.41 for intensity estimation [24]. The OpenFace toolkit 2.0 [1] is a CV pipeline for facial and head behaviour identification. Its behaviour analysis pipeline includes landmark detection, head pose and eye gaze estimation, and facial action unit recognition. This algorithm detects AU 1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25, 26 with an average accuracy of 59 %. According to a benchmark conducted by the authors of a Python facial expression analysis toolbox PyFeat [4], OpenFace is the best performing algorithm for AU6 and AU12 with F1 scores of 0.81 and 0.83, respectively. Moreover, a Twin Cycle Auto Encoder can be used to extract representations for AUs in a self-supervised manner. The results show F1 scores for 3 datasets between 0.48 and 0.75 for AU6 and 0.76 and 0.85 for AU12 [22]. Model-agnostic meta-learning in combination with few-shot learning achieved accuracies of 0.87 for AU6 and 0.83 for AU12 on the DISFA and 0.81 for AU6 and 0.86 for AU12 on the BP4D dataset [21]. JAA-Net (Joint Facial Action Unit Detection and Face Alignment via Adaptive Attention) is an end-to-end deep learning framework with an attention learning module that proposes a novel AU detection method that combines face alignment with facial landmarks and local AU detection to improve AU detection performance. This method yields F1-scores of 0.78 and 0.87 for AU6 and AU12 respectively [36], [38]. Convolutional Neural Network evaluations on multiple views of the face have shown F1-scores of 0.77 for AU6 and 0.88 for AU12 [32] on the BP4D database. Considering multi-view algorithms is an important step to use CV in the wild. The recent Matlab framework Automated Facial Affect Recognition framework (AFAR) is an attempt to bridge the gap between expensive commercial tools of unknown validity and hard-to-use open-source tools. It automates AU detection by relying on pretrained models, and provides the option for users to fine-tune the network performance with their own datasets [13].

III. DATA SET

This data is a subset of the data generated in a previous study exploring posed and spontaneous smiles [28]. Since we are only interested in detecting AUs, regardless of the nature of the smiles, we collapsed the data from different experiment conditions into one.

A. Participants

41 producers took part in the study (19 female, average age $= 25.03 \pm 3.83$ years). All the participants had normal or corrected-to-normal vision. This research was approved by the Institutional Ethical Committee of the University of Tsukuba with review code 2017R176. From these, 10 participants were removed as they did not show enough AU6 samples to allow for cross-validation.

B. Experiment design

The experiment consisted of several blocks. All the producers completed all the experimental blocks in the same order. This was to keep the purpose of the experiment hidden during the spontaneous block.

1) Spontaneous Block (S-B): A positive affective state was induced using a 90 s humorous video. After the stimuli, a standardised scale assessing emotional experience was answered. Next, producers were asked to tag any facial expressions that they had made.

2) Posed Block (P-B): Producers were requested to make similar smiles as they did in the S-B. However, this time, a 90 s slightly negative video was presented. Their instruction was: “Please perform the smiles you video coded. This is for a contest. We are going to show the video we record to another person, who is unknown to you, and if she or he cannot guess what video you were watching, then you are a good actor. Please do your best to beat the evaluator”. After watching the video and performing the task, they completed the same standardised scale assessing emotional experience. They were also asked to tag their own expressions.

C. Measurements

Smile-reader: Four channels total of distal facial EMG were measured from both sides of the face using dry-active electrodes (Biolog DL4000, S&ME Inc) sampled at 1 kHz (Figure 2).

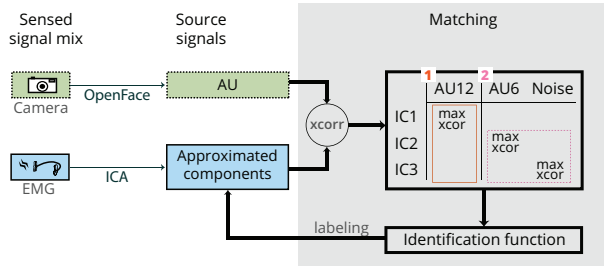

Fig. 3. Feature engineering. AU labels are independently extracted from CV. Then the information derived from CV is used to identify which ICA components of the EMG are likely to correspond to each AU type. The matching is done by cross-correlating the three independent components (ICs) derived from EMG to CV_AU12, CV_AU6 and to Gaussian noise representing a mix of other facial movements. Finally, a threshold of the mean plus 20 standard deviations are used to determine the AU presence in the identified ICA components, independently of the CV labels.

Video recordings: A video of the producer's facial expressions was recorded using a Canon Ivis 52 camera at 30 FPS.

Certified FACS labels: Two certified FACS coders labelled the intensity of AU6 and AU12 on a frame-by-frame basis. The coding of both FACS coders was combined with an OR logic operator. Coded intensities were averaged. To score AU presence, a minimum intensity value of 1 (level A) was used.

IV. DATA ANALYSIS

A. EMG Pre-processing

The four EMG channels were first passed through a custom Hanning window with a ramp time of 0.5 s to avoid the introduction of artificial frequencies by the filtering at the start and the end of the signal. Afterwards, the signals were (1) linear detrended, (2) transformed to have zero mean and one standard deviation, (3) band-pass filtered from 15 to 490 Hz and (4) rectified. These bandpass frequencies were previously reported to be optimal for detecting facial muscle movement. Frequencies below 15 to 25 Hz and above 400 to 500 Hz contain undesired artefacts [41].

B. CV-based AU labelling

Using AFAR: AFAR is a toolbox that provides an automated AU detection pipeline, which consists of face tracking, face registration, AU detection and visualization [13]. The toolbox uses a deep neural network to perform AU detection using pre-trained models [5]. The model returns probabilities for the presence of AUs in a continuous manner. We only used the outputs for AU6 and AU12.

Using JAA-Net: JAA-Net is a deep learning based AU detection framework that exploits the common features between AU detection and face alignment tasks to improve the robustness of AU detection [37], [38], [39]. The model output was in the same format as from the AFAR toolbox and only AU6 and AU12 were used.

Using OpenFace: The Facial Behaviour Analysis Toolkit OpenFace 2.0 was used to identify several facial features including AUs. This is an end-to-end deep learning framework.

TABLE I
AUC METRIC OF OPENFACE, AFAR AND JAA-NET COMPARED TO FACS CODING FOR AU6, AU12 OF OUR DATASET

	AU6	AU12
OpenFace	0.81	0.77
AFAR	0.71	0.77
JAA-Net	0.58	0.50

AU identification is given both as a continuous output of intensity ratings; and a binary output indicating AU presence. The intensity and presence predictors have been trained separately and on slightly different datasets, which means that they are not always consistent [1]. In this work, we choose to use the continuous output of the algorithm, as it allows us to correlate the outputs with the EMG, and it is comparable to the continuous outputs of AFAR and JAA-Net.

C. Selection of the CV reference for the matching algorithm

To select a CV-based reference model or baseline model, we selected and evaluated three different CV algorithms using the Area Under the (Receiving Operating) Curve (AUC) on our dataset. OpenFace, AFAR, and JAA-Net were selected as candidate baseline models. We used AUC because previous research showed that all performance metrics except the AUC were attenuated by skewed data distributions [19], [26]. These comparisons considered both threshold metrics (i.e., accuracy, F-score, Cohen's kappa, Krippendorf's alpha), and rank metrics (i.e., AUC, the precision-recall curve). Therefore, we choose the area under the receiver operating curve rank metric as evaluation criteria for the selection of our baseline model. OpenFace was finally chosen as the baseline model because it outperforms AFAR and JAA-Net, based on the AUC metric. Table I lists the AUC score for OpenFace, AFAR and JAA-Net.

D. Blind-source separation

Independent component analysis (ICA) [17] was used to automatically estimate different muscle activity sources from the recorded EMG signals. The wearable used to collect the data has four channels. Thus, we set the number of decomposed components to three.

E. Feature engineering with component matching to CV-generated labels

We propose a **Component Matching (CM)** method (*engEMG* in Figure 6 and Table II). This method aims to identify different sources or muscle groups from the recorded distal EMG, and to assesses their similitude to AU labels estimated with CV (Figure 3). First, blind-source separation is used to estimate sources of facial movement; then cross-correlation is applied to match CV-based AU output with the estimated sources from EMG; next, ICs are tagged as AU6 and AU12. The AU presence can be then detected from the tagged components using machine learning. Our matching algorithm assumes that the ICA components of the EMG signal contain AU6, AU12 and noise. Noise is defined as electrical interference as well as other muscular source's

activity (e.g., other AUs, chewing, jaw clenching). This assumption is made because the participants were mostly smiling and keeping a neutral expression otherwise. At times, other AUs were displayed, but we are not interested in them at this point. We calculated the cross-correlation of the three ICA components; the continuous AU6, AU12 OpenFace CV-labels; and a uniformly distributed random noise distribution. Since AU12 stems from the large and strong ZM muscle, the independent component (IC) with the highest correlation is chosen to correspond to AU12. The other two ICs get assigned to be AU6 and noise in order of maximum correlation value. Afterwards, the ICs are downsampled from 1 kHz to 30 Hz to match the sampling frequency of the FACS coding. Further smoothing is applied on the individual ICs by means of a first-order Savitzky-Golay filter with a length of 1 second. Figure 4 shows EMG components thresholded based on standard deviation (SD) to determine the presence of the relevant AU. For thresholding, the latter half of a $\approx 10\ s$ or 300 samples long neutral phase of the IC was used as a baseline. We calculate the signal average $\overline{m}$ and standard deviation σ from the baseline. The whole signal then is turned into a binary vector where samples that cross the threshold of $\overline{m} + k\sigma$ with $k = 20$ are set to one (Figure 4). The delay between EMG activity and visible AU label was also quantified using cross-correlation between the selected AU EMG components (before binarization) and the AU continuous label output from OpenFace (Figure 5). The values set to one are thought to correspond to the activity of the respective AU assigned to the IC during the process of AU identification. In other words, our matching algorithm is using an analytical technique applied on a continuous-time series to transform the data into more relevant information that can be thresholded by its SD or other machine learning techniques.

F. Comparison to other Machine Learning algorithms

The baseline model was OpenFace. We fitted Support Vector Machines (SVM) and Random Forests (RF) using sklearn in Python. A test set with 20% data was set aside to score the models. Subject-dependent models were fitted using a 5-fold cross-validation and later scored using the test set. AU6 and AU12 were fitted using separate binary classifiers. The SVMs were fitted using a radial basis function (RBF) kernel, a maximum of 1000 iterations and with a tolerance of 0.0001. RF had a maximum depth of 2 and was initialised with the same random state. Input features to the models were mean, standard deviation and kurtosis, extracted with a sliding window of length $0.25\ s$ of (a) each of the EMG channels (EMG); and (b) the two ICA channels identified as previously described as AU6 and AU12 (engEMG). Furthermore, feature fusion was performed on the combinations of (c) the EMG features and CV (CV+EMG); as well as (d) the ICA features and CV (CV+engEMG). CV features are the AU intensity values extracted using OpenFace.

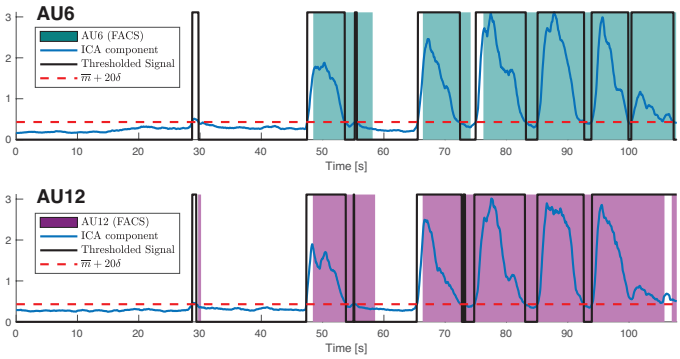

Fig. 4. AU detection by the human-coded FACS labels (shaded areas), and the selected components by the engEMG algorithm (blue line). By using a threshold of the signal's mean plus 20 SD (red dotted line), we can identify parts of the AU signal that are relevant including the ones that were not visible from the camera. These are areas that are not necessarily a misclassification. However, they do not match the visual ground truth, and therefore, EMG-based results would be penalised.

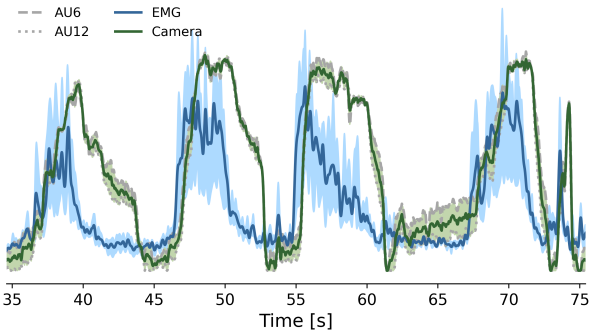

Fig. 5. The plot shows exemplary data from one participant posing smiles. The activation patterns of EMG and CV-based AUs are similar to each other, with the muscle activation measured with distal EMG leading camera-detected AU movement. The four channels of raw distal EMG activate on average 374 ms before the detected CV-based AU labels. The blue line shows the mean of the four EMG channels plus Standard Deviation (SD). AU6 and AU12 (grey lines) often co-occur, as shown by the green line representing the average CV-based output. This makes identifying which muscle produced the measured activity challenging, as EMG measures a mix of muscle activity throughout the face.

G. Agreement with the ground truth

Human-coded FACS labels, CV-labels, and EMG-labels were transformed to have a matching sampling rate (1 kHz). Then the agreement between different measurements was calculated using accuracy and the F1 score that encompasses both precision and recall. Additionally, we calculated Area Under the Curve (AUC) as a metric that penalises for class imbalances.

H. Comparison between models

Comparisons were performed with the Kruskal-Wallis rank sum test. AUC was used given the unbalanced nature of the data, i.e., there is an unequal sample of no expression, AU6 and AU12. For reference, comparisons using F1 are also provided.

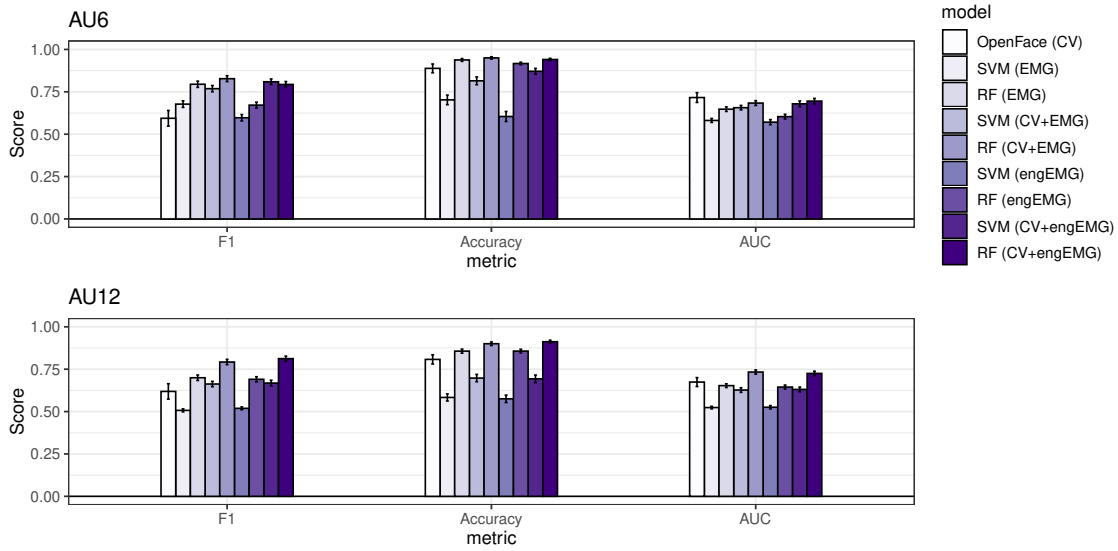

Fig. 6. The performance of the different methods as measured by F1 score, accuracy and Area Under the Curve (AUC). Error bars represent standard errors. The baseline's (CV) performance is quite good for AU12, but it can be improved by including EMG information. AU6's detection increases greatly by considering the input from EMG. The engEMG corresponds to a Component Matching feature engineered EMG as input.

TABLE II
MEAN AGREEMENT OF AU6 AND AU12 BETWEEN HUMAN FACS-CODED AND MACHINE-DETECTED AUS.

		CV	EMG SVM	EMG RF	CV+EMG SVM	CV+EMG RF	engEMG SVM	engEMG RF	CV+engEMG SVM	CV+engEMG RF
AU6	F1	0.59	0.68	0.79	0.77	0.83	0.60	0.67	0.81	0.79
	Accuracy	0.89	0.70	0.94	0.81	0.95	0.60	0.92	0.87	0.94
	Recall	0.39	0.68	0.79	0.77	0.82	0.60	0.66	0.81	0.78
	Precision	0.33	0.68	0.81	0.77	0.85	0.60	0.70	0.81	0.83
	AUC	0.72	0.58	0.65	0.66	0.68	0.57	0.60	0.68	0.70
AU12	F1	0.62	0.51	0.70	0.66	0.79	0.52	0.69	0.67	0.81
	Accuracy	0.81	0.58	0.86	0.70	0.90	0.58	0.86	0.69	0.91
	Recall	0.35	0.52	0.69	0.66	0.78	0.52	0.69	0.66	0.79
	Precision	0.63	0.51	0.71	0.68	0.82	0.52	0.71	0.68	0.84
	AUC	0.67	0.52	0.65	0.63	0.73	0.53	0.64	0.63	0.73

V. RESULTS AND DISCUSSION

Using our feature engineering method and a simple threshold method, we observed that EMG activity precedes visual movement onset (Figure 4). We compared EMG activity to the OpenFace output (Figure 5). There was a delay between CV AU activation and EMG activation, with EMG activation leading by 374 ms. This was expected as EMG originates skin displacement. This delay was larger than that observed from proximal EMG measurements (average of 230 ms) [6].

We proposed to identify AU6 or the Duchenne Marker, and AU12 or the movement of the mouth during smiling with a distal EMG wearable device. Our results show that AU6 and AU12 can be identified using distal EMG. Furthermore, we compared the performance of several algorithms. The baseline was the OpenFace CV algorithm. The baseline did not perform as well in our dataset as in previous reports [4]. As shown in Table II, AU6 had an F1-score of 0.59 and AU12 an F1-score of 0.62 in our dataset as compared to 0.81 and 0.83 reported for the Extended DISFA dataset [25]. Using EMG only and RF had an advantage (F1-score 0.79 for AU6, and 0.70 for AU12). The performance scored with F1 was also increased by fusing features from CV and EMG (0.83 for AU6, and 0.79 for AU12), and by fusing CV and engineered EMG features (0.79 for AU6, and 0.81 for AU12).

Using F1 as the success metric, we observed generalised differences across classifiers (AU6: $\chi^2(9) = 148.02, p < .001$; AU12: $\chi^2(9) = 284.05, p < .001$). RF is better than SVM at identifying AU12 ($\chi^2(1) = 106.64, p < .001$). This also holds marginally for AU6 ($\chi^2(1) = 4.67, p = .03$). This might indicate that RF is better at detecting Action Unit elements from facial expressions than SVM. Overall, the fusion of two modalities is better than EMG or CV alone, for both AU6 ($\chi^2(1) = 29.82, p < .001$) and AU12 ($\chi^2(1) = 87.55, p < .001$). The best performing combination was RF and fusion of CV and EMG features with F1 of 0.83 for AU6; and RF and fusion of CV and engineered EMG features F1

of 0.81 for AU12. These values were significantly above the CV-only baseline for AU6 only (AU6: $\chi^2(1) = 6.57$, $p < .05$; AU12: $\chi^2(1) = 0.12$, $p = .73$). Furthermore, the feature engineering method we devised was helpful for both AU6 ($\chi^2(1) = 7.29$, $p < .01$), and AU12 ($\chi^2(1) = 31.48$, $p < .001$).

Using AUC as the success metric, we observed generalised differences across classifiers as well (AU6: $\chi^2(9) = 82.02$, $p < .001$; AU12: $\chi^2(9) = 262.91$, $p < .001$). RF is again better than SVM at identifying AU6 ($\chi^2(1) = 7.85$, $p < .01$) and AU12 ($\chi^2(1) = 110.03$, $p < .001$). Overall, the fusion of two modalities is better than EMG or CV alone, for both AU6 ($\chi^2(1) = 31.80$, $p < .001$) and AU12 ($\chi^2(1) = 63.84$, $p < .001$). The best performing model for AU6 only was the CV baseline, OpenFace with AUC of 0.72 followed by a RF trained on CV and EMG features processed with our method (0.68). However, for AU12, RF and fusion of CV and EMG features achieved the highest AUC score (0.733), followed closely by RF with CV and engineered EMG (0.725). We did not find a significant difference with the CV-only baseline (AU6: $\chi^2(1) = 3.79$, $p = .05$; AU12: $\chi^2(1) = 2.21$, $p = .13$). Furthermore, the feature engineering method we devised was helpful only for AU12 (AU6: $\chi^2(1) = 0.82$, $p = .36$, AU12: $\chi^2(1) = 17.86$, $p < .001$).

The F1 scores from OpenFace are reported on their original benchmark on a specific dataset. It is common that benchmark results only hold for the dataset they were obtained on. Many of the available datasets are recorded in good conditions optimized for computer vision and different classes are balanced. In this dataset, participants' movement was not constrained, and they were not instructed to behave in a certain way beneficial for the face detection algorithm. This was a choice to increase ecological validity, even if it entails that the detection becomes harder. Therefore, a drop in performance is to be expected. In the dataset we used in this paper, there is head rotation and sometimes occlusion by participants covering their mouth. Furthermore, when there is data imbalance such as in the case of AU presence assessment in a video, metrics such as accuracy and F1 score have been reported to be biased. A classifier could achieve high performance just by saying that there is no AU all the time, because indeed, AUs might be rare in between many frames of neutral expression [19].

Previous work using wearable distal EMG aimed to detect smiles. Smiles are often a combination of AU6 and AU12. Therefore, detecting smiles is an easier task than detecting the more fine-grained action units. In particular, because the Orbicularis Oculi and the Zygomaticus Major often move in synchrony, and we are measuring their activity distally. Figure 4 shows little difference between the selected components for AU12 and AU6. Future work should explore to what extent this synchrony is expected due to the shape of the smile, or if it is a measurement limitation. Moreover, previous studies with this device used a different experimental paradigm to elicit smiles, and it did not make any performance comparison to a computer vision baseline [30].

Overall, our method presents little improvement with respect to the state-of-the-art in computer vision. This might have been because when performing ICA, we decomposed the data into three time series, from which only the two most relevant ones were used as input for the algorithm. When our matching was not successful, important information might have been lost. Our feature engineering method helped the detection only marginally. We expect that as the number of EMG channels increases, this method would be useful for dimensionality reduction. Future work should aim to improve algorithm performance in such scenarios.

Furthermore, we can observe in Figure 6 that the EMG activity starts before it can be observed by the FACS coder. This might be one of the reasons why results of the EMG and engineered EMG methods are sometimes in less agreement with the ground truth FACS labels determined by visual cues. However, it does not necessarily mean that it is less valid. Future work should explore these differences between visible and invisible activity in more detail.

One of the limitations of this study was the number of electrodes provided in the EMG wearable. Four electrodes provide a good trade-off between wearability, smile and AU detection; but they are limited to estimating multiple muscle sources. Increasing the electrode number will enable us to explore more AUs. In this case, we opted to model mainly AU6 and AU12, and to consider other AUs in the EMG as "noise".

Our results suggest that in situations when the participants are seated in front of the camera, there is no difference between CV and wearable EMG. Therefore, CV might be the method of choice, given its unobtrusiveness. Nevertheless, we argue that there are some situations where using wearable distal EMG would be beneficial, and would potentially outperform CV-based methods. These are situations where the people to be tracked are constantly moving, lighting conditions vary, and there is constant occlusion. An example of this is the case of children playing outdoor games. Further, we believe that by combining both methods during calibration, we can aim to achieve a system that needs little to no input from a human coder, and still achieves an acceptable level of accuracy. Future work should compare the performance of both modalities in high movement scenarios.

VI. Conclusions

We demonstrated the use of distal EMG to detect individual facial movements during a smile. By detecting AUs instead of facial expressions, we can explore facial movements before making inferences of their affective meaning. Uncovering those movements with a high temporal resolution will help shed light on the intended and perceived affective meaning. Therefore, this technology would enable researchers to investigate facial social signal behaviour in a more ecologically valid manner, and to compare the results measured with this device to the majority of psychological research on facial expressions. So far psychological research on this aspect has been restricted to highly controlled environments where all the dynamism of facial expressions might have been altered by demand characteristics.

REFERENCES

[1] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. OpenFace 2.0: facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, may 2018.

[2] V. Bettadapura. Face expression recognition and analysis: the state of the art. *CoRR*, pages 1–27, 2012.

[3] H. S. Cha, S. J. Choi, and C. H. Im. Real-time recognition of facial expressions using facial electromyograms recorded around the eyes for social virtual reality applications. *IEEE Access*, 8:62065–62075, 2020.

[4] J. H. Cheong, T. Xie, S. Byrne, and L. J. Chang. Py-Feat: Python Facial Expression Analysis Toolbox. page 25. ZSCC: 0000000.

[5] J. F. Cohn, L. A. Jeni, I. Onal Ertugrul, D. Malone, M. S. Okun, D. Borton, and W. K. Goodman. Automated affect detection in deep brain stimulation for obsessive-compulsive disorder: A pilot study. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 40–44, New York, NY, USA, 2018. Association for Computing Machinery.

[6] J. F. Cohn and K. Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2:121–132, 2004.

[7] P. Comon. Independent component analysis, A new concept? *Signal Processing*, 36(36):28–314, 1994.

[8] P. Ekman. Basic Emotions. In T. Dalgleish and M. Power, editors, *Handbook of cognition and emotion*, chapter 3, pages 45–60. John Wiley & Sons, Ltd., 1999.

[9] P. Ekman, W. Friese, and R. Davidson. The Duchenne Smile: Emotional Expression And Brain Physiology II. *Journal of Personality and Social Psychology*, 58(2):342–353, 1988.

[10] P. Ekman, W. Friesen, and J. Hager. FACS investigator's guide, 2002.

[11] P. Ekman and W. P. Friesen. Measuring facial movement with the Facial Action Coding System. In P. Ekman, editor, *Emotion in the human face*, chapter 9, pages 178–211. Cambridge University Press, second edition, 1982.

[12] P. Ekman, W. V. Friesen, and M. O'Sullivan. Smiles when lying. *Journal of Personality and Social Psychology*, 54(3):414–420, 1988.

[13] I. O. Ertugrul, L. A. Jeni, W. Ding, and J. F. Cohn. AFAR: A Deep Learning Based Tool for Automated Facial Affect Recognition. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–1, May 2019. ZSCC: 0000007.

[14] A. Funahashi, A. Gruebler, T. Aoki, H. Kadone, and K. Suzuki. Brief report: The smiles of a child with autism spectrum disorder during an animal-assisted activity may facilitate social positive behaviors - Quantitative analysis with smile-detecting interface. *Journal of Autism and Developmental Disorders*, 44(3):685–693, 2014.

[15] A. Gruebler and K. Suzuki. Design of a Wearable Device for Reading Positive Expressions from Facial EMG Signals. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2014.

[16] M. Hamedi, S.-H. Salleh, M. Astaraki, and A. M. Noor. EMG-based facial gesture recognition through versatile elliptic basis function neural network. *Biomedical engineering online*, 12(1):73, 2013.

[17] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks: the official journal of the International Neural Network Society*, 13(4-5):411–30, 2000.

[18] L. Inzelberg, M. David-Pur, E. Gur, and Y. Hanein. Multi-channel electromyography-based mapping of spontaneous smiles - IOPscience. *Journal of Neural Engineering*, 17(2), apr 2020.

[19] L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 245–251, 2013.

[20] M. Knierim, M. Schemmer, and M. Perusquía-Hernández. Exploring the recognition of facial activities through around-the-ear electrode arrays (ceegrids). pages 57–65, 2021.

[21] M. Lee, O. Rudovic, V. Pavlovic, and M. Pantic. Fast Adaptation of Deep Models for Facial Action Unit Detection Using Model-Agnostic Meta-Learning. In *Workshop on Artificial Intelligence in Affective Computing*, pages 9–27. PMLR, Nov. 2020. ZSCC: 0000000 ISSN: 2640-3498.

[22] Y. Li, J. Zeng, S. Shan, and X. Chen. Self-Supervised Representation Learning From Videos for Facial Action Unit Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10916–10925, June 2019. ZSCC: 0000024 ISSN: 2575-7075.

[23] J. D. Martin, A. Wood, W. T. L. Cox, S. Sievert, R. Nowak, E. Gilboa-Schechtman, F. Zhao, Z. Witkower, A. T. Langbehn, and P. M. Niedenthal. Evidence for Distinct Facial Signals of Reward, Affiliation, and Dominance from Both Perception and Production Tasks. *Affective Science*, Feb. 2021.

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

[25] M. Mavadati, P. Sanger, and M. H. Mahoor. Extended DISFA Dataset: Investigating Posed and Spontaneous Facial Expressions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1452–1459, Las Vegas, NV, USA, June 2016. IEEE. ZSCC: 0000045.

[26] S. Namba, W. Sato, M. Osumi, and K. Shimokawa. Assessing automated facial action unit detection systems for analyzing cross-domain facial expression databases. *Sensors*, 21(12), 2021.

[27] L. Oberman, P. Winkielman, and V. Ramachandran. Face to face: blocking facial mimicry can selectively impair recognition of emotional expressions. *Social neuroscience*, 2(3-4):167–78, sep 2007.

[28] M. Perusquía-Hernández, S. Ayabe-Kanamura, and K. Suzuki. Human perception and biosignal-based identification of posed and spontaneous smiles. *PLOS ONE*, 14(12):e0226328, dec 2019.

[29] M. Perusquía-Hernández, S. Ayabe-Kanamura, K. Suzuki, and S. Kumano. The Invisible Potential of Facial Electromyography: A Comparison of EMG and Computer Vision when Distinguishing Posed from Spontaneous Smiles. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–9, New York, New York, USA, 2019. ACM Press.

[30] M. Perusquía-Hernández, M. Hirokawa, and K. Suzuki. A wearable device for fast and subtle spontaneous smile recognition. *IEEE Transactions on Affective Computing*, 8(4):522–533, 2017.

[31] M. Perusquía-Hernández, M. Hirokawa, and K. Suzuki. Spontaneous and posed smile recognition based on spatial and temporal patterns of facial EMG. In *Affective Computing and Intelligent Interaction*, pages 537–541, 2017.

[32] A. Romero, J. Leon, and P. Arbelaez. Multi-View Dynamic Facial Action Unit Detection. *arXiv:1704.07863 [cs]*, Aug. 2018. ZSCC: 0000011 arXiv: 1704.07863 version: 2.

[33] K. Schmidt, S. Bhattacharya, and R. Denlinger. Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises. *Nonverbal Behaviour*, 33(1):35–45, 2009.

[34] K. L. Schmidt and J. F. Cohn. Dynamics of facial expression: Normative characteristics and individual differences. In *IEEE Proceedings of International Conference on Multimedia and Expo.*, pages 728–731, Tokyo, 2001. IEEE.

[35] I. Schultz and M. Pruzinec. *Facial Expression Recognition using Surface Electromyography*. PhD thesis, Karlruhe Institute of Technology, 2010.

[36] Z. Shao, Z. Liu, J. Cai, and L. Ma. Deep Adaptive Attention for Joint Facial Action Unit Detection and Face Alignment. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, volume 11217, pages 725–740. Springer International Publishing, Cham, 2018. ZSCC: NoCitationData[s0] Series Title: Lecture Notes in Computer Science.

[37] Z. Shao, Z. Liu, J. Cai, and L. Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *European Conference on Computer Vision*, pages 725–740. Springer, 2018.

[38] Z. Shao, Z. Liu, J. Cai, and L. Ma. Jâa-net: Joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 2020.

[39] Z. Shao, Z. Liu, J. Cai, and L. Ma. JÂA-Net: Joint Facial Action Unit Detection and Face Alignment Via Adaptive Attention. *International Journal of Computer Vision*, 129(2):321–340, Feb. 2021.

[40] Y. Takano and K. Suzuki. Affective communication aid using wearable devices based on biosignals. In *Proceedings of the 2014 conference on Interaction design and children - IDC '14*, pages 213–216, New York, New York, USA, 2014. ACM Press.

[41] A. van Boxtel. Optimal signal bandwidth for the recording of surface emg activity of facial, jaw, oral, and neck muscles. *Psychophysiology*, 38(1):22–34, 2001.

[42] A. van Boxtel. Facial EMG as a Tool for Inferring Affective States. In A. Spink, F. Grieco, Krips OE, L. Loijens, L. Noldus, and P. Zimmerman, editors, *Proceedings of Measuring Behavior*, pages 104–108, Eindhoven, 2010.