Ordinal Scale Evaluation of Smiling Intensity using Comparison-Based Network

Kei Shimonishi[1,*], Kazuaki Kondo[1], Hirotada Ueda[1] and Yuichi Nakamura[1]

[1]*Kyoto University, Yoshida-honmachi, Sakyo, Kyoto, Japan*

Abstract

The ability to evaluate both explicit facial expressions and intermediate expressions is helpful for human monitoring. Since intermediate facial expression is out of the scope of traditional studies, evaluation scores obtained from traditional facial expression recognition techniques are unreliable. In this paper, we propose an ordinal scale-based evaluation scheme for facial expression based on a comparison. The proposed framework is based on an ordinal scale; it is challenging to construct a standard scale that can be applied to multiple individuals. However, it is expected to be effective enough to track changes in the facial expressions of the specific individual, including intermediate expressions. We also propose an algorithm for selecting reference images from the data by taking into account the consistencies of the strong-weak relationships between reference images because the reference image selection significantly impacts the ordinal evaluation. Our approach is evaluated by conducting experiments with human annotators.

Keywords

Facial expression recognition, Siamese network, ranking, ordinal scales

1. Introduction

Monitoring an individual's Quality of Life (QOL) is becoming increasingly important to maintain good mental conditions and detect early trends in harmful conditions. Because direct QOL inquires are bothering and it is difficult to accurately represent one's internal state, estimating internal state from external nonverbal information is desired. Facial expression is one of the modalities that reflects an individual's internal state and is expressed with being influenced by mental condition. For example, when an individual is not feeling well, the same smile may appear weaker than usual. Therefore, monitoring facial expressions in daily life is a crucial clue to estimating an individual's QOL.

The research field of facial expression recognition (FER) has a long history, and it has already been put into practical use as a technology, such as smiling shutters. While traditional FER mainly focuses on recognizing whether a clear facial expression is represented or not, from the viewpoint of monitoring in daily life, evaluating the degree of expression for the individual is rather crucial, especially for patients with dementia who have little or no facial expressions. Based on this point of view, this research aims to draw a curve of transitions of the individual's degree of facial expressions, particularly smiling intensity, as shown in Figure 1.

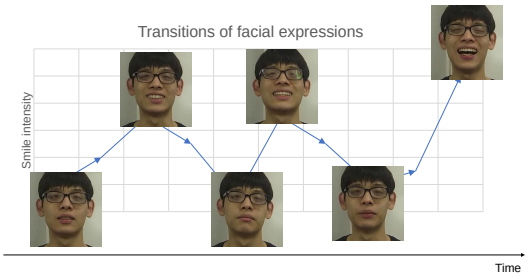


Figure 1: An example of transition curve of smiling intensity in daily life

Though the traditional algorithm of FER seems able to evaluate intermediate facial expressions as a probability that a specific facial expression is represented, the probability values are not so reliable, especially for evaluating intermediate expressions. This is because the intermediate facial expression was out of the scope of the traditional studies; learning is likely to output a value close to the binary value of either no expression (0) or an expression (1). As a result, for example, when the degree of smile expression is estimated for a series of facial expressions, the value may change abruptly over a series of times as shown in Figure 2. In addition, it is also difficult for the machine learning algorithm to directly learn the intermediate facial expressions since it is difficult for even humans to give appropriate absolute values for intermediate facial expressions.

Kondo et al. [1] proposed a network for recognizing smiling based on "comparison" to address the issues of recognizing intermediate facial expressions. Their work

Machine Learning for Cognitive and Mental Health Workshop (ML4CMH), AAAI 2024, Vancouver, BC, Canada

*Corresponding author.

shimonishi@i.kyoto-u.ac.jp (K. Shimonishi);
kondo@ccm.media.kyoto-u.ac.jp (K. Kondo);
ueda.hirotada.2r@kyoto-u.ac.jp (H. Ueda);
yuichi@media.kyoto-u.ac.jp (Y. Nakamura)

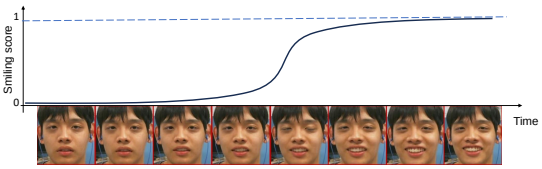


Figure 2: An example of sudden jump of evaluation scores for intermediate facial expressions by traditional facial expression recognition technique

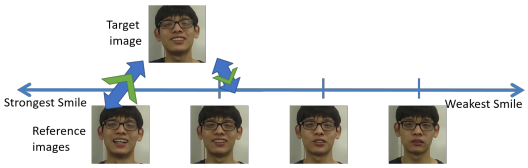


Figure 3: Overview of evaluation method of facial expression intensity based on comparison

is based on the assumption that the problem of relatively evaluating which of two images represents more smiles by comparing two images is easier than absolutely evaluating a degree of smiling from only one image.

By borrowing this comparison-based idea to evaluate facial expressions, we propose an approach to evaluate smiling intensity with an ordinal scale. The basic idea of this approach is that if we have multiple reference face images for a specific individual and a method for comparing facial expressions, we can evaluate the smiling intensity of a new image of the individual through pairwise comparison with the reference images, as shown in Figure 3.

Since the expression ratings in this method are based on an ordinal scale, the degree of each rating is not the same for multiple individuals. However, this ordinal scale-based approach may satisfy our need to capture changes in facial expressions for each individual.

In addition, reference image selection is crucial for this ordinal-based evaluation because they are considered an evaluation space for facial expressions. Therefore, we also propose an algorithm of reference image selection from a large number of face image data of each individual based on consistencies of comparison results within images.

In summary, the contributions of this paper are as follows:

- We propose an approach to evaluating intermediate smiling intensity by ordinal scales based on comparisons.
- We propose an algorithm for selecting appropriate reference images to construct appropriate evaluation space.

We briefly introduce related work in the next section. Then, we introduce an approach to evaluate facial expressions by ordinal scales and an algorithm of reference image selection. We evaluate our approach and algorithm with human annotators, and finally, we conclude our research.

2. Related Work

2.1. Facial expression recognition

Facial expression recognition is widely utilized in several fields. Traditional studies mainly focused on determining whether a specific expression is represented or not.

2.1.1. Facial Action Coding Systems

Facial Action Coding Systems (FACS) [2] is a framework proposed by Ekman et al. that classifies a face into several parts (Action Units; AUs) based on the basic action units of individual muscles and describes facial expressions as a combination of these AU actions. Many facial expression recognition applications have used FACS as features, and for example, OpenFace [3] can analyze multiple facial expressions in near real-time by automatically recognizing the actions of AUs.

2.1.2. Deep neural network based approach

Although the FACS-based FER approach has been successful, it has the limitation that the final results are affected by the accuracy of FACS detection. This limitation can become a problem, especially when trying to capture subtle differences in facial expressions because the effect of observation noise cannot be ignored. On the other hand, the end-to-end approach by the deep neural network can be expected to reduce the effect of such observation noise by eliminating the necessity of explicit feature detection. For example, VGGNet [4] is a traditional deep neural network, but it is known that human facial features can be extracted well, and recent research of FER also utilized VGGNet [5, 6]

2.2. Siamese structure-based recognition technique

Siamese network [7] is one of the deep neural networks of metric learning. This network acquires two inputs and returns the distance between the two inputs. By applying the same structures and the same weights to feature extraction layers of these two inputs and the distance of the two inputs to the loss function, the network can learn the distance space. The Siamese Network is a network that determines whether two inputs are similar or different and has been applied to handwritten signature

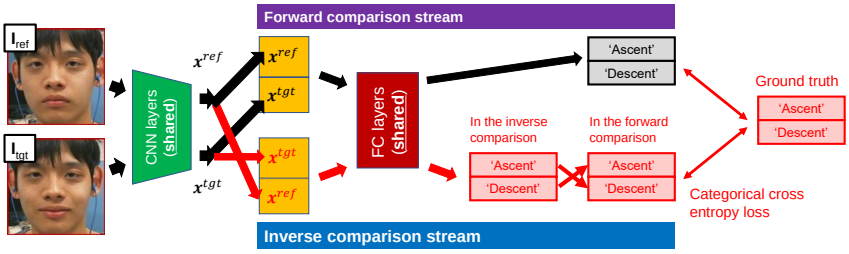


Figure 4: Siamese-based network to compare face images to evaluate the degree of smiling

recognition [8] and used as a framework for anomaly detection [9]. As one of its features, it is known as a network that can be trained from a small number of training data compared to conventional networks that perform multi-valued discrimination and regression [10].

Kondo et al. [1] proposed an approach to the evaluation of facial expressions based on comparison inspired by the Siamese structure. Their approach compares two facial images and returns which of one image represents more smiles, and they showed that the approach has the potential to distinguish subtle facial expression differences. In addition, Zhang et al. [11] extended their work from a positive-neutral direction to a negative-neutral direction.

3. Comparison-based smiling evaluation by ordinal scales

3.1. Overview of the proposed framework

As introduced in the Introduction, the basic idea of our approach is a comparison-based evaluation. Kondo et al.[1] has developed a Siamese-based smiling recognition network that takes two face images as input and recognizes which one is expressing smiling more. By borrowing this idea, once we develop a network that can determine which of two images represents more smiles, and if we have multiple reference images, we can evaluate the smile intensity of a new image through pairwise comparison with the reference images as also introduced in the Introduction. When it comes to determining smiling intensity based on ordinal scales, although all the comparison results are ideally consistent, the results are sometimes inconsistent due to an ambiguity of slightly different face images. Therefore, we apply a voting-based evaluation and determine smiling scores by merging multiple comparison results. In addition, we propose an algorithm to select appropriate reference images to reduce the ambiguity between reference images in the following section.

3.2. A network for facial expression comparison

In this paper, we defined the recognition task as a simple two-category classification problem (i.e., determining which of two input images represents the greater degree of smiling) and construct a Siamese-based network to recognize smiling similar to the network Kondo et al. have developed [1].

Figure 4 shows the structure of the proposed network that accepts two input images and returns two likelihood values corresponding to ascension and descension labels relative to the degree of smiling. We employed the CNN component of VGG16 [4] and two fully connected layers with rectified linear units, a 0.25 dropout rate, and SoftMax in the proposed method. The ground-truth likelihood values for an input image pair were represented as a two-element one-hot vector, with its element corresponding to the ground truth label set to 1 and the other element set to 0, respectively. We used categorical cross-entropy loss to optimize the network parameters, as follows:

$$L_{cat} = -\sum_{i} \{y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)\}\,, \tag{1}$$

where $i = \{0, 1\}$, y_i, and $\hat{y}_i$ denotes ascension and descension labels relative to the degree of smiling, the ground-truth label, and the predicted likelihood values, respectively.

The previously proposed network by Kondo et al. was not designed to consider the order of inputs, resulting in instances where swapping the order of two inputs led to contradictory outputs. To address this issue, we input a permuted version of the two features extracted from two input images by the CNN component into the fully connected layer in the latter stage and calculate the categorical cross-entropy loss of inverted input, L_{inv}, as same as L_{cat}, as shown in red arrows in Figure 4. Also, a loss of consistency of these two types of input is calculated as

$$L_{con} = 1 - \{min(P_f^{As}(tgt, ref), P_i^{As}(tgt, ref))$$

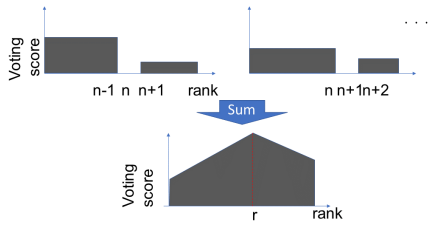


Figure 5: Voting-based evaluation

$$+min(P_f^{Des}(tgt, ref), P_i^{Des}(tgt, ref))\}, \quad (2)$$

where $P^{As}(n, m)$ and $P^{Des}(n, m)$ represent probabilities that the degree of smiling of image I_n is larger or smaller than that of image I_m, respectively. In other words, $P(sn(I_n) > sn(I_m))$ and $P(sn(I_n) < sn(I_m))$, where $sn(I)$ represents a degree of smiling of image I. Also, P_f and P_i represent the likelihoods of the forward comparison stream and inverse comparison stream, respectively.

In total, our network is trained to decrease the following loss function:

$$L = L_{cat} + L_{inv} + L_{con}. \quad (3)$$

Here, we expected that the CNN and the fully connected components would be trained to compare extracted features in order to project the results onto the likelihood values of the ascension and descension labels, respectively.

3.3. Voting-based evaluation

Since the reference images may include some ambiguity between neighboring images, it is difficult to directly determine the degree of smiling of the new target image in a reference image set. Therefore, we apply a voting technique to determine the final rank of the image. The algorithm votes to possible ranks using the result of each comparison of reference images and a target image. As a result, the most likely rank should have a maximum number of votes.

In particular, the procedure is as follows. Suppose that we have N reference images with its order of degree of smiling, i.e., $sn(I_i) > sn(I_j), \forall i < j$. A new target image I_{new} is compared to all reference images, and likelihood values that the degree of smiling of target image is larger than that of a reference image $P^{As}(new, n)$ and likelihood that the degree of smiling of target image is lower than that of a reference image $P^{Des}(new, n)$ for all reference images ($n \in \{1, \ldots, N\}$) are obtained. Because if $sn(I_{new}) < sn(I_n)$ is estimated, the smile rank of I_{new} is estimated as larger than n, large values are voted to ranks larger than n. In practice, add likelihood values of "ascend" and "descend" to the ranks lower than and higher than n, respectively.

Simply thinking, the degree of smiling in the reference images can be determined by searching the position whose scores are maximum. Here, the position r can be derived as:

$$r = \arg\max_r \left\{ \sum_{n=1}^{r-1} P^{Des}(new, n) + \sum_{n=r+1}^{N} P^{As}(new, n) \right\}. \quad (4)$$

In addition, we here apply a mean-shift algorithm to determine the evaluation score based on these probability values.

4. Reference image selection

To apply voting-based ordinal-scale evaluation as described above, we first need to construct an evaluation space with several reference images. Since the proposed approach utilizes ordinal scales, the construction of the evaluation space is crucial for the capability of the approach. Although a straightforward way is to utilize all the face data as reference images, the evaluation space constructed by very similar or subtle different images is unreliable due to the ambiguity of these images.

In this paper, we first consider all the data as baseline images and take pair-wise comparisons to sort all the data in a dataset and construct a baseline ranking. Then, we select several images from the baseline ranking as reference images and quantize the evaluation space by taking into account consistency to address the issues due to ambiguity.

4.1. Baseline ranking construction

Figure 6 (a) shows a comparison table of the result of all pair-wise comparisons in baseline images. Each color shows a probability of how a target image has a stronger smile than a reference image, i.e., $P^{As}(tgt, ref)$. The blue shows a pair whose target image has a stronger smile than the reference image, i.e., $P^{As}(tgt, ref) > P^{Des}(tgt, ref)$. In contrast, the red area shows a pair whose reference image has a stronger smile than a target image, i.e., $P^{As}(tgt, ref) < P^{Des}(tgt, ref)$. The white area represents that the target and the reference images represent similar facial expressions.

By sorting the baseline images based on the sum of the probability values in each column of this table, a baseline ranking considering the consistency of the strong-weak relationship can be constructed (Figure 6 (b)). In particular, suppose we have N baseline images $\{I_1, \ldots, I_N\}$ in total, and denote images sorted in descending order by smiling intensity as $\{I_1^{rank}, \ldots, I_N^{rank}\}$. Since the

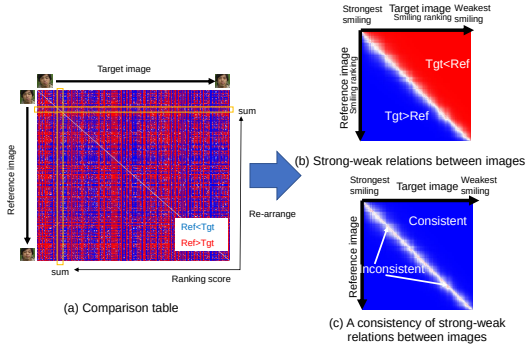


Figure 6: A baseline ranking made from a comparison table

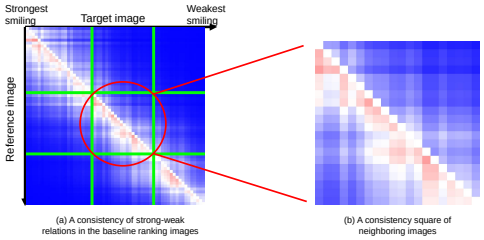


Figure 7: Consistencies of large small relationships in neighbor images as a part of consistency table in baseline ranking.

strong-weak relationship between each image I_n to other images $I_{\hat{n}}, \hat{n} \in \{1, \ldots, N\}_n$ are calculated as probability values P^{As} and P^{Des}, the total consistency values in the baseline ranking is derived as

$$L = \sum_{n}^{N} \left\{ \sum_{\{\hat{n}|I_{\hat{n}} \in \{I_1^{rank}, \ldots, I_{n-1}^{rank}\}\}} P_{Des}(n, \hat{n}) + \sum_{\{\hat{n}|I_{\hat{n}} \in \{I_{n+1}^{rank}, \ldots, I_N^{rank}\}\}} P_{As}(n, \hat{n}) \right\} \quad (5)$$

By maximizing this total consistency, baseline ranking images $(I_1^{rank}, \ldots, I_N^{rank}) = \arg\max I_1^{rank}, \ldots, I_N^{rank} L$ can be obtained. From now on, the subscript n will be used to sort the images in descending order of smiling degree.

An example of the consistency of the strong-weak relations in this rearranged table is shown in Figure 6 (c) by replacing $P^{As}(tgt, ref)$ into $P^{Des}(tgt, ref)$ when $sn(I_{tgt}) < sn(I_{ref})$. We here denote probabilities $C_{tgt,ref}$ to indicate this consistency as follows:

$$C_{tgt,ref} = \begin{cases} P^{As}(tgt, ref) & \text{if } sn(tgt) > sn(ref), \\ P^{Des}(tgt, ref) & \text{if } sn(tgt) < sn(ref). \end{cases} \quad (6)$$

Ideally, all cells would be blue, i.e., consistency is nearly equal to 1. However, due to the ambiguity of comparison results for similar facial expressions, there is also ambiguity in the consistency between the neighboring baseline ranking. Therefore, reference image selection is important to construct appropriate evaluation space.

4.2. Reference image selection

An important factor in selecting reference images is the consistency of the strong-weak relationships within reference images. That is, when the consistency table is calculated the same as the bottom-right figure of Figure 6, the less red and white colors area is a better sign of reference image selection.

To realize that, we focus on a square region of neighbor images as shown in Figure 7, and call this square consistency square. Suppose the differences between images are significant, and a strong-weak relationship is evident in the images. In that case, the consistency values in the consistency square are also expected to be large, i.e., cells in the square become blue. In contrast, when images are similar, and therefore the difference between images is ambiguous, the consistency values in the consistency square become low, i.e., cells in the square become white and red.

The basic idea of building a consistent evaluation space is to quantize images with low consistency values in the consistency square into a single class. As a result, the ambiguity between these images becomes “don’t care” in the evaluation space, and the consistency of the evaluated value becomes significant. That is, it is good to select images where the total consistency values in the sum of consistency squares, as shown in the right of Figure 7, becomes low.

In addition, neighbor images in the baseline ranking should not be selected as reference images. In other words, to select good reference images, the evaluation space is better to be divided evenly. To realize that, select a group of reference images so that the sum of the area of consistency space becomes small. To sum up, it is better to choose a group of images for which both the sum of consistency values and the sum of areas within the consistency square is small. Here, selecting a group of images with high consistency values within the consistency square is equivalent to selecting a group with low inconsistency. In summary, a group of reference images should be selected to maximize the total values of inconsistencies in consistency squares divided by the sum of the areas of consistency squares.

In practice, a procedure of reference image selection is the following. Suppose there are N images in total, and we want to select one image as a reference image.

At first, baseline ranking is constructed as introduced in the previous subsection and obtains consistency values $C_{i,j}$ $(i, j < N)$ for all pairs in the ranking. When evaluation space is divided into two with image $I_m (m < N)$, the sum of inconsistency values of the two spaces divided by the sum of the area is calculated as:

$$D_m^{(2)} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m}(1 - C_{i,j}) + \sum_{i=m}^{N}\sum_{j=m}^{N}(1 - C_{i,j}}{m^2 + (N-m)^2} \tag{7}$$

By searching the position whose $D_m^{(2)}$ are maximum, the best reference image I_m that divides evaluation space into two can be obtained. Similar to this, by calculating the sum of inconsistency in the consistency squares divided by the sum of the area with different division numbers N_{ref} + 1, N_{ref} reference images can be obtained. When it comes to calculating these values, we apply a scheme of dynamic programming to reduce calculation costs.

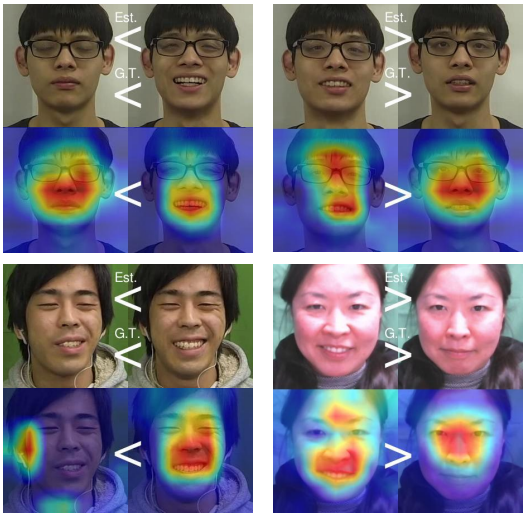


Figure 8: Examples of prediction results by comparison network. The first row shows label and prediction results, and the second row shows regions on which a network focuses by Grad-CAM.

5. Experiment

We conducted an experiment to evaluate the following things:

- How a proposed network can evaluate image pair.
- How appropriately reference images can be selected regarding consistency of both network and human annotators' evaluations.
- How the selected reference images evaluate face images.

5.1. Dataset construction

At first, the face image dataset is constructed by capturing participants' face images. We conducted two types of experiments to construct datasets with different situations. In the first type of dataset, we asked a participant to sit in front of the camera and to listen to funny radio. In the second type of dataset, we asked a participant to sit in front of the laptop PC and play a simple game. We captured facial images of these participants during the experiment. The second experiment was still experimental but closer to a natural scene than the first one. Each dataset was constructed only by one participant because our focus was to build a model to evaluate each individual. We collected two datasets of the first type, and one dataset of the second type.

Then, we added labels between image pairs that showed which of the two images expressed more smiles. The annotation between images with a slight difference in the degree of smiling was difficult, even for humans. It might cause a mistake in giving the correct labels. Therefore, we utilized image pairs with a clear difference as training data in this paper. In particular, we manually annotated segments that we thought the degree of smiling ascended or descended monotonically. We then picked each segment's start and end frame to construct one pair with its label. The number of image pairs of each dataset were 216, 174, and 123, respectively. Also, all face images in these pairs were utilized as baseline images. That is, the size of the dataset is twice the size of the image pairs; 432, 348, 246, respectively.

5.2. Evaluation scheme

The procedure of evaluation was constructed the following four steps; (1) the proposed network was trained for each individual by collected data and was evaluated by the cross-validation scheme; (2) we constructed the baseline ranking and evaluated the voting-based algorithm by determining the rank within the images in the baseline ranking; (3) reference images were selected from baseline images as proposed in the previous section and evaluated by human annotators on how they were consistent; (4) we confirmed the smiling intensity of face images in the evaluation space constructed by reference images.

As for training our network, we utilized pre-trained feature extraction layers of VGG-Face [12], which was trained on millions of face images for person identification, and trained only fully connected layers.

Regarding the evaluation of the voting-based algorithm, the rank of each baseline image was determined by the baseline ranking itself. In this evaluation, the grand truth of the rank of each baseline image was given as the

original rank of baseline ranking.

As for evaluating selected reference images, nine reference images were selected from the baseline images. Then, human annotators were asked to evaluate which of the two images represented more smiling for the pair of neighboring images in reference images. The images up to the third nearest neighbor images were considered a pair, and all image pairs were annotated twice by swapping the left and right sides of the comparison image. After comparing the image pairs within each dataset, the participants moved on to the next dataset. The order of evaluated image pairs was randomized for each dataset, but the order of the datasets was constant. We evaluated the consistency of the reference images by how accurate and consistent annotators evaluated the image pair. A group of 9 images regularly extracted every $N/10$ from the baseline images $\{I_1^{rank}, \ldots, I_9^{rank}\}$ was used as the reference image for comparison. Seven participants between the ages of 21 and 27 (6 male and 1 female) were recruited as annotators.

Figure 9: Selected reference images of dataset 1. More smile images are located on the left side.

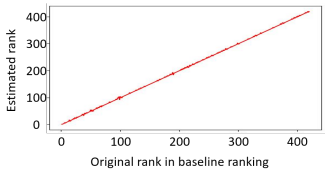


Figure 10: Consistency of estimated rank and original rank in baseline ranking

5.3. Results

5.3.1. Prediction accuracy

We first show the evaluation results of the trained comparison network in terms of accuracy. In this evaluation, five-fold cross-validation was applied, and prediction accuracies of three datasets were 99.5% $(215/216)$, 100% $(174/174)$, 98.3% $(121/123)$, respectively. Figure 8 shows examples of prediction results with Gradient-weighted Class Activation Mapping (Grad CAM) [13]. In each figure, the first row shows the grand truth label and estimation results, and the second row shows regions on which a network focuses for the prediction as a heat map. From these results, we can see that the network returns accurate prediction results by correctly focusing on face regions, including the mouth and eyes, which are well known as corresponding to smiling, even from the small dataset.

5.3.2. Consistency of voting based evaluation

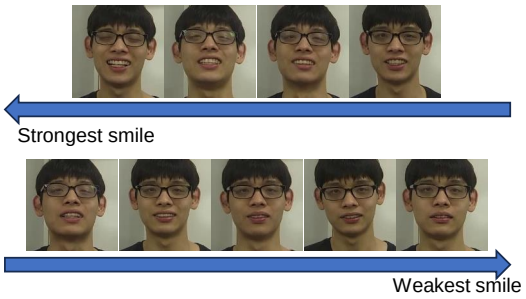
Figure 9 shows four examples of estimated ranks of images in the baseline ranking, rank 1, 100, 200, and 400 of dataset 1. The total number of baseline images was $216 \times 2 = 432$. Each image was estimated as rank 1, 103, 199, and 398, respectively, and an almost correct rank can be estimated by the voting algorithm. Figure 10 shows all pairs of estimated rank and grand truth label of this evaluation in dataset 1. These results show the consistency and effectiveness of the voting-based algorithm, as it predicted almost consistent values for all baseline images.

5.3.3. Selected reference images

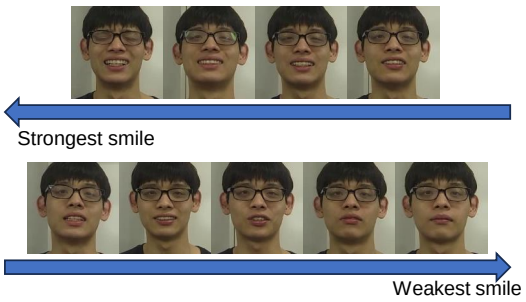
Figure 11 shows selected reference images by the proposed algorithm and equally picked up from the baseline ranking. The consistency table correlated to this result of dataset 1 is shown in Figure 12. In this figure, the green line shows where the algorithm divides baseline ranking. These results show that there still appears to be some ambiguity between adjacent images, even with the proposed approach. However, it appears to be reduced compared to a group of images acquired at regular intervals.

The consistency table calculated by these selected reference images is shown in Figure 13. We can see that almost all the cells represent blue. This result shows that ambiguities within reference images are small.

Figure 14 and Figure 15 show the quantitative evaluation results of selected reference images by annotators. In each figure, “proposed” and “baseline” represent the result for the images up to the third nearest neighbor reference images of the proposed algorithm and baseline algorithm, respectively, and “proposed_adjacent” and “baseline_adjacent” represent the result for the images only the nearest neighbor reference images. That is, the difficulty of the evaluation becomes hard. Figure 14 shows a prediction accuracy of evaluation results. Here, we consider the order given by the proposed network as the grand truth of the prediction. Therefore, this result also shows a correlation between network prediction and human perceptions. Figure 15 shows the consistency of each participant’s evaluation. In particular, it shows how much the same evaluation was given when the same image pair displayed with the left and right sides swapped. This high consistency indicates a low degree of ambiguity between image pairs. In almost all cases, reference



(a) Selected reference images by proposed algorithm



(b) Selected reference images by regularly picked up from baseline ranking

Figure 11: Selected reference images of dataset 1. More smile images are located on the left side.

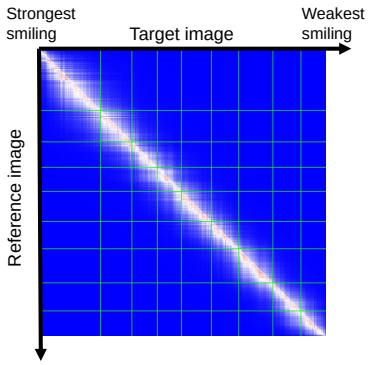


Figure 12: Consistency table of dataset 1. The green line shows where the algorithm divides baseline ranking.

images selected by the proposed algorithm obtain higher accuracies and higher consistencies. Since the smiles expressed in the experimental time were quantized into ten levels and the maximum value of the smile was not very high, both methods have a certain degree of similarity between the neighboring reference image pairs. Therefore, evaluation by humans may be somewhat difficult even with the proposed method. However, even in such a situation, we can confirm that the proposed method selects

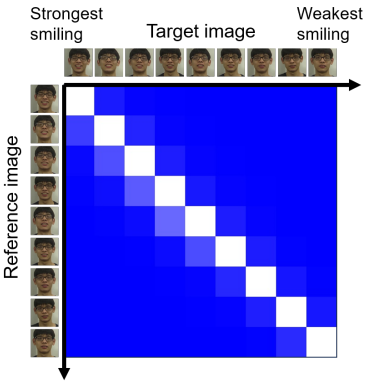


Figure 13: Consistency table calculated by selected reference images of dataset 1.

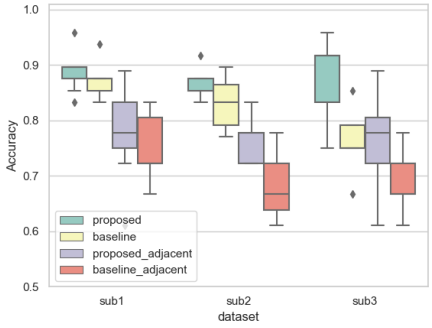


Figure 14: Accuracies of annotators evaluation. Adjacent means the image pair consists of the nearest neighbor images.

image pairs with higher accuracy than the comparison method.

5.3.4. Smiling intensity evaluation

Examples of face images evaluated by selected reference images and the proposed network are shown in Figure 16. Since it is sometimes hard to qualitatively evaluate two adjacent images in a row, the four reference images skip one rank at a time. The images with a smile level one class lower than the reference image are listed, and each row shows the same evaluation value. The images on the left side of the figure are recognized as having a higher degree of smiling. This result confirms that the proposed method effectively evaluates the degree of smiling within the ordinal scale.

Finally, a part of the transition of the smiling intensity during the experiment is shown in Figure 17. In this result, an evaluation score was smoothed by the median filter to trace the trend of transitions. We can see that

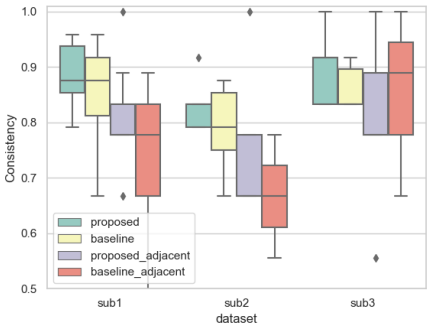


Figure 15: Consistencies of annotators evaluation of the same image pair. Adjacent means the image pair consists of the nearest neighbor images.

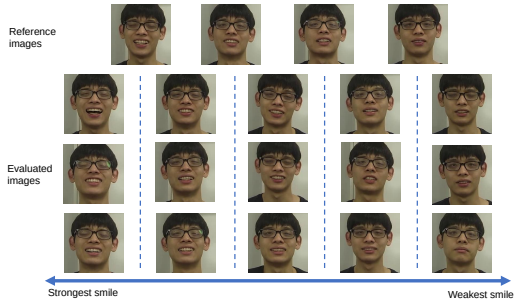


Figure 16: Example results of the evaluation of the degree of smiling

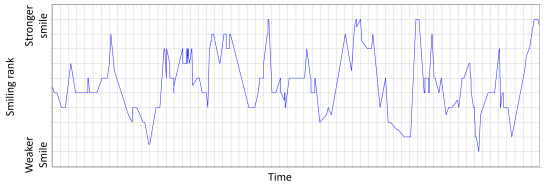


Figure 17: A part of the transition of the degree of the smiling. Each grid line of time shows 10 seconds.

the participant smiles several times in this period. It can be seen that smiles of slightly stronger intensity than the middle level occurred several times in succession in the first half of this period. In comparison, smiles of considerably stronger intensity occurred with a short interval in the second half of this period.

6. Conclusion

In this paper, we propose an approach to evaluate the degree of smiling of individuals by ordinal scales based on multiple comparisons for the purpose of monitoring individuals. Suppose that we have enough data from individual face images; we also propose an algorithm for selecting appropriate reference images for the ordinal evaluation.

Experimental results show that our ordinal scale-based evaluation can successfully give the degree of not only clear smiling but also intermediate facial expressions. In addition, we can see that an evaluation space constructed by selected reference images by our algorithm is more consistent and, therefore, considered to be reasonable.

One of the future works is to map the proposed and constructed ordinal scale to some physical index. Although this paper proposed a method of selecting reference images that are somewhat reasonable when evaluated by humans, the validity of the scale would be improved if it could be mapped to some physical index. For example, by measuring the myoelectricity of facial muscles, the degree of muscle activity could be used as an index. In addition, the other future work is to apply this technique to people whose facial expressions do not change much, e.g., dementia patients, as we described in the introduction section.

Ethics

Our method aims to monitor the daily health conditions of a specific individual by evaluating the smiling intensity using a model trained specifically for the individual's facial images. Since data for model training and smiling intensity evaluation can be collected and processed at terminals installed in each individual's environment, it is expected to reduce the risk of leakage of particularly strong personal information such as facial images being stored in the cloud in practical applications.

References

[1] K. Kondo, T. Nakamura, Y. Nakamura, S. Satoh, Siamese-structure deep neural network recognizing changes in facial expression according to the degree of smiling, in: Proc. of ICPR2020, 2021, pp. 4605–4612. doi:10.1109/ICPR48806.2021.9411988.

[2] P. EKMAN, Facial action coding system (facs), A Human Face (2002). URL: https://ci.nii.ac.jp/naid/10025007347/.

[3] B. Amos, B. Ludwiczuk, M. Satyanarayanan, OpenFace: A general-purpose face recognition library with mobile applications, Technical Report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[4] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. of ICLR 2015, San Diego, CA, USA, May 7-9, 2015.

[5] C. C. Atabansi, T. Chen, R. Cao, X. Xu, Transfer learning technique with vgg-16 for near-infrared facial expression recognition, Journal of Physics: Conference Series 1873 (2021) 012033. URL: https://dx.doi.org/10.1088/1742-6596/1873/1/012033. doi:10.1088/1742-6596/1873/1/012033.

[6] Y. Liu, Facial expression recognition model based on improved vggnet, in: 2023 4th International Conference on Electronic Communication and Artificial Intelligence (ICECAI), 2023, pp. 404–408. doi:10.1109/ICECAI58670.2023.10177007.

[7] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a "siamese" time delay neural network, in: Proc. of NIPS'93, 1993, pp. 737–744.

[8] J. Bromley, J. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Sackinger, R. Shah, Signature verification using a "siamese" time delay neural network, International Journal of Pattern Recognition and Artificial Intelligence 7 (1993) 25. doi:10.1142/S0218001493000339.

[9] X. Zhou, W. Liang, S. Shimizu, J. Ma, Q. Jin, Siamese neural network based few-shot learning for anomaly detection in industrial cyber-physical systems, IEEE Transactions on Industrial Informatics 17 (2021) 5790–5798. doi:10.1109/TII.2020.3047675.

[10] G. Koch, R. Zemel, R. Salakhutdinov, Siamese neural networks for one-shot image recognition, in: Proc. of the deep learning workshop in the 32nd International Conference on Machine Learning, volume 2, 2015.

[11] J. Zhang, K. Shimonishi, K. Kondo, Y. Nakamura, Facial expression change recognition on neutral-negative axis based on siamese-structure deep neural network, in: Cross-Cultural Design. Product and Service Design, Mobility and Automotive Design, Cities, Urban Areas, and Intelligent Environments Design: 14th International Conference, CCD 2022, Held as Part of the 24th HCI International Conference, HCII 2022, 2022, pp. 583–598.

[12] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: British Machine Vision Conference, 2015.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, International Journal of Computer Vision 128 (2019) 336 – 359. doi:10.1007/s11263-019-01228-7.