Automatic Smile and Frown Recognition with Kinetic Earables

Seungchul Lee*
KAIST, Republic of Korea
seungchul@nclab.kaist.ac.kr

Chulhong Min
Nokia Bell Labs, UK
chulhong.min@nokia-bell-labs.com

Alessandro Montanari
Nokia Bell Labs, UK
alessandro.montanari@nokia-bell-labs.com

Akhil Mathur
Nokia Bell Labs and UCL, UK
akhil.mathur@nokia-bell-labs.com

Youngjae Chang*
KAIST, Repulic of Korea
yjchang@nclab.kaist.ac.kr

Junehwa Song
KAIST, Repulic of Korea
junesong@nclab.kaist.ac.kr

Fahim Kawsar
Nokia Bell Labs, UK and TU Delft
fahim.kawsar@nokia-bell-labs.com

ABSTRACT

In this paper, we introduce inertial signals obtained from an earable placed in the ear canal as a new compelling sensing modality for recognising two key facial expressions: smile and frown. Borrowing principles from Facial Action Coding Systems, we first demonstrate that an inertial measurement unit of an earable can capture facial muscle deformation activated by a set of temporal micro-expressions. Building on these observations, we then present three different learning schemes - shallow models with statistical features, hidden Markov model, and deep neural networks to automatically recognise smile and frown expressions from inertial signals. The experimental results show that in controlled non-conversational settings, we can identify smile and frown with high accuracy (F_1 score: 0.85).

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

KEYWORDS

smile and frown recognition; earable; FACS; kinetic modeling.

1 INTRODUCTION

Facial expressions are one of the most powerful and essential non-verbal signals for social interactions, conveying cues about human affect, empathy, and intention. Naturally, building systems capable of automatically recognising facial expressions has been an intense interdisciplinary area of study with many and diverse applications including HCI, affective communication, behavioural science, medicine, entertainment, education, and security.

Vision coupled with powerful deep learning algorithms has been the conventional approach for building automatic recognition systems of facial expression [7, 11]. Although highly accurate, vision-based sensing requires a camera directed towards a user's face, it often has little tolerance to occlusion, camera angle, and changing lighting conditions. Moreover, it lacks a wearable form due to ergonomics and privacy. Electromyographic (EMG) signals obtained through electrodes placed on the facial surface capturing electrical activities of the muscle tissues have been studied [4]. Beyond clinical settings, this approach is impractical, unnatural, uncomfortable, and undesirable, despite demonstrating strong recognition performance. In the wearable computing literature, multiple works have explored smart glasses augmented with piezoelectric sensors [10] and photo-reflective sensors [6]. While these approaches showed promising results, further studies are needed to fully assess their capabilities in diverse real-world settings for continuous monitoring of facial expressions describing human emotional states.

We turn our attention onto kinetics for recognising facial expressions. In particular, for the first time, we explore the capability of a wireless earable placed in the ear canal and augmented with an inertial measurement unit (accelerometer and gyroscope) in understanding facial muscle movements. Earables are discreet, privacy preserving, and already integrated into our lives; they allow people to make hands-free calls and provide access to high-definition music during work, commuting, and exercise. If earables can recognise facial expressions continuously and unobtrusively, it will uncover a wide area of personal-scale applications in the domains of affective communication, education, health and well-being.

To this end, we first look at Facial Action Coding Systems and specifically its constituent Action Units that descriptively encode various facial expressions to establish whether temporal muscle deformations triggered by micro-expressions can be captured by our sensory earable [5]. We observe that *zygomaticus major* (activated while smiling) and *corrugator supercilii* (activated while frowning) muscles' movements are strongly visible in inertial signals on earables. Building on this observation, we develop three different learning schemes - shallow models with statistical features, hidden Markov model (HMM), and deep neural networks (DNN) to automatically recognise these two expressions — *smile* and *frown*. To benchmark those, we construct a brand-new dataset of 1,620 samples up to 3 seconds-long cross-talk free (i.e., no other facial

*This work was done when these authors were on an internship at Nokia Bell Labs.

AH2019, March 11–12, 2019, Reims, France

ACM ISBN 978-1-4503-6547-5/19/03…$15.00
https://doi.org/10.1145/3311823.3311869

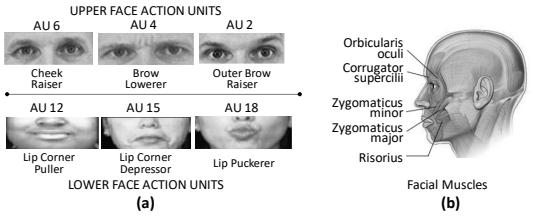

Figure 1: (a) Examples of upper face and lower face AUs in FACS and (b) typical muscles of a human face.

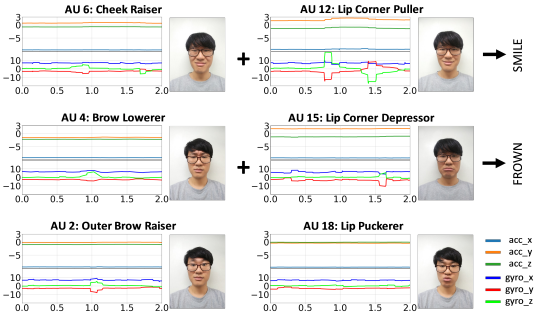

Figure 2: Patterns of inertial measurements corresponding to different AUs in FACS; top 3 lines are accelerometer values (m/s^2) and bottom 3 lines are gyroscope values (deg/s).

activities such as talking and eating are present) facial expression segments from 9 participants seen through three different sensor modalities: audio, vision, and motion. The experimental evaluations show that we achieve an average F_1 score of 0.85 with the motion dataset and the best performing HMM across both learning targets in a cross-talk free setting. As such, we consider the contribution of this work is two-fold.

- We present a first-of-its-kind study that shows the relationship between facial muscles and ear-worn IMU while smiling and frowning. We show three different learning algorithms that model this relationship for automatic recognition of smile and frown.
- We contribute a brand-new multi-modal dataset with 1,620 time-synchronised samples of audio, vision and motion segments capturing various facial expressions.

2 BACKGROUND: FACIAL EXPRESSIONS AND INERTIAL SIGNALS

Analysis and understanding of facial signals have been a long-sought research agenda across multiple disciplines. Ekman and Friesen [3] developed a descriptive Facial Action Coding System (FACS) for explaining facial expressions. In essence, FACS encodes movements in facial muscles at different parts of a human face by a set of Action Units (AUs). 30 of these AUs are anatomically related to the contraction of specific facial muscles - 12 for upper face and 18 for lower face. AUs can also occur in combination and sequence either in an additive or non-additive form. Collectively, these AUs provide a complete and powerful coding scheme to describe the details of human facial expressions. Figure 1 (a) illustrates examples of lower and upper face AUs in FACS.

In our study, we assessed which of the 30 AUs and their corresponding muscle movements are strongly visible in inertial signals captured by the accelerometer and gyroscope, placed in the ear

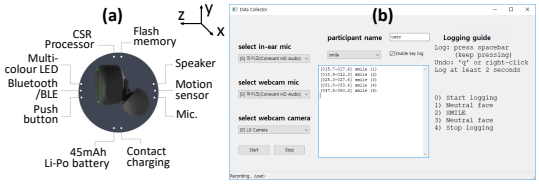

Figure 3: (a) Sensory earable and (b) data collection interface.

canal. We collected the inertial signals from 30 micro-expressions as described by 30 basic AUs in a cross-talk free setting, i.e., no other facial activities such as eating and talking were present during the task. Through careful, repeated experimentation and post analysis, we observed three recurring phenomena. First, only a subset of 30 basic AUs is clearly visible in the captured inertial signals. Second, those AUs that are strongly visible activate two specific muscles — *zygomaticus major* and *corrugator supercilii* (see Figure 1 (b)). For any AU that activates either of two muscles, their impulse response is carried to the ear canal and can be captured within the earable. Last, two upper face AUs - AU 4 (Brow Lowerer), AU 6 (Cheek Raiser) and two lower face AUs - AU 12 (Lip Corner Puller), AU 15 (Lip Corner Depressor) invariably trigger the highest impulse response in the inertial signals as shown in Figure 2; on the contrary, two AUs (AU 2 and AU 18) which do not strongly activate aforementioned muscles are not clearly visible. Interestingly, the additive combination of these AUs represents two first order facial expressions, e.g., a combination of AU 6 and AU 12 together constitutes a *smile* activating *zygomaticus major* (positive emotional valence), and a combination of AU 4 and 15 represents a *frown* activating *corrugator supercilii* (negative emotional valence). Based on these findings, we decided to develop our automatic recognition system primarily for two macro expressions — *smile* and *frown*.

3 EARABLE AND DATA DESCRIPTION

We describe the details of our earable hardware and the data collection procedure.

Earable: We prototyped a wireless earable device with multiple sensing modalities [5]. It is instrumented with a microphone, a 6-axis inertial measurement unit (3-axis accelerometer and 3-axis gyroscope), a dual-mode Bluetooth/Bluetooth low energy, and powered by a CSR processor. Figure 3 (a) illustrates the earable.

Data: We recruited 9 participants for the data collection. All were male and in 20s except one in 30s. They were asked to wear our earbud prototype and to conduct two types of facial expressions, *smile* and *frown*. The participants performed each expression 20 times while performing three activities, sitting, standing, and walking. We further obtained the *none* events by randomly segmenting the stream in the middle of consecutive expressions. In total, we collected 1,620 segments (9 participants × 3 expressions × 20 trials × 3 activities) with varying duration up to 3 seconds.

A key challenge in collecting the IMU data is the segmentation of each facial expression. Manual segmentation is extremely time and effort consuming as facial expressions are mostly subtle and momentary. Also, the duration varies even for the same type of facial expression from the same person. To address the challenge, we implemented a custom data collection tool. Figure 3 (b) shows its interface. The participants self-record the exact moments of their expressions by pressing a button. A sample video clip for each facial

Table 1: Average F_1 score of shallow classifiers

	SVM	RF	kNN	DT	NB	QDA
F_1 score	0.64	0.70	0.69	0.63	0.62	0.61

expression is provided to help the participants. In the background, the tool collects the accelerometer and gyroscope data stream from our earbud prototype and automatically segments the stream based on the button events; the sampling rate is set to 130 Hz. It also records video streams of the face as ground truth verification.

4 LEARNING ALGORITHMS

In this section, we present three different learning schemes with their experimental performance — shallow models with statistical features, HMM, and DNN. To understand the base performance, we first used the motion data collected while the participants were sitting, i.e., the most stable situations, and validated it with a 10-fold cross-validation method. We further performed an in-depth analysis with the full dataset and other validation methods. As a performance metric, we used the average of F_1 scores of three expression classes — *smile*, *frown*, and *none*.

4.1 Shallow Models with Statistical Features

Overview: As an initial attempt of developing learning algorithms with IMU signals, we investigated the traditional sensing pipelines for IMU signals, which is commonly used for physical activity recognition. The typical pipeline consists of two key components; extracting statistical features from raw IMU signals and using a machine learning classifier for the final classification [1].

Pipeline: We built a three-stage recognition pipeline; feature extraction, dimensionality reduction, and classification. As input, we took 6-axis IMU values from the expression segment. For each axis, we extracted *time-domain* features, e.g., mean and root mean square, and *frequency-domain* features, e.g., DC component and information entropy, reported in [2]. We further applied principal component analysis (PCA) to reduce the dimensionality of the data into 10 dimensions. The output vectors are fed into classifiers.

Results and implications: Table 1 shows the average F1 scores of six popular shallow classifiers: SVM, random forest (RF), kNN, decision tree (DT), naive Bayesian (NB), and quadratic discriminant analysis (QDA). We found the optimal hyperparameter values using grid search. The results show that the performance of statistical feature-based models is not satisfactory. The best performing classifier is RF, but its average F_1 score is 0.70; the F_1 score of *smile*, *frown*, and *none* classes is 0.81, 0.57, and 0.71, respectively. We conjecture that this is mainly because those statistical features could not well capture the temporal signal patterns from micro-expressions.

4.2 Hidden Markov Model

Overview: We investigate a hidden Markov model (HMM)-based learning scheme. HMM is effective to characterise sequential data with an embedded structure (here, Action Units) and also robust to variable input size [8].

Pipeline: As input, the pipeline takes a list of 8-dimensional vectors including raw values and magnitude of the 3-axis accelerometer and 3-axis gyroscope data. For the training, it constructs HMM models for each class and trains them using the standard Baum-Welch algorithm. For the recognition, we calculate the log-likelihood of an observed sequence for each class using the forward algorithm. We consider the model with the maximum log-likelihood as the recognised class. In our current implementation, we configure HMM using a 12-hidden-state left-right model with Gaussian emissions.

Table 2: Confusion matrix of HMM (F_1 score: 0.85)

		Predicted		
		smile	frown	none
Actual	smile	162	3	15
	frown	2	165	13
	none	20	29	131

Results and implications: The HMM classifier outperforms shallow models for smile and frown recognition. Table 2 shows its confusion matrix. It shows an average F_1 score of 0.85, which implies that HMM classifiers are able to capture intermittent and microscopic muscle movements during facial expressions. An average F_1 score of *smile*, *frown*, and *none* are 0.89, 0.87, 0.77, respectively. Interestingly, most of the false recognition occurs in distinguishing none expressions. We conjecture that the classifier well understands different signal patterns of *smile* and *frown*, resulting in high recognition accuracy between the two. However, those expressions with weak movements are sometimes recognised as *none* and *none* with strong movements are sometimes recognised as *smile* or *frown*.

4.3 Deep Neural Network

Overview: We further investigate the performance of the state-of-art deep neural networks (DNNs), which have shown supreme performance on many machine learning tasks. However, there is yet no well-established network for facial expression with IMU signals. Thus, we examine a convolutional network designed for physical activity recognition [9], which consists of a chain of temporal convolution layer and pooling layer prior to top-level fully connected layer and softmax group. We further devise a new architecture, namely *ConvAttention*, for performance improvement.

Pipeline: A key feature of ConvAttention is to adopt the long short-term memory (LSTM) and attention mechanism. We used LSTM to leverage the temporal pattern from micro-expression. We further adopted the attention mechanism to cope with irrelevant and noisy inputs. ConvAttention comprises two convolution layers followed by an LSTM layer that returns attention weights for each time point. The probabilities are multiplied with the feature vectors from the convolution layers, and averaged to result in a single feature vector. The feature vector is then converted into class likelihood through a fully-connected layer. To deal with the variable size of the segment, we put a global average pooling layer between the convolution layers and the fully-connected layer. We used Adam optimizer with $\alpha = 0.005, \beta_1 = 0.9, \beta_2 = 0.999$ and put Dropout ($p = 0.3$) at the output of the last convolution layer.

Results and implications: The performance of DNNs is not satisfactory, yet. The average F_1 score of CNN and ConvAttention is 0.46 and 0.75, respectively; for ConvAttention, the F_1 score of *smile*, *frown*, and *none* is 0.90, 0.79, and 0.58, respectively. We conjecture that the main reason is insufficient amount of data to effectively train the network. In case of CNN, the such low F_1 score is because it does not well characterise the facial expression, similarly to shallow models with statistical features. However, interestingly, we can

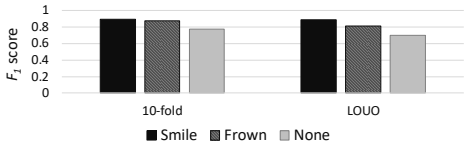

Figure 4: Impact of different validation methods.

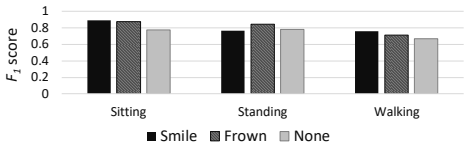

Figure 5: Recognition robustness to moving situations.

observe the performance improvement of ConvAttention. It implies that the architecture of ConvAttention well captures the signal patterns from *smile* and *frown*. We expect that higher performance would be achievable if we have a larger dataset and optimise the hyperparameter values more carefully.

4.4 Additional Benchmark on HMM

We further conducted additional benchmark on HMM, the best performing scheme.

Effect of validation methods: We compare the performance of HMM with two cross-validation methods, 10-fold and leave-one-user-out (LOUO). The average F_1 scores of 10-fold and LOUO are 0.85 and 0.80, respectively. Figure 4 shows the F_1 score of each class. Unlike *smile*, the detection performance of *frown* drops from 0.87 (10-fold) to 0.81 (LOUO). Such a drop would be mainly due to individual difference in performing frowning. According to our observation on video recordings, most of the participants made frowning as a combination of AU 4 (brow lowerer) and AU 15 (lip corner depressor), but some did only with AU 4. It would be solved by training the classifier with a larger dataset including various styles of frown expressions.

Robustness to moving situations: We further examine the performance of HMM in noisy environments, i.e., standing and walking. In standing situations, the participants were requested to perform facial expressions while standing but having natural movements of an upper body. In walking situations, they made facial expressions while they were walking around the office.

Figure 5 shows the result. We can observe non-negligible performance drops in noisy environments. The average F_1 score in standing and walking is 0.80 and 0.71, respectively. Such a drop happens in all facial expressions, indicating that the classifier does not accurately distinguish microscopic facial muscle movements from macroscopic body movements, yet. We believe that periodic patterns made by macroscopic body movements can be removed using filtering techniques such as high-pass filtering and independent component analysis (ICA). We leave it as future work.

5 OUTLOOK

While our results are promising, we conducted the experiments in a controlled setting with a few strong assumptions. Firstly, we gather the facial expressions in a cross-talk free setting, i.e., in the absence of other facial activities such as speaking and eating. We consider this as a significant limitation of this work and we intend to address this as our future work applying techniques such as ICA.

Next, in the FACS literature, AUs are associated with an intensity measure that further qualifies the strength of muscle deformations. In our current work, we did not consider AU intensity. We concur with adequate techniques this information can help substantially to improve the recognition performance.

Finally, although available in our earable platform, we have not considered audio as an additional sensing modality. Audio has been shown as a reliable indicator of various facial expressions, e.g., laugh, yawn, snore, etc. We envision to apply it together with inertial signals in the immediate future to improve the recognition capabilities of our work. We also consider such a multi-modal approach combined with AU intensity will enable as to recognise more and fine-grained facial expressions.

In this work, we explored automatic smile and frown recognition with kinetic earables. Borrowing principles from FACS, we demonstrated that IMU of an earable can capture facial muscle deformation activated by a set of temporal micro-expressions. We developed three learning schemes and evaluated their performance with the dataset collected in controlled non-conversational settings. The results show that the HMM-based scheme accurately identifies smile and frown (F_1 score: 0.85). We believe, such findings will help to uncover opportunities for brand new applications, especially in understanding the dynamics of human behaviour in the real world.

6 ACKNOWLEDGEMENTS

This research was partly supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017R1A2B3010504, NRF-2017M3C4A7066473).

REFERENCES

[1] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)* 46, 3 (2014), 33.

[2] Davide Figo, Pedro C Diniz, Diogo R Ferreira, and João M Cardoso. 2010. Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing* 14, 7 (2010), 645–662.

[3] E Friesen and P Ekman. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* (1978).

[4] Anna Gruebler and Kenji Suzuki. 2010. Measurement of distal EMG signals using a wearable device for reading facial expressions. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE, 4594–4597.

[5] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Allesandro Montanari. 2018. Earables for Personal-Scale Behavior Analytics. *IEEE Pervasive Computing* 17, 3 (2018), 83–89.

[6] Katsutoshi Masai, Yuta Sugiura, Masa Ogata, Kai Kunze, Masahiko Inami, and Maki Sugimoto. 2016. Facial expression recognition in daily life by embedded photo reflective sensors on smart eyewear. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, 317–326.

[7] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. 2016. Going deeper in facial expression recognition using deep neural networks. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–10.

[8] Lawrence R Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 2 (1989), 257–286.

[9] Fernando Moya Rueda, René Grzeszick, Gernot A. Fink, Sascha Feldhorst, and Michael ten Hompel. 2018. Convolutional Neural Networks for Human Activity Recognition Using Body-Worn Sensors. *Informatics* 5, 2 (2018), 26. https://doi.org/10.3390/informatics5020026

[10] Jocelyn Scheirer, Raul Fernandez, and Rosalind W Picard. 1999. Expression glasses: a wearable device for facial expression recognition. In *CHI'99 Extended Abstracts on Human Factors in Computing Systems*. ACM, 262–263.

[11] Inchul Song, Hyun-Jun Kim, and Paul Barom Jeon. 2014. Deep learning for real-time robust facial expression recognition on a smartphone. In *Consumer Electronics (ICCE), 2014 IEEE International Conference on*. IEEE, 564–567.