

Automatic dataset generation for recognition of subtle facial expression changes for smiling faces

Kei SHIMONISHI¹[0000-0002-2413-8237], Kazuaki KONDO¹, Hirotada UEDA¹,
and Yuichi NAKAMURA¹[0000-0001-5103-7194]

Academic Center for Computing and Media Studies, Kyoto University,
Sakyo, Kyoto, Japan shimonishi@i.kyoto-u.ac.jp

Abstract. Recognizing the transitions of intensity of facial expressions in daily life is crucial to monitoring the individual’s mental state. We have investigated a two-input facial expression recognizer specialized for individuals to evaluate facial expressions in detail with personalized scales. Although this recognizer can accurately recognize facial expressions based on comparisons, it has the limitation of requiring the manual provision of supervised data for each individual to train the recognizer. The goal of this study is to automatically generate the datasets required for training the two-input facial expression recognizer for smiles from videos that capture an individual’s facial changes. To achieve this goal, we propose a method for extracting facial image pairs with explicit expression differences using a conventional facial expression recognizer. Additionally, we introduce a three-input facial expression change recognizer to extract facial image pairs with subtle expression differences. Experimental results showed that our proposed framework could successfully generate a dataset for training the two-input facial expression recognizer.

Keywords: Automatic dataset generation · Facial expression · smile intensity.

1 Introduction

Facial expressions and their temporal changes provide valuable insights into a person’s internal state, making them a foundational technique for state monitoring. For example, measuring facial expressions over time is particularly useful for individuals such as dementia patients, who may have difficulty recalling or accurately expressing their conditions.

While a much amount of research has been conducted on facial expression recognition and has developed generalized classifiers for broad applicability, most studies focus on significant expressions, aiming to identify which facial expression is being represented. To capture changes in facial expression intensity in daily life, however, it is necessary to recognize not only significant expressions but also intermediate or subtle ones, such as a slight smile. Capturing such subtle facial expression changes has been reported to result in low accuracy and sensitivity, even with current state-of-the-art facial expression recognizers [11]. Moreover,

facial expression patterns vary between individuals, and specifically, in cases of mental disorders such as depression, expressions tend not to appear. Therefore, it is essential to develop methods that can detect subtle changes and evaluate expressions on a personalized scale.

To address these challenges, our research group has developed a two-input facial expression recognizer [6] for evaluating facial expressions, which adapt to individual characteristics using a comparative approach. Although this recognizer achieved high recognition accuracy for changes in the facial expressions of specific individuals, this method has practical limitations that require collecting extensive facial expression samples from specific individuals for training.

In this study, we introduce a novel framework that automatically generates the necessary training data for our previously proposed two-input facial expression recognizer from videos capturing an individual’s facial expressions. This framework utilizes existing facial expression recognition models to detect distinct expressions and create pairs of images: one with a pronounced expression and another with a less pronounced expression. Additionally, we introduce a novel three-input facial expression change recognition network to identify small differences in expression intensity and augment the dataset with such subtle variations.

2 Related Work

2.1 Facial Expression Recognition Using FACS

Facial expression recognition methods have traditionally focused on features defined by facial muscle movements, such as in the Facial Action Coding System (FACS) [4, 7]. These methods mainly classify expressions with prominently expressed emotions, making it challenging to apply them to general expressions that appear in everyday life. In our daily lives, not only prominent expressions but also subtle or ambiguous expressions are frequently represented. When facial expressions are subtle, changes caused by factors other than expression, such as head orientation and lighting, become dominant, making it difficult to detect landmark transitions solely due to expression changes. In fact, Hamm et al. [5], revealed that the detection error of facial landmarks used in FACS becomes significant when identifying subtle expressions in schizophrenia patients.

2.2 Facial Expression Recognition Using CNNs

Convolutional neural networks (CNNs) have been increasingly used for facial expression recognition and have demonstrated superior performance compared to landmark-based recognition methods. For example, Neagoe et al. trained a CNN to recognize seven categories of emotions (six basic emotions plus neutral) [8]. Even with a small dataset of 210 facial expression images from Japanese women, their CNN model, which consisted of three convolutional layers and three pooling layers, achieved a high recognition accuracy.

While these studies demonstrate the effectiveness of CNNs for facial expression recognition, the expressions studied are mostly prominent, and little research has examined subtle expressions in daily life, which is the focus of our purpose. In fact, as described in Section 1, accuracy and sensitivity are still insufficient for subtle facial expressions even with state-of-the-art facial expression recognition methods.

2.3 Comparison-Based Facial Expression Recognition

To capture subtle expression changes, our research group has investigated a two-input facial expression recognizer to estimate the relative intensity of expressions within an individual by comparing two expressions of the individual [6]. Using this recognizer, if we have multiple typical facial images of an individual, the intensity of a facial expression can be evaluated within a personalized scale by comparing the facial image with a group of reference images [9]. However, as noted in Section 1, this approach requires manually annotated data for each individual for training. In particular, annotations for small changes in the training data are necessary to capture subtle facial expression changes, and that causes limitations for practical uses due to the considerable effort involved.

3 Automatic training for facial expression recognition network

This study aims to automatically train a recognizer capable of capturing facial expression changes in everyday life. The proposed framework is a fully automatic learning of a two-input facial expression recognizer developed by our group, which will be described in detail below. The recognizer is trained to recognize the differences in expressions of a specific individual. In this paper, we specifically focus on recognizing smiles, which are mainly related to the feeling of pleasure.

3.1 Comparison-based facial expression recognition network

It is difficult to give absolute values for the intensity of facial expressions even by hand. In contrast, relatively evaluating which of two images represents more smiles by comparing two images is much easier. Based on this concept, our group has developed a two-input facial expression recognizer [6].

Fig.1 shows the overview of the network structure. This network is based on a Siamese network architecture [2], which acquires two inputs and returns one value. The two feature extraction layers share weights and extract features from their respective facial images. Those two extracted features are then concatenated and passed through the fully connected layer to determine the relative intensity of the expressions. We here employed the CNN component of VGG16 [10] to extract features from facial images.

Note that training data consists of two facial images and a label indicating which of the two images represents the expression with greater intensity (Fig.2).

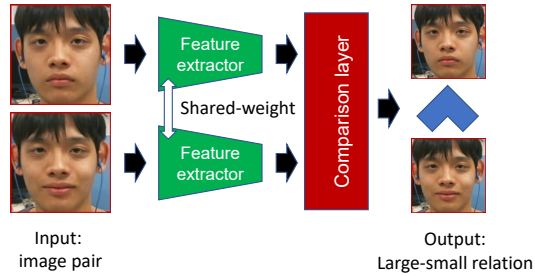


Fig. 1. A comparison-based smile intensity recognition network

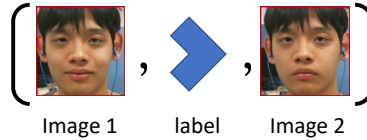


Fig. 2. Examples of paired data

In short, the question to answer in this paper is, “How can we automatically select pairs of facial images and assign appropriate labels indicating the relative intensity of facial expression between them, given a sequence of facial images of a specific individual?”

3.2 Framework of automatic training for facial expression recognition network

The specific steps of the automated training process are as follows and consist of the following three steps:

1. A sequence of facial images of an individual facing a camera is extracted using the BlazeFace [1], one of the face detectors, from a video capturing the individual’s facial expression changes.
2. A dataset consisting of paired data is created from the sequence of facial images using the proposed algorithm, which will be described in the next section.
3. The recognizer is trained by the created dataset.

4 Automatic dataset generation

As introduced in the previous section, the training data for our two-input facial expression recognizer consists of two facial images and a label indicating which represents the expression with greater intensity. These three elements are collectively referred to as “paired data” hereinafter.

Here, we assume that a sequence of facial expression data for a specific individual has been obtained from the video, and we introduce a method for generating paired data from the sequence of facial expression images. This paper specifically focuses on smiling as a target facial expression.

The proposed algorithm for automatically generating a dataset consists of three main methods:

Baseline Method Significant smiles and neutral expressions are identified using a conventional single-input facial expression recognizer that can detect prominent facial expressions. Then, the image pair is assigned a label, “smile > neutral,” and these images and the corresponding label construct one paired data.

Beam Method Similar to the baseline method, a conventional single-input facial expression recognizer is used to detect significant smiles and a variety of other expressions besides smiles. For each image pair, a label of “smiling > otherwise” is assigned, and the paired data is constructed along with the corresponding label.

Triplet Method We introduce a novel three-input facial expression change recognizer to recognize subtle differences in facial expressions. This recognizer identifies the consistency of expression change directions and generates paired data containing subtle differences in expressions.

These methods are based on the assumption that conventional facial expression recognizers can successfully detect prominent facial expressions. Furthermore, these three methods are combined and utilized, given the necessity of diverse training data and the inclusion of subtle differences in facial expressions. Each method is described in detail in the following sections.

4.1 Paired Data Generation Using the Baseline Method

For the facial images extracted from a video, the smile intensity of each image is estimated using a conventional single-input smile recognizer, Smilenet [3]. Smilenet is a network that outputs whether a facial image represents a smile (1) or a neutral expression (0) using activation from the sigmoid function in the final layer. In this study, the pre-activation value is regarded as the smile intensity estimated by the network.

As mentioned in Section 1, recognizers trained for binary classification have limitations in reliably estimating smile intensity for intermediate expressions. However, expressions with strong or weak intensity are likely significant smiles or neutral expressions similar to those in the training data. In other words, by setting an appropriate threshold for the estimated smile intensity and ignoring facial images within the threshold, i.e., estimated as intermediate expressions, we expect the smile intensity to be somewhat trusted.

Fig.3 shows an example of smile intensity transitions estimated by Smilenet. Here, we set appropriate thresholds for the smile intensity values and classified the facial images into three categories: “smile,” “neutral,” and “unknown.” Based

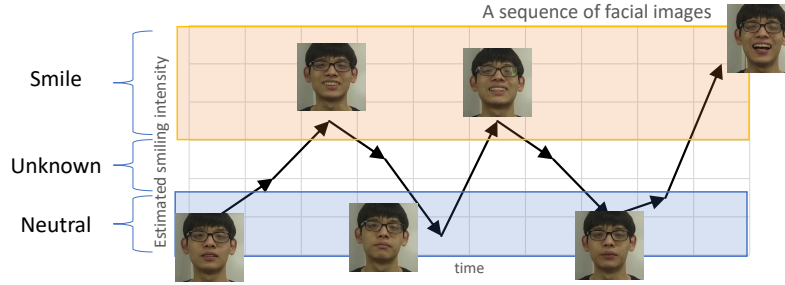


Fig. 3. An example of the transition of smiling intensity estimated by the Smilenet

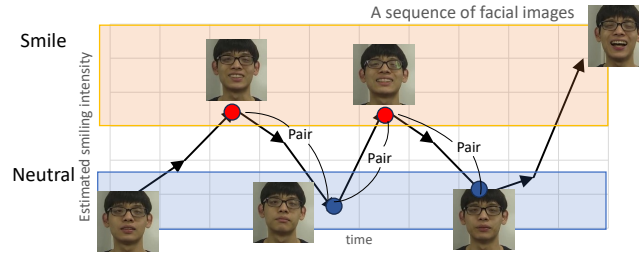


Fig. 4. Paired data generated by the Baseline method

on this classification and the estimation results, a time segment in which the state transitions sequentially from neutral to unknown and then to a smile, or conversely, from a smile to unknown and then to neutral, is defined as a slope. If the transition involves multiple cycles, such as (smile, unknown, smile, unknown, neutral), the minimum segment, i.e., (smile, unknown, neutral) is considered the slope.

From these slopes, image pairs are extracted from the smile-side and neutral-side peaks within each slope as shown in Fig.4 Then, the image pair is assigned a label, “smile > neutral.”

4.2 Paired Data Generation Using the Beam Method

The Baseline method allows for creating paired data consisting of smiles and neutral expressions. However, this method is limited in that the number of detected slopes constrains the number of generated pairs. Additionally, while Smilenet can achieve a certain level of accuracy in determining whether an input facial image represents a significant smile, it frequently produces false positives when identifying neutral expressions. This may be because the network is trained primarily to recognize smiles, with less emphasis on distinguishing neutral expressions.

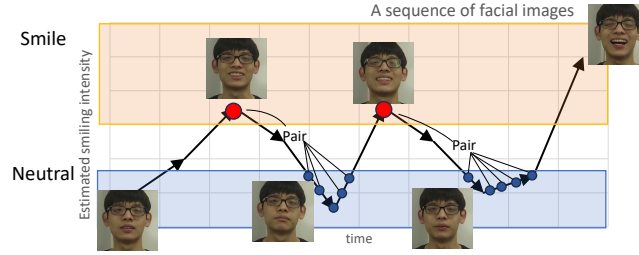


Fig. 5. Paired data generated by the Beam method

To tackle these limitations, we introduce the Beam method (see Fig.5). In this approach, for each slope, the peak facial image on the smile side (indicated by the red circle in the Fig.5) is paired with a group of temporally adjacent facial images that are estimated to be neutral expressions. These pairs are labeled “smile > otherwise.” This allows us to make a large amount of paired data with a significant smile and multiple images that are considered to be close to a neutral face.

4.3 Paired Data Generation Using the Triplet Method

Two proposed methods described in the previous sections enable the creation of paired data consisting of significant smiles and neutral expressions. In addition, to account for capturing subtle changes in facial expressions in daily life, the training data should include image pairs with such subtle differences in intensity. However, intensity values estimated by Smilenet are unreliable, especially for expressions of intermediate intensity, as discussed above. Therefore, using such values to generate paired data that includes subtle facial expression differences is undesirable.

To address this issue, we introduce a new network that recognizes the consistency of directional changes between three input facial images. Suppose the three images are arranged in ascending or descending order regarding facial expression intensity; the two magnitude relations are in agreement when compared two by two, as shown in Fig.6 (a). Conversely, suppose the order of intensities in the three images is inconsistent, i.e., the middle image has a stronger or weaker smile than the images at either end; the two magnitude relations do not match as shown in Fig.6 (b). The network makes it possible to determine whether the smile intensity of a new facial image lies between the smile intensities of the two images in the existing paired data. Based on this idea, by repeatedly selecting facial images with intermediate smile intensities for the existing paired data, it becomes possible to gradually add pairs containing subtle changes as training data, starting from those with more significant changes.

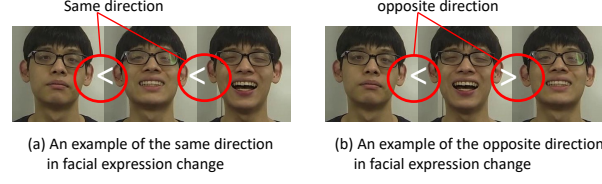


Fig. 6. Examples of the direction of facial expression change

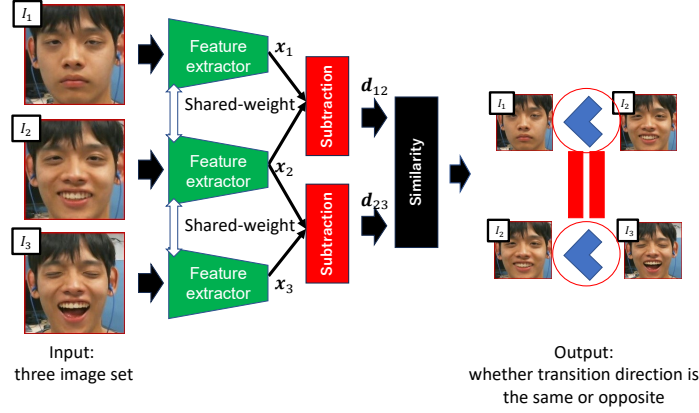


Fig. 7. A network structure of the three-input facial expression changes recognition network

A three-input facial expression change recognition network Fig.7 shows the structure of the proposed three-input facial expression change recognition network. At first, feature vectors ($\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$) are extracted from each facial image (I_1, I_2, I_3), respectively, using a shared weighted feature extraction layer as shown in the green block in the figure. Here, we employed the CNN component of VGG16 for the feature extraction layers in the same way as the proposed two-input facial expression recognizer. Then, the feature differences $\mathbf{d}_{12} = \mathbf{x}_1 - \mathbf{x}_2$ and $\mathbf{d}_{23} = \mathbf{x}_2 - \mathbf{x}_3$ for each pair of images are calculated. Finally, the cosine similarity between the two calculated feature difference vectors $\mathbf{d}_{12}, \mathbf{d}_{23}$ is calculated. With this similarity, the model is trained using the MSE loss to output 1 if the facial expression changes in (I_1, I_2) and in (I_2, I_3) are in the same direction (Fig.6 (a)) and -1 if they are in opposite directions (Fig.6 (b)). This training approach is expected to enable the model to learn feature extraction capabilities for recognizing individual facial expression changes.

Training the three-input facial expression change recognition network The newly introduced three-input facial expression change recognition network should also be trained automatically to achieve a fully automated generation of

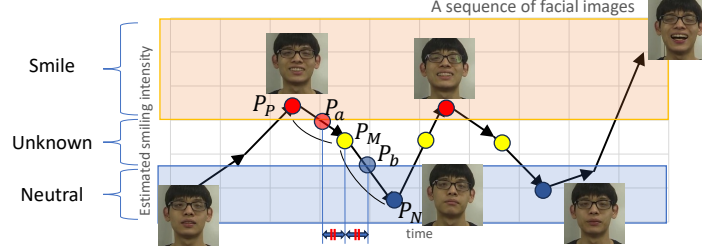


Fig. 8. Training data for the three-input facial expression changes recognition network

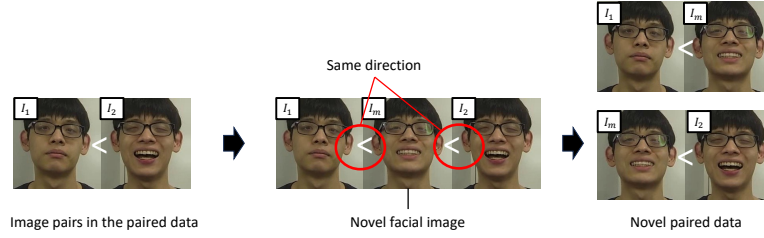


Fig. 9. A concept of adding novel paired data by triplet network

training data for the two-input facial expression recognizer. The basic idea to realize this is similar to the Baseline method: One image is selected from each of {smile, unknown, neutral} estimated by Smilinet and used as a triplet. Although the Smilinet estimates themselves are considered to be unreliable, the smile intensity in the face images evaluated as neither smiling nor neutral expressions is expected to come somewhere between the images predicted as smiling and the images predicted as neutral. This is because these images are likely to reflect a level of expression intensity that is neither as strong as those predicted as smiles nor as minimal as those predicted as neutral.

In particular, facial images are selected for each of the slopes detected by Smilinet, as follows: We select the smile peak image P_P , the neutral peak image P_N , along with an image P_a near the smile-side threshold and an image P_b near the neutral-side threshold as shown in Fig.8. Since all facial images between P_a and P_b lie within the “unknown” images, the smile intensity of the facial image at temporal midpoint P_M between P_a, P_b is expected to fall between those of P_P and P_N . Three-image triplets (P_P, P_M, P_N) from the descending slope and (P_N, P_M, P_P) from the ascending slope for each slope are extracted to construct training data for the three-input expression change recognition network.

Paired Data Generation Using the Three-Input Facial Expression Change Recognition Network The basic concept of paired data generation using the three-input expression change recognition network is as previously described. For

each existing paired data, the network evaluates whether the smile intensity of a facial image obtained from the same slope as the paired data lies in between the intensities of the two images, as illustrated in Fig.9. If this relationship holds, the new facial image, the two images in the paired data, and their corresponding labels are used to generate new paired data.

This process is repeated until either the frame interval between the paired data becomes less than a certain threshold or an image with a direction of change inconsistent with the pair is encountered. Through this iterative procedure, paired data incorporating finer differences in facial expressions can be generated.

5 Experiment

We conducted an experiment to confirm that the proposed methods can successfully generate appropriate training datasets for the two-input facial expression recognizer. In this experiment, we compared the recognition accuracy of the recognizer between two training conditions: A manually prepared dataset and an automatically generated dataset using the proposed method. Additionally, we confirmed that combining the three methods described in the previous section improved recognition accuracy.

5.1 Dataset

Two video categories, A and B, were prepared, with three videos in each category: A1, A2, and A3 in category A and B1, B2, and B3 in category B. The first category (Category A) consisted of videos captured participants playing a game alone, and contained multiple smiling with a significant intensity. In contrast, the second category (Category B) consisted of videos captured participants watching various short clips, during which their facial expressions were recorded. The facial expression changes in this category were generally subtle.

Each video lasted approximately 20 to 30 minutes. We manually annotated paired data from the sequence of facial images obtained from each video by BlazeFace. In this experiment, we manually annotated the expression change slopes and used the images at the start and end of the slope and the labels for the two points as paired data.

Adding the Beam method to the Baseline method is expected that the recognition accuracy with large expression differences (i.e., Category A) will improve. Furthermore, adding the Triplet method is expected to improve recognition accuracy with small expression differences (i.e., Category B).

5.2 Recognition Accuracy Evaluation for Expression Changes

The results for each case are shown in Tables 1 and 2. Because the data for validation consisted of manually annotated paired data, 5-fold cross-validation was applied for accuracy evaluation by manual annotation.

Table 1. Recognition accuracies for video category A

Dataset	A1	A2	A3
Baseline	78.6	87.0	72.2
Baseline + Beam	85.7	93.5	90.6
Baseline + Triplet	83.9	93.5	86.7
All	90.2	93.5	89.8
Manual annotation	94.6	95.2	98.4

Table 2. Recognition accuracies for video category B

Dataset	B1	B2	B3
Baseline	61.7	69.2	69.5
Baseline + Beam	69.7	72.8	79.7
Baseline + Triplet	76.7	82.9	75.7
All	82.4	84.1	81.4
Manual annotation	80.9	73.9	63.8

These results indicated that for the dataset containing only large expression differences (category A), the increase in the number of paired data generated by the Beam method mainly contributes to improving accuracy, as we expected. For the dataset that also includes smaller expression differences (category B), the addition of paired data with small differences using the Triplet method mainly contributes to improved accuracy for two of three datasets (videos B1 and B2). Regarding Video B3, the effect of the Beam method was more significant than that of data augmentation using the Triplet method. However, even for this data, further accuracy improvement was observed when the Triplet method was also applied in addition to the Beam method.

For the dataset containing only large expression differences, the training data automatically generated by the proposed method could not achieve the same level of recognition accuracy as the manually annotated data. This may be because the effect of false positives in the automatically generated dataset was more significant for datasets that can be annotated relatively easily by human operators. On the other hand, for datasets containing small expression differences, the proposed method achieved higher recognition accuracy than the manually annotated results. This suggests that human annotation is limited to datasets containing small differences, and the ability to acquire subtle expression changes through automatic recognition is expected to be useful.

6 Conclusion

In this paper, we proposed a method for automatically generating a dataset to train a two-input facial expression recognizer that evaluates smiling intensity based on individual-specific comparisons. The proposed method applies the conventional network for recognizing salient facial expressions that can accurately

detect “significant smiles,” and automatically generate training data consisting of smiling and non-smiling faces. In addition, we introduced a new network that recognizes the identity of the direction of change in smiles and added a dataset that includes detailed changes in facial expressions.

The experimental results showed that the recognizer trained on the dataset automatically generated by the proposed method achieved high recognition accuracy, although it was not as good as the recognizer trained on the manually labeled dataset that consisted only of large facial expression differences. Furthermore, when evaluated on a dataset containing subtle facial expression differences, the automatically trained recognizer outperformed the manually trained one.

Future works include investigating the robustness of the proposed method under various conditions within videos and exploring possible extensions to enable automatic training of facial expression recognizers from more general video content.

References

1. Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., Grundmann, M.: Blazeface: Sub-millisecond neural face detection on mobile gpus (2019)
2. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature Verification using a “Siamese” Time Delay Neural Network. In: *Advances in Neural Information Processing Systems*. vol. 6. Morgan-Kaufmann (1993)
3. Chen, C., Tokuhara, K., Arakawa, Y., Watanabe, K., Ishimaru, S.: Quantifying a multi-person meeting based on multi-modal micro-behavior analysis. In: *Proc. of JSAI2022* (2022)
4. EKMAN, P.: Facial action coding system (facs). *A Human Face* (2002)
5. Hamm, J., Kohler, C.G., Gur, R.C., Verma, R.: Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of Neuroscience Methods* **200**(2), 237 – 256 (2011)
6. Kondo, K., Nakamura, T., Nakamura, Y., Satoh, S.: Siamese-structure deep neural network recognizing changes in facial expression according to the degree of smiling. In: *Proc. of ICPR2020*. pp. 4605–4612 (01 2021)
7. Lien, J.J., Kanade, T., Cohn, J.F., Ching-Chung Li: Automated facial expression recognition based on facs action units. In: *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. pp. 390–395 (1998)
8. Neagoe, V.E., Barar, A., Sebe, N., ROBITU, P.: A deep learning approach for subject independent emotion recognition from facial expressions (2013)
9. Shimonishi, K., Kondo, K., Ueda, H., Nakamura, Y.: Ordinal scale evaluation of smiling intensity using comparison-based network. In: *Machine Learning for Cognitive and Mental Health Workshop (ML4CMH), AAAI2024, Vancouver, BC, Canada*. pp. 64–73 (2024)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *Proc. of ICLR 2015, San Diego, CA, USA, May 7-9* (2015)
11. Zhang, J., Shimonishi, K., Ueda, H., Kondo, K., Nakamura, Y.: Evaluating subtle positive-negative facial expression transitions for monitoring changes in personal internal states. In: *PRICAI 2024: Trends in Artificial Intelligence*. pp. 392–404. Springer Nature Singapore, Singapore (2025)