Consistent Smile Intensity Estimation from Wearable Optical Sensors

Katsutoshi Masai*†, Monica Perusquía-Hernández*‡, Maki Sugimoto†, Shiro Kumano* and Toshitaka Kimura*
*NTT Communication Science Laboratories
NTT Corporation, Atsugi, Japan
Email: {katsutoshi.masai.fa,shirou.kumano.yc,toshitaka.kimura.kd}@hco.ntt.co.jp
†Interactive Media Lab, Keio University, Yokohama, Japan
Email:sugimoto@imlab.ics.keio.ac.jp
‡Nara Institute of Science and Technology, Nara, Japan
Email:perusquia@ieee.org

***Abstract*—Smiling plays a crucial role in human communication. It is the most frequent expression shown in daily life. Smile analysis usually employs computer vision-based methods that use data sets annotated by experts. However, cameras have space constraints in most realistic scenarios due to occlusions. Wearable electromyography is a promising alternative; however, issue of user comfort is a barrier to long-term use. Other wearable-based methods can detect smiles, but they lack consistency because they use subjective criteria without expert annotation.**

We investigate a wearable-based method that uses optical sensors for consistent smile intensity estimation while reducing manual annotation cost. First, we use a state-of-art computer vision method (OpenFace) to train a regression model to estimate smile intensity from sensor data. Then, we compare the estimation result to that of OpenFace. We also compared their results to human annotation. The results show that the wearable method has a higher matching coefficient (r=0.67) with human annotated smile intensity than OpenFace (r=0.56). Also, when the sensor data and OpenFace output were fused, the multimodal method produced estimates closer to human annotation (r=0.74). Finally, we investigate how the synchrony of smile dynamics among subjects and their average smile intensity are correlated to assess the potential of wearable smile intensity estimation.

***Index Terms*—wearable computing, affective computing, smart eyewear, optical sensors, smile intensity estimation**

I. Introduction

Humans are social animals. Nonverbal cues, such as facial expressions, play an important role in social communication. In particular, smiling is frequent in daily life, not only as a happy emotion, but also as a means to mediate social relationships [1]. The positive and social effect of smiling at others or oneself is applied in the area of affective computing [2] and human-computer interaction to improve human well-being. Tsujita and Rekimoto developed a device that stimulates smiling to produce a positive mental effect on the user [3]. Nakazato et al. studied the effect of computer-assisted deformation of users' facial expressions to make them appear to be smiling during a remote brainstorming session to enhance creativity [4]. Other applications include evaluating video content preferences [5]; extracting important scenes in life-logging videos [6]; and promoting the sharing of smiles between children with autism spectrum disorders (ASD) and their parents and supporters [7].

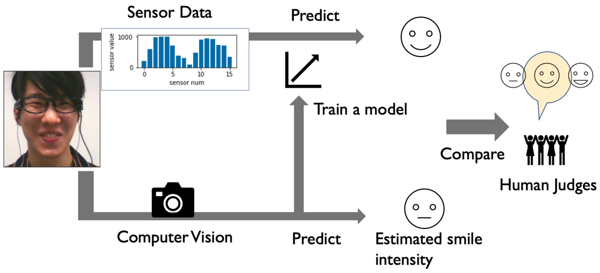

Fig. 1. Concept of our work. First, we use a state-of-art computer vision method to train a regression model to estimate smile intensity from wearable optical sensors. Then, we compare the estimation results. Finally, we compared the results to human annotation.

It is, of course, important to decode the meaning of a smile. The predominant approach is based on the Facial Action Coding System (FACS) [8]. FACS enables an objective description of visible facial movements by action units (AUs) that correspond to the movement of specific facial muscles or groups of muscles. The AU12 (Lip Corner Puller, the activation of the *Zygomaticus Major* muscle) is the prototypical movement present in every smile. Besides labeling the presence of one or more AUs, FACS describes facial expressions in terms of their AU intensity. Intensity is coded according to the degree of muscle contraction using a six-level scale, including absence. The dynamics of smile intensity are useful for understanding the implications of smiles [9]–[11].

Since the manual annotation of AUs is time-consuming and cumbersome, automatic approaches by computer vision (CV) have emerged. CV researchers have achieved high accuracy in basic emotion recognition, including smiles and AU detection in controlled conditions, as confirmed by comparison to annotations by experts. However, recognizing facial expressions in natural environments remains challenging due to privacy concerns, occlusions, and facial direction and positioning.

Electromyography (EMG) has long been used to measure facial expressions. The advantage of the method is that it can recognize subtle expressions by placing electrodes on the skin directly above the facial muscles to be measured.

978-1-6654-5908-2/22/$31.00 ©2022 IEEE

However, placing many electrodes on the face is seldom accepted in everyday situations. For this reason, attempts have been made to recognize facial expressions by placing electrodes on more insignificant parts of the face. For instance, distal EMG sensors are a promising wearable alternative to the CV-based approach as their recognition performance is comparable with that of CV [12]. However, concerns about the long term comfort of the device remain as it is directly in contact with the skin surface. In addition, the signal is sensitive to changes in electrode position and skin conditions, such as individual differences and device reattachment, making stable measurement difficult.

Optical sensors combine the properties of CV and EMG sensors. They enable wearable measurement with simple configuration and high wearing comfort. In fact, optical sensors have been used for both recognizing basic facial expressions [13] and estimating the intensity and genuineness of facial expressions [6], [14]. However, annotation consistency remains a problem as annotation was by non-experts. Furthermore, no comparison or combination with CV-based methods has been investigated. The comparison is important to indirectly utilize the human annotation from CV to understand and handle sensor characteristics in smile intensity estimation. Wearable optical sensors that can well estimate smile intensity will expand the range of everyday application, such as monitoring the wearer's mood, aiding communication for people with ASD, and ambulatory analysis of smile in social interactions.

This study attempts to estimate smile intensity consistently and continuously by using photo-reflective sensors on smart eyewear while minimizing manual annotation costs. The sensors measure the skin deformation caused by facial muscle activity from reflective intensity. We use the dataset from [14] which includes the sensor data from smart eyewear and a video of facial expressions while the participants watched a comedy video. First, we annotated smile intensity using OpenFace [15], [16], state-of-the-art in CV AU intensity estimation. Then, we built a multiple regression model with OpenFace's AU12 intensity as the explained variable for smile intensity and sensor data as the explanatory variable to predict smile intensity from sensor data. Next, we compared this sensor-based method to OpenFace. Then, we determined whether the sensor-based or OpenFace estimates were closer to smile intensity as annotated by humans. Finally, we performed a case analysis of smiles in the dataset to show the potential of the method.

The key findings are: 1) The sensor data-based estimates and OpenFace output were highly correlated (r=0.93 on average). This result shows that smile intensity can be estimated consistently by wearable optical sensors with at least as much accuracy as OpenFace. 2) Human annotation revealed that the sensor-data-based estimates were closer to annotated smile intensity than those of OpenFace. In particular, the sensor-based estimate is better under face occlusion and when the participants smile with a closed mouth like a chuckle and a subtle smile. 3) We analyzed how the synchronicity of smile dynamics among subjects and their average smile intensity are correlated. If smile dynamics were similar, their smile intensity increased.

This paper offers a new contribution to the field of wearable smile intensity estimation: 1) targeting spontaneous smiles using photo reflective sensors, 2) comparing the intensity of smiles with CV-based state-of-art methods, and 3) showing that wearable methods can estimate smile intensity consistently and continuously by utilizing the OpenFace annotation.

II. Related Work

This section summarizes the methods of smile detection and the intensity estimation from the sensor perspective: computer vision, wearable sensors, and optical sensors.

CV-based methods have been used in many studies of facial expression analysis, including smiles. Recently, recognizing subtle changes in smiles has become an active research topic, such as distinguishing between forced and spontaneous smiles [10] and estimating the intensity of smiles [17], [18]. However, a camera may not be able to capture facial expressions stably due to obstructions such as hands or due to head motion of the subject. Therefore, it is not suitable for daily life situations. In addition, it has difficulty in measuring subtle expressions. Furthermore, constantly recording by a camera raises privacy concerns, and the anxiety of being filmed may prevent natural facial expressions.

Wearable sensors for facial expressions, including smiles, can overcome these limitations. Many researchers have proposed wearable systems such as EMG for detecting positive expressions [19], emotional valence [**?**], subtle smiles [21], smile-related action units [22], "enjoyment", "social" and "masked" smiles for long-term recordings [23], and classifying posed and spontaneous smiles [24], EOG glasses to detect upper action units [25], electric field sensing technology using electrodes on-ear canal [26], capacitive sensors [27], ultrasonic sensors [28], and optical sensors [6], [13], [14] to measure facial expression movement. Other studies use the movement of the abdomen and diaphragm when laughing and detect it as pressure changes at the abdomen by using capacitive e-textile [29] or voice [30].

Wearable technologies for estimating the intensity of a smile or smile-related action units have been investigated. Rantanen et al. identified three levels of AU12 intensity using multichannel capacitance measurements based on EMG value at maximum smile intensity [31]. Iravantchi et al. used an acoustic interferometry technique with ultrasonic transducers to estimate an arbitrary 4-level smile intensity that users found easy to repeat and generate [28]. Yet, these collected the data of intentional facial expressions under controlled conditions in a laboratory setting. Spontaneous smiles have more varied dynamics than posed smiles and contain many noise factors such as head movements [32]. The method of Fukumoto et al. [6] detects smile intensity with a simple system configuration and algorithm, illustrating the potential of optical sensors. However, the intensity of the ground truth movement was unclear because their algorithm applied an arbitrary threshold to cheek and eye muscle movements. In addition, eye movements alone were not regarded as smiles, which limits the variations of smiles that can be detected. It may not be able to detect small smiles. Furthermore, this method is not suitable for analyzing the dynamics of smiles because it only provides discrete classification.

Among these methods, optical sensors have great potential as wearable devices because they are small, highly responsive, offer low processing costs, and high comfort due to non-contact measurements. Furthermore, the small sensors can yield socially acceptable devices such as eye-glasses, enabling private sensing hidden from observers. In addition to Fukumoto et al.'s method [6], Masai et al. [13] showed that wearable optical sensors can identify facial expressions associated with basic emotions, Saito et al. [14] discriminated between spontaneous and posed smiles, and Asano et al. [33] reproduced facial expression geometry. Unfortunately, these studies did not estimate the smile intensity continuously. Also, no comparison between optical sensors and CV/human annotation was made in smile intensity estimation. We believe that a comparison with CV, the most common method for estimating expression intensity, would clarify the advantages and disadvantages of the methods and allow us to build a consistent estimator without human annotation. Of particular importance, both photo-reflective sensors and cameras measure optical information and have high data affinity to human vision perception.

III. Dataset

We used the dataset from [14], which provided time series data from 16 photo-reflective sensors on smart eyewear and the concurrent image data when the participants watched a comedy video that induced smiles and laughter. The data collection was approved by an ethical committee in Keio university. First, we describe the recording setup, followed by a detailed description of the dataset.

Fig. 2 shows the appearance of the device and sensor layout. The 16 photo-reflective sensors were embedded in smart eyewear [14]. The photo-reflective sensors consist of infrared (IR) LEDs and IR phototransistors. The sensors work as proximity sensors to measure facial expression activity. The skin deformation caused by facial muscle movements changes the distance between the sensors on smart eyewear and the wear's skin surface. For example, in a smiling expression, the skin around the cheek is deformed and the distance between the corresponding sensors and the skin surface becomes shorter than in a neutral expression. Therefore, the intensity of the reflection measured by each sensor varies in response to the muscle movements. We estimate smile intensity from the sensor features.

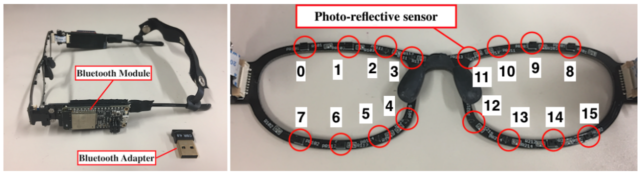

Fig. 2. Eyewear device and its sensor layout. Source: Adapted from [14].

The data was collected from 12 participants (8 males, 4 females, average age = 25, all Japanese) while watching a 13 minute 28 second comedy video alone. The sensor data from the eyewear device (30Hz sampling frequency) and image data (size of 630X320) from the built-in camera of a PC (30 fps) were recorded. These data were synchronized with the frames of the viewed video via time stamps. Any missing data were offset by using the data from one frame earlier. Overall, each participant yielded 24245 data samples.

Smiling is defined by the FACS as including movements of AU6 (Cheek Raiser) and AU12. According to the benchmark results from Cheong et al. [36], the estimate of AU6 and AU12 from OpenFace [15], [16] is best, with F1 scores of 0.81 and 0.83, respectively. The performance of OpenFace was also confirmed by Perusquia-Hernandez et al.'s comparison [22]. Therefore, we used OpenFace to output the intensity of AU6 and AU12 for each image frame in the dataset. The correlation between these two values was high, with an average of 0.934 (SD=0.0357) for the 12 users. OpenFace can estimate an intensity ranging from 0 (absent) to 5 (maximum), and we consider values greater than 1 to be smiling. Since OpenFace only extracts the intensity of pure facial movements, our study focus on this definition of intensity.

We used AU12 intensity as the ground truth of smile intensity of facial features movements, because it is the prototypical movement that signals a smile. We also expect a high correlation with AU6, given the sensors' close proximity to the eyes, and the nature of smiles. AU6 often co-appears with smiles, but it not always present. Many researchers define smile intensity by considering both AUs or only AU12. The Smiling Intensity Scale [34] by Gironzetti et al. uses these AUs to classify smile intensity into five categories. Witzig et al. also classifies smile intensity into four stages [35] to create a dataset for smile intensity detection by the deep learning method [18]. Both research present an annotation scale only on smiles and not laughs. For the data set, trained coders annotated the intensity by checking AU6 and AU12 visually. On the other hand, these two AUs are said to co-occur in expressions of happiness, but AU12 may occur independently at low intensity [37]. According to Ruan et al. [38], attention to mouth movement (AU12) improved the ability to classify spontaneous smiles and posed smiles. Girard et al. [17] classified the smile intensity using the multi-class SVM with the results of AU12 annotation from the certified coder as the ground truth. Moreover, the output of AU6 may be less accurate due to shielding by the eyewear device.

The comedy video elicited spontaneous smiles and laughter from the participants, with 40.6% (5 minutes 28 seconds) of the scenes achieving an AU12 intensity of 1 or greater and 13.8% (1 minute 51 seconds) achieving an AU12 intensity of 2 or greater on average. The video includes pre-recorded audience laughter and other reactions to the humorous materials as a social proof that the materials were funny [39]. Therefore, one of the authors annotated the onset and offset of the laughter reactions in the video using ELAN [40]. The laughter reactions were assigned to 71 scenes in the video footage. The data for two seconds before and after the laugh tracks showed that the ratio of AU12 intensity 1 or higher was 53.4% and the ratio of AU12 intensity 2 or higher was 22.7%, which is more laughter than the average for the entire video.

IV. Analysis

The optical sensing method focuses on the facial expression changes around the eyes during smiling because the device measures the skin deformation around the eyes. We, however, selected AU12 as the video-based ground truth, because that is the distinctive element of a smile. By learning AU12 as the ground truth, we expect that the sensor-based estimation will reflect both the eye and mouth areas.

First, we compare this method with the OpenFace output to see how well the device can estimate smile intensity. Then, we investigate the differences between the sensor-based estimation and OpenFace through a qualitative analysis and human annotation comparisons.

A. How similar is the sensor data and OpenFace output?

We examined the correlation between each of the 16 sensors and the AU12 intensity of OpenFace to understand the characteristics of the sensors on the device used for data recording. Fig. 3 shows the correlation coefficients. The six sensors with average correlation coefficients above 0.7 (sensor numbers: 5, 6, 7, 13, 14, and 15) measured skin deformation from under the eyes to the cheeks (see Fig. 2 for sensor number and placement). The correlation coefficient of the sensors on the upper part of the device with the OpenFace output was around 0.5 at maximum (sensor numbers: 2, 3, 10, and 11). We observed the individual differences were due to hair entrapment between the sensors and the skin, the fitting of the glasses, and the power supply noise.

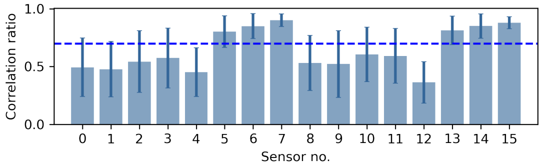

Fig. 3. Correlation coefficient between each sensor and AU12 Output of OpenFace. The error bar indicates the standard deviation.

We compared the AU12 output of OpenFace to the sensor-based estimates to evaluate their similarity. First, the values of each sensor were z-scored in the temporal domain and compressed into an n-dimensional feature vector using principal component analysis (PCA). Next, we made a multiple linear regression model with this vector as the explanatory variable and OpenFace's AU12 intensity output as the explained variable. Finally, we segmented the time series data into five segments and applied one-time segment-leave-out cross-validation to obtain a sensor-based estimation of smile intensity. Since most of the training data are facial expression data without smiles, we classified the training data into eight levels of smile intensity by 0.5 (0.0-0.5, 0.5-1.0, etc.) and randomly undersampled the data so that the number of data in each category was equal to or less than 400. We corrected the cases where the estimated intensities were negative to zero.

To avoid overfitting, we compared the number of dimensions of the feature vectors reduced by PCA. Fig. 4 shows the values of the average mean squared error (MSE) of the estimation (error bars show the maximum and minimum values out of 12 participants, respectively) when the sensor data was compressed to 1-15 dimensions. From the figure, we chose eight dimensions, since the MSEs for all users were less than 0.1. Fig. 5 shows the maximum of the cross-correlation between the time series of estimated values and the AU12 intensity of OpenFace when training and fitting were done for each individual. The average correlation was 0.923 (SD=0.043), and the average MSE was 0.065 (SD=0.028). The cross-correlation showed an average lag of 11.1 ms (SD=28.3 ms) in the output from the OpenFace compared to the estimation from the sensors. We considered this lag to be related to the difference in synchronization or the difference in the characteristics of smiles, given the standard deviation of the lag. We do not think it is a difference due to the nature of the measurement method [22] as is found in the distal facial EMG approach.

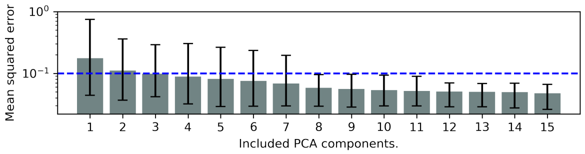

Fig. 4. MSE between sensor-based estimation and AU12 Output of OpenFace with different principal components. Error bar indicates the max and min of MSE among the participants.

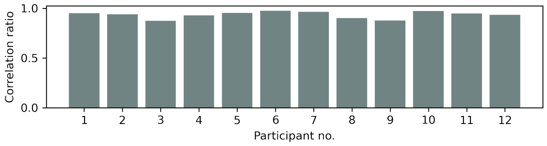

Fig. 5. Cross Correlation between sensor-based method and AU12 Output of OpenFace for each participant. The sensor data contains information about smile intensity. By learning the results of OpenFace, the sensor-based method can estimate the smile intensity to the same level as OpenFace.

B. How different is the sensor data and OpenFace output?

These results show a strong correlation between OpenFace and sensor data. However, there are cases where the output of OpenFace differs from the sensor-based estimation. For cases where the difference was detected by a peak detection algorithm, we compared the pros and cons of sensor data over OpenFace output through a qualitative analysis and a human annotation comparison. To understand the differences between OpenFace and sensor-based estimation, we extracted data points that differed in trend, not instantaneous differences, by using a low-pass filter and peak detection.

Fig. 6 shows an example of a change in sensor values, OpenFace outputs, their low-pass outputs, and the sequences of facial expressions at that time. From the sensor-based estimation, some participants showed strong blinking and movements of facial muscles around the nose (e.g., AU10) that were not associated with smile intensity. These noises occurred in a higher frequency band than the temporal changes in sensor values caused by smiling movements. Therefore, we applied a low-pass filter to the estimation so that high-frequency bands above 10 Hz were attenuated. The same procedure was applied to the OpenFace output when detecting peaks to ensure data processing consistency.

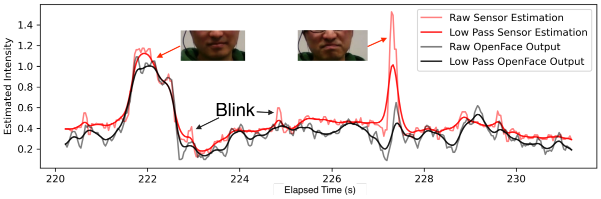

Fig. 6. A 10 Hz low-pass filter was used to remove blinks and irrelevant movements.

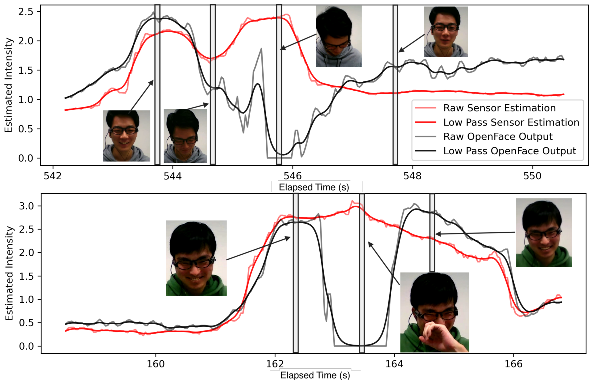

Fig. 7. Examples of gaps in the OpenFace detection are smoothly detected by the wearable sensor.

We applied the peak detection algorithm (using Scipy [41], prominence=10, i.e., 0.5 on average in the time window, width=1) to the time series of the absolute sum of the differences within a time window of 20 frames. Then, we extracted the maximum difference points between OpenFace output and the sensor-based estimates from each detected peak width. Finally, we excluded the data points determined to be non-smile (i.e., estimated value of less than 1) by both methods, resulting in 135 samples from 12 users.

We identified two cases in which the sensor-based estimates and OpenFace were extremely different (Fig. 7). The first is the case where the face could not be detected due to head motion (2 samples), as shown at the top of the figure, and the second is the case where the face was partially occluded by a hand (1 sample). In these cases, we can infer from the temporal contexts that the smiling state continues, but the output from OpenFace is zero or close to zero – the prediction was not accurate. On the contrary, the sensor-based estimation can still make predictions as the device is worn. This indicates that the sensor-based estimates are more reliable than OpenFace in such cases. Therefore, we excluded them from the subsequent analysis as outliers.

We analyzed the data qualitatively. We applied the unsupervised clustering method (k-means) to classify 132 samples (135 minus outliers) into five classes. Heuristics using the elbow method determined the number of classes. As humans change how they perceive facial expressions depending on dynamic perceptions, we designed the features to consider the temporal changes. We used 6-dimensional features: the OpenFace estimates, the sensor estimates, each data point before and after ten frames.

The left figure is a two-dimensional representation of the six-dimensional features with their first and second principal components. The right figure shows the axes of the estimated values from the sensors and the output from OpenFace. For the color selection of the visualization, we used Colorgorical [42]. Fig. 9 shows the representative image for each class, with the data points for each image closest to the centroid of each class in k-means clustering.

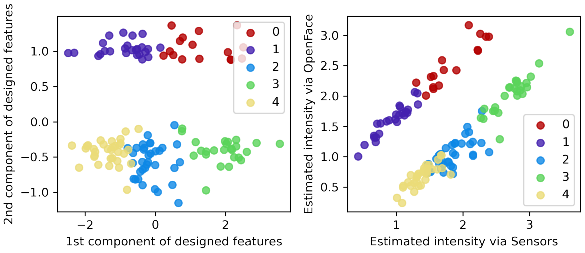

Fig. 8. The five categorical classes using k-means clustering.

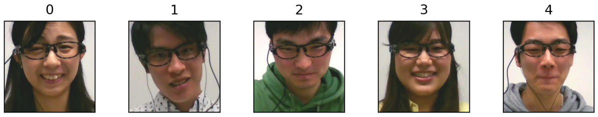

Fig. 9. Representative images for each class.

TABLE I
THE AVERAGE ESTIMATES FOR EACH CLASS.

Class	0	1	2	3	4
Sensors	1.84	0.93	1.85	2.75	1.40
OpenFace	2.53	1.6	1.04	2.05	0.69

The Table I shows the average estimates from OpenFace and the sensors for each class. Fig. 9 shows the representative images. They illustrate the differences in characteristics that the two methods measure. The five classes fall into two main patterns: cases where the sensor estimate is less than the OpenFace (class 0 and 1) and cases where the sensor estimate is greater than the OpenFace output (class 2, 3, and 4). Class

0 included intense smiles showing teeth at the onset. The class also included the activities of other facial action units and head movements that influenced the estimation results. With head movements, the proportion of mouth in the face tended to be more prominent, depending on the light exposure and the angle of the face. The smiles in Class 1 are more pronounced around the mouth than around the eyes. Class 2 and 4 included closed-mouth smiles and chuckles shown at the offset of smiles. Class 4 also included a small smile at the onset. Class 3 has big smiles showing the teeth, but the sensor-based estimates showed higher intensity than OpenFace.

C. Human Annotation Comparison

We collected a human annotation to examine which automatic facial recognition method, the sensor-based or OpenFace, better reflected the smile intensity of facial movements annotated by laypersons. Four Japanese people (2 males, 2 females) volunteered to annotate by answering the questions on the implemented program from their personal computers. All participants had normal vision or corrected-to-normal vision, and they reported that they had no difficulty in reading facial expressions. The annotation program was web-browser based, and implemented using the javascript-based jsPsych library [43]. To save the annotation cost, we focused on data points within 30 frames before and 60 frames after the laughter reactions, the scenes where people smiled frequently. Also, we assumed that data points from the same person in the same class would have similar characteristics, so we selected a maximum of five data points from each class of each participant in order of the largest estimation error. Overall, we used 83 data points for the annotation.

Since the annotators were not trained, we created a smile intensity scale of facial movements, see Fig. 10, for each individual in the dataset, which consisted of a series of images for different smile intensity levels to establish a clear criterion for the smile intensity assessment. The images were those in which the sensor-based estimation and OpenFace outputs had small error less than 0.1). We consider the intensity from the images were close to human annotated smile intensity of facial movements, as the sensors and OpenFace agreed. The scale has a five-step intensity, ranging from 0.5 to 2.5, where the corresponding smile intensity was present in all experimental participants. We used the face_recognition[1] program based on dlib library [44] to extract facial area, exclude closing eyes, and enlarge and adjust the brightness to increase visibility. Fig. 10 is an example of the smile intensity scale.

The annotation procedure was as follows. First, the overall process was explained, and the consent was obtained. Next, the following procedure was repeated for all the data points. The annotators were allowed to play the video only once to measure immediate impressions. The order of the videos was randomized. The videos did not include audio.

1) The annotators watched two seconds videos that ended at the images corresponding to one of the 83 data points.

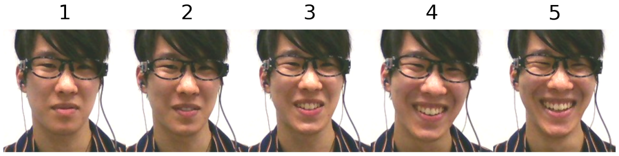

Fig. 10. An example of smile intensity scale used for human annotation experiment. The images are retrieved when the sensor-based estimation and OpenFace output matched.

2) They rated the impression of smile intensity of facial movements in the last frame of the video, referring to the scale mentioned earlier. The rating ranges with an eleven-level Likert scale, including five images of the scale and the middle and before and after.
3) They rated how natural and genuine they perceived those smiles with a five-step Likert scale from their video impression.

D. Results

There was a high correlation (r=0.82) between the annotated smile intensity of facial movements and perceived smile genuineness. The correlations between the smile intensity/genuineness and the sensor-based estimates and OpenFace output were 0.67/0.56 and 0.59/0.36, respectively. The mean absolute error between the average human intensity score and each method was 0.37 for the sensors and 0.48 for CV estimation. Then, we considered a simple error optimization equation to combine the sensor and CV estimation results under the constraint that the matched parts are not affected as follows: $sensor * (1 - \alpha) + CV * \alpha$ subject to $0 \leq \alpha \leq 1$. This shows a minimum error of 0.29 when alpha is 0.34. The correlations between the smile intensity/genuineness and combined estimation was 0.74/0.56. The results for each class also showed that in Class 4, 92% of the data were closer to the results estimated from the sensor data to human annotation than those of OpenFace. This means the method using optical sensors can estimate the smile intensity better than the CV method, especially in the case of closed-mouth smiles, such as a giggle and subtle smile.

E. Case Study: Smile Dynamics Analysis using Sensors

The sensor-based results trained with the CV-based approach outcome showed that the smile intensity estimation method is reliable. This section further examines the potential of sensor-based estimation by observing the data in detail. We focus on the relationship between the smile synchronicity among users and the maximum rate of increase in smile intensity using these estimation results.

First, at a given time, *t*, we define the change in the smile intensity estimated by the sensor over the previous ten frames as *the smile change rate* at *t*. Fig. 11 shows *the smile change rate* before and after laughter reaction onsets for all smile scenes, cropped and averaged, and arranged in time series for each user. The onset of laughter reactions is fixed to the 60th frame, and *the smile change rate* at that frame is the data

[1] https://github.com/ageitgey/face_recognition by Adam Geitgey

calculated from frame 51 to frame 60. This figure shows that *the smile change rate* increases roughly from the onset of the reactions and peaks at the 81.5 frame (SD =2.90) on average. This trend is observed for all users. The laughing reactions accelerated the smile intensity of those who watched the video. We used these onsets as reference points of the video stimulus.

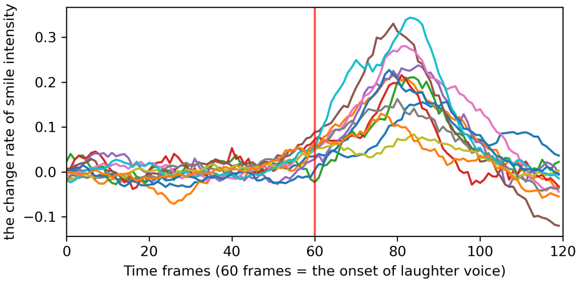

Fig. 11. Smile intensity dynamics on average estimated from wearable optical sensors. A red vertical line indicates the onset of the laughter reactions.

Next, we used these reference points to examine the relationship between *the smile change rate* and smile intensity dynamics among the participants in the dataset. The estimated smile intensity in the 60 frames before and after each laugh reaction was used to calculate and average the correlations for all 12 pairs. We took this as an indicator of the degree of synchronization of the smile dynamics among subjects. Fig. 12 maps these data and the maximum rate of smile change in this interval. From this figure, the higher the degree of synchronization of smile, the higher the maximum smile change rate for all participants (Coefficient of determination: 0.57). This result indicates that many subjects laughed similarly if the rate of increase in smile intensity was high. This ability of wearable devices to dynamically estimate and analyze the intensity of such smiles has the potential in examining the role of smiles in social situations such as daily communications and collaborative tasks and understanding its mechanism.

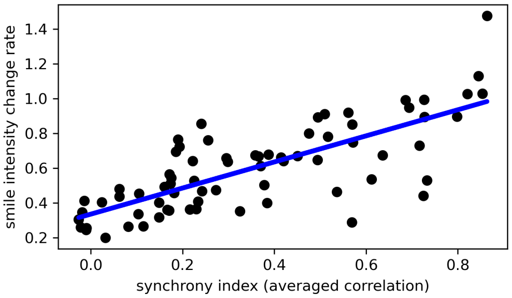

Fig. 12. Smile synchronicity analysis for wearable optical sensors. The higher the degree of synchronization of the smile, the higher the maximum smile change rate for all participants.

V. Discussion and Future Work

The trained sensor-based estimation has high correspondence (r =0.923) with AU12 intensity estimated from OpenFace. This indicates that the optical sensor-based method can estimate the smile intensity consistently without manual annotation by learning the OpenFace output. We also confirmed that sensor-based estimation is closer to human annotation than OpenFace and that the best estimation result was the combination of the two. This suggests that the accuracy and applicability of the method could be improved in two ways: 1) using the sensor data results to fine-tune the image-based results, and 2) calibrating the sensor data results from the image-based results and using them in more realistic scenarios.

The OpenFace output was unstable when the lighting environment or the position of the face changed. For example, even if changes in facial expression could not be observed, the OpenFace output increased when the size of the mouth relative to the face increased due to face angle change. Furthermore, OpenFace did not detect the subtle smiles that sensor-based estimation could catch. Hence, a future challenge is understanding the sensor characteristics by directly comparing sensor values and human annotations as a ground truth. Also, in the present dataset, few factors could be considered noise for the optical sensors other than head motion and blinks. The participants did not move their mouths for conversation and did not show facial movements associated with other emotions, such as AU9 (Nose Wrinkler) and AU10 (Upper Lip Raiser). We would like to validate the wearable method in more natural interactions. When several categories of facial expressions are present, the facial expression recognition method [13] may be applied beforehand to accommodate noise factors for enhanced smile intensity estimation.

We did not distinguish laughs from smiles in this work. The annotators' perception of smiles might be affected by the presence of laughter. Future work should investigate the influence of other body movements associated with laughter on the perception of the intensity of smiles as a facial expression. Also, future work should explore whether our results will hold valid for other people than Japanese participants.

VI. Conclusion

We presented a wearable-based method to continuously estimate smile intensity using optical sensors without manual annotation. The method can output intensity estimates matching between the state-of-art CV method (OpenFace) and sensor data, when trained with a linear regression. The results showed the method can output estimates even in situations where the facial expressions cannot be recognized by OpenFace. When the two estimation methods output different intensity levels, human annotated smile intensity was closer to that of the sensors than to OpenFace. This showed the potential of the wearable method in estimating smile intensity dynamics. Furthermore, we analyzed the relationship between the rate of change of smile intensity and the degree of synchronization of smile intensity dynamics. When the participants smiled synchronously based on the canned laughter in the stimuli, they showed more intense smiles. These methods would be helpful in understanding the synchronization of smiles in more realistic social interactions such as face-to-face communication.

References

[1] J. Martin, M. Rychlowska, A. Wood, and P. Niedenthal, ”Smiles as multipurpose social signals,” *Trends in cognitive sciences*, 21(11), pp.864-877. 2017.

[2] Picard, Rosalind W. Affective computing. MIT press, 2000.

[3] H. Tsujita and J. Rekimoto, ”Smiling makes us happier: enhancing positive mood and communication with smile-encouraging digital appliances,” in *Proc. 13th Int. Conf. Ubiquitous Comput.*, 2011. pp. 1-10.

[4] N. Nakazato, S. Yoshida, S. Sakurai, T. Narumi, T. Tanikawa and M. Hirose, ”Smart face: enhancing creativity during video conferences using real-time facial deformation,” in *Proc. 17th ACM Conf. on Computer supported cooperative work and social Comput.*, ACM,2014, pp. 75-83.

[5] D. McDuff, R. E. Kaliouby, J. F. Cohn and R. W. Picard, ”Predicting Ad Liking and Purchase Intent: Large-Scale Analysis of Facial Responses to Ads,” in *IEEE Trans. on Affect. Comput.*, vol. 6, no. 3, pp. 223-235, 1 July-Sept. 2015.

[6] K. Fukumoto, T. Terada, and M. Tsukamoto, ”A smile/laughter recognition mechanism for smile-based life logging,” in *Proc. of the 4th Augmented Human Int. Conf.*, ACM, 2013, pp. 213-220.

[7] Y. Takano and K. Suzuki. 2014. ”Affective communication aid using wearable devices based on biosignals,” in *Proc. of the 2014 Conf. on Interaction design and children (IDC ’14)*. ACM, New York, NY, USA, 213–216.

[8] P. Ekman, and W. V. Friesen, ”Facial action coding system,” *Environmental Psychology Nonverbal Behavior*. 1978.

[9] J. F. Cohn and K. L. Schmidt, ”THE TIMING OF FACIAL MOTION IN POSED AND SPONTANEOUS SMILES,” *Int. Journal of Wavelets, Multiresolution and Information Processing,* 02, 02 (2004), 121–13.

[10] M. Kawulok, J. Nalepa, J. Kawulok, B. Smolka B, ”Dynamics of facial actions for assessing smile genuineness,” *PLOS ONE* 16(1): e0244647. 2021.

[11] E. G. Krumhuber, A. Kappas, A. S. R. Manstead, ”Effects of Dynamic Aspects of Facial Expressions: A Review,” *Emotion Review*. 2013;5(1):41-46.

[12] M. Perusquia-Hernandez, 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 *Proc. of the 2019 CHI Conf. on Human Factors in Comput. Systems- CHI’19*, 2019, pp. 1-9.

[13] K. Masai, K. Kunze, Y. Sugiura, M. Ogata, M. Inami, and M. Sugimoto, ”Evaluation of Facial Expression Recognition by a Smart Eyewear for Facial Direction Changes, Repeatability, and Positional Drift,” *ACM Trans. Interact. Intell. Syst.* 7, 4, Article 15 (December 2017), 23 pages.

[14] C. Saito, K. Masai and M. Sugimoto, ”Classification of spontaneous and posed smiles by photo-reflective sensors embedded with smart eyewear,” in *Proc. of the 14th Int. Conf. on Tangible Embedded and Embodied Interaction TEI ’20*, 2020, pp. 45-52.

[15] T. Baltrusaitis, A. Zadeh, Y. C. Lim and L. Morency, ”OpenFace 2.0: Facial Behavior Analysis Toolkit,” in *13th IEEE Int. Conf. on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 59-66.

[16] T. Baltrušaitis, M. Mahmoud and P. Robinson, ”Cross-dataset learning and person-specific normalisation for automatic Action Unit detection,” in *11th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG 2015)*, 2015, pp. 1-6,

[17] J. M. Girard, J. F. Cohn, and F. De la Torre, ”Estimating smile intensity: A better way”, *Pattern recognition letters*, 66, 2015, 13-21.

[18] P. Witzig, J. Kennedy and C. Segalin, ”Smile Intensity Detection in Multiparty Interaction using Deep Learning,” in *8th Int. Conf. on Affect. Comput. and Intelligent Interaction Workshops and Demos (ACIIW)*, 2019, pp. 168-174.

[19] A. Gruebler and K. Suzuki, ”Design of a Wearable Device for Reading Positive Expressions from Facial EMG Signals,” in *IEEE Trans. on Affect. Comput.*, vol. 5, no. 3, pp. 227-237, 1 July-Sept. 2014.

[20] W. Sato, K. Murata, Y. Uraoka, K. Shibata, S. Yoshikawa and M. Furuta, ”Emotional valence sensing using a wearable facial EMG device,” *Sci Rep* 11, 5757 (2021).

[21] M. Perusquía-Hernández, M. Hirokawa and K. Suzuki, ”A Wearable Device for Fast and Subtle Spontaneous Smile Recognition,” in *IEEE Trans. on Affect. Comput.*, vol. 8, no. 4, pp. 522-533, 1 Oct.-Dec. 2017.

[22] M. Perusquía-Hernández, F. Dollack, C. K. Tan, S. Namba, S. Ayabe-Kanamura and K. Suzuki, ”Smile Action Unit detection from distal wearable Electromyography and Computer Vision,” *16th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG 2021)*, 2021, pp. 1-8.

[23] L. Inzelberg, D. Rand, S. Steinberg, M. David-Pur and Y. Hanein, ”A Wearable High-Resolution Facial Electromyography for Long Term Recordings in Freely Behaving Humans,” *Sci Rep* 8, 2058 (2018).

[24] M. Perusquía-Hernández, M. Hirokawa and K. Suzuki, ”Spontaneous and Posed Smile Recognition Based on Spatial and Temporal Patterns of Facial EMG”, *Proc. of the 7th Affect. Comput. and Intelligent Interaction Conf.*, pp. 537-541. 2017.

[25] S. Rostaminia, A. Lamson, S. Maji, T. Rahman, and D. Ganesan. ”W!NCE: Unobtrusive Sensing of Upper Facial Action Units with EOG-based Eyewear,” in *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 23 (March 2019), 26 pages.

[26] D. J. C. Matthies, B. A. Strecker, and B. Urban, ”EarFieldSensing: A Novel In-Ear Electric Field Sensing to Enrich Wearable Gesture Input through Facial Expressions,” in *Proc. of the 2017 CHI Conf. on Human Factors in Comput. Systems (CHI ’17)*. ACM, New York, NY, USA, 2017, 1911–1922.

[27] D. J.C. Matthies, C. Weerasinghe, B. Urban, and S. Nanayakkara, ”CapGlasses: Untethered Capacitive Sensing with Smart Glasses,” in *Proc. of Augmented Humans Conf. 2021 (AHs’21)*, ACM, New York, NY, USA, 2021, 121–130.

[28] Y. Iravantchi, Y. Zhang, E. Bernitsas, M. Goel, and C. Harrison, ”Interferi: Gesture Sensing using On-Body Acoustic Interferometry,” in *Proc. of the 2019 CHI Conf. on Human Factors in Comput. Systems (CHI ’19)*. ACM, New York, NY, USA, Paper 276, 2019, 1–13.

[29] A. Shimasaki and R. Ueoka, ”Laugh Log: E-textile Bellyband Interface for Laugh Logging,”. in *Proc. of the 2017 CHI Conf. Extended Abstracts on Human Factors in Comput. Systems (CHI EA ’17)*. ACM, New York, NY, USA, 2017, 2084–2089.

[30] J. Gillick, W. Deng, K. Ryokai, D. Bamman. ”Robust Laughter Detection in Noisy Environments,” in *Proc. Interspeech 2021*, 2481-2485.

[31] V. Rantanen et al., ”Capacitive Measurement of Facial Activity Intensity,” in *IEEE Sensors Journal*, vol. 13, no. 11, pp. 4329-4338, Nov. 2013.

[32] M. Perusquía-Hernández, S. Ayabe-Kanamura and K. Suzuki, ”Posed and spontaneous smile assessment with wearable skin conductance measured from the neck and head movement,” in *8th Int. Conf. on Affect. Comput. and Intelligent Interaction (ACII)*, 2019, pp. 199-205.

[33] N. Asano, K. Masai, Y. Sugiura, and M. Sugimoto, ”Facial performance capture by embedded photo reflective sensors on a smart eyewear,” in *Proc. of the 27th Int. Conf. on Artificial Reality and Telexistence and 22nd Eurographics Symposium on Virtual Environments (ICAT-EGVE ’17)*, Eurographics Association, Goslar, DEU, 21–28.

[34] E. Gironzetti, S. Attardo and L. Pickering, ”Smiling, gaze, and humor in conversation,” *Metapragmatics of Humor: Current research trends*, 14, 235, 2021.

[35] L. Heron, J. Kim, M. Lee, K. El Haddad, S. Dupont, T. Dutoit, K. Truong. ”Dyadic Conversation Dataset on Moral Emotions,” in *13th IEEE Int. Conf. on Automatic Face Gesture Recognition (FG 2018)*, pp. 687-691.

[36] J. H. Cheong, T. Xie, S. Byrne and L. J. Chang, ”Py-feat: Python facial expression analysis toolbox,” arXiv preprint arXiv:2104.03509, 2021.

[37] Y. Fan, J. Lam, and V. Li, ”Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution”, in *Proc. of the AAAI Conf. on Artificial Intelligence*, (2020, April), (Vol. 34, No. 07, pp. 12701-12708).

[38] Q. N. Ruan, J. Liang, J. Y. Hong and W. J. Yan, ”Focusing on Mouth Movement to Improve Genuine Smile Recognition,” *Frontiers in psychology*, 11, 1126, 2020.

[39] R. B Cialdini and L. James, ”Influence: Science and practice,” Vol. 4. Pearson education Boston, 2009.

[40] ELAN (Version 6.3) [Computer software]. (2022). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

[41] P. Virtanen et al.,”SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” *Nature Methods*, 17(3), 2020, 261-272.

[42] C. C. Gramazio, D. H. Laidlaw and K. B. Schloss, ”Colorgorical: Creating discriminable and preferable color palettes for information visualization,” in *IEEE Trans. on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 521-530, Jan. 2017.

[43] J.R. de Leeuw, ”jsPsych: A JavaScript library for creating behavioral experiments in a Web browser,” *Behav Res* 47, 1–12 (2015).

[44] Davis E. King. 2009. ”Dlib-ml: A Machine Learning Toolkit”. J. Mach. Learn. Res. 10 (12/1/2009), 1755–1758.