Facial Expression Change Recognition on Neutral-Negative Axis Based on Siamese-Structure Deep Neural Network

Junyao Zhang(✉), Kei Shimonishi, Kazuaki Kondo, and Yuichi Nakamura

Kyoto University, Kyoto, Japan
zhang.junyao.65m@st.kyoto-u.ac.jp, shimonishi@i.kyoto-u.ac.jp

Abstract. Facial expressions are very crucial in everyday communication, since a person's internal state of emotion swings can be assessed based on changes in expressions. Because there have been previous studies on positive expressions, and people with diseases are more inclined to show negative expressions, we focused our study on negative expressions. And we provided a new objective measure of facial expression that can be applied to dementia care, rehabilitation support, QOL assessment, etc. We analyzed the process of "change" in facial expressions for obtaining the degree of negative expressions, rather than the extreme facial expressions previously studied in various fields. Unfortunately, since there is no database on expression change picture sequences, we collected a small-scale database on our own in the experiment. We applied Siamese-structure deep neural network for classification of the pictures. The results show that the muscle parts affected by the change of facial expressions from neutral to negative, such as lip corners and forehead, can also be extracted accurately by the network, and a recognition accuracy of over 92% is obtained.

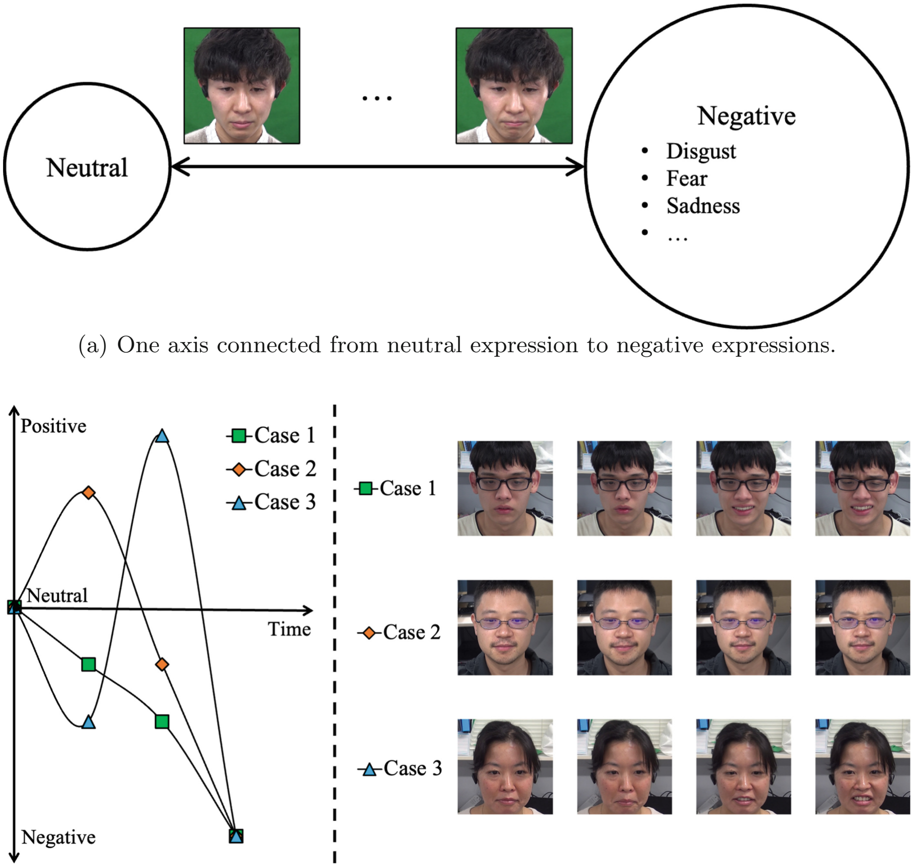
Keywords: Facial expression change · Siamese network · HRQOL

1 Introduction

Face is a mirror of the soul, reading face is a useful and essential to estimate person's internal emotion states. This kind of "reading" is crucial not only in socialization, but also in doctor-patient relationship. For example, dementia becomes one of the main causes of disability in elderly, which has led to verbal communication with them becoming increasingly difficult, and they needed to communicate with more often through non-verbal means. Therefore, automatically detecting human's facial expressions to analyze their inner emotions is desirable for families or caregivers to understand the patients' true feelings in a timely manner.

To trace face and analyze expressions in daily life, our research aim is to construct axes connecting two different expressions in expression space. Through

P.-L. P. Rau (Ed.): HCII 2022, LNCS 13314, pp. 583–598, 2022.
https://doi.org/10.1007/978-3-031-06053-3_40



(a) One axis connected from neutral expression to negative expressions.

(b) There may be multiple ways to change the expression from neutral to negative.

Fig. 1. Neutral - Negative expression axis.

these axes, we can obtain the intensity of expressions represented by any point in this space. In this paper, we pay special attention to the neutral-negative axis, because patients tend to show negative expressions more often. To find this axis, intermediate expression needs to be recognized because the axis is made up of the points represented by the intermediate expressions. While current facial expression recognition (FER) methods mainly focus on clear expressions, for example, methods that analyzing the movement of facial muscles [1], or extracting features by machine learning [2–4]. To recognize intermediate expressions, we borrow the idea from Kondo et al. [5], which comparison based method can be used to obtain the smiling level by recognizing facial expression changes. And, we extend the application of the network to neutral-negative expression axis.

On this neutral-negative axis (see Fig. 1(a)), we connect the neutral expression and negative expression, and can arrange the facial expression images in order to get different expression intensity. For facial expression changes, there are non-linear transitions, as shown in Fig. 1(b) for Case 2 and Case 3, where the former first goes from neutral expression to positive expression and then drops to negative expression; the latter first drops from neutral expression, followed by a relatively larger expression fluctuation, rising to positive expression, and finally dropping to negative expression again. There are also linear transitions existed as in Case 1, with a gradual decline from neutral expression to negative expression. Although the transition is not always monotonic, we assume that a linear monotonic expression change sequence can be obtained by cutting the non-monotonic sequence.

2 Related Work

This research is closely related to the 3 branches of study; Emotion and expression, Facial action coding system (FACS), and facial expression recognition (FER).

2.1 Emotion and Expression

In 1964, Tomkins and McCarter [6] demonstrated that facial expressions were reliably associated with certain emotional states. When a person perceives certain emotions, such as smiling and frowning, facial muscles will make corresponding movements, these muscles are commonly called muscles of facial expression, their specific location and attachments enable them to produce movements of the face. Based on this, the American psychologist P. Ekman et al. [7] first defined 7 basic emotion expressions, that is anger, disgust, fear, enjoyment, sadness, surprise and contempt.

2.2 Facial Action Coding System

With the development and in-depth study of facial expressions. They proposed the Facial Action Coding System (FACS) [8], which interprets expressions by analyzing the motion features and regions of corresponding facial action units. Facial muscles are like pieces in a jigsaw puzzle, and by moving them to quantitatively analyze the true emotions of the puzzle's owner.

Subsequently, in recent years, Facial action unit (AU) provide information in describing comprehensive facial expression, such as enjoyment is calculated from the combination of AU6 and AU12, separately describes *Cheek Raiser* and *Lip Corner Puller*, correlated to orbicularis oculi and zygomatic major muscle. K. Zhao et al. [9] proposed deep region and multi-label learning (DRML), which is able to identify more specific regions for different AUs than conventional patch-based methods. The important innovation in DRML is the use of a feed-forward function to induce the network to focus on the crucial regions, and better classification accuracy was obtained by constructing relationships between important facial regions and multiple AUs. J. He et al. [10] designed

CNN models to do facial view recognition and AU occurrence recognition, and achieved 97.7% accuracy in 9 facial views classification. In the next year, A. Romero et al. [11] took advantages of the ability of CNN to do large-scale classification, presented HydraNets and AUNets and did detection and classification on 12 different AUs. G. M. Jacob et al. [12] used the now-hot transformer to classify AU, and achieved absolute improvement on BP4D and DISFA datasets, which contain input images and action unit labels.

2.3 Facial Expression Recognition

Recognition for Clear Expressions. To extract suitable and accurate features is crucial for FER research, for features are important basis for classification. And expression feature extraction is mainly based on mathematical models to extract the continuous deformation and the muscle movement of different regions on the face. For example, Gabor wavelet method [13] and Local Binary Pattern (LBP) operator method [14] for single image, and optical flow method [15] for dynamic sequence.

The great achievement of deep learning in the field of image recognition has provided a new idea for FER. Unlike the traditional approach, deep networks can perform FER using an end-to-end approach, the network can then directly output the predicted probability for each sample. In addition to the end-to-end learning approach, another option is to use deep neural networks (especially CNNs) as a feature extraction tool and then apply additional independent classifiers, such as SVM or random forests.

Most applications of deep learning in face expression recognition are based on VGGNet, GoogleNet and ResNet network models, whose core structure is a deep convolutional neural network (DCNN). Dynamic sequence FER adds analysis of temporal and spatial changes in expressions. And most approaches focus on identifying expressions with peak emotional intensity in a sequence, while ignoring frames with low amplitude emotional expression. This causes several frames with obvious expressions can be accurately recognized, but how expressions change in this sequence cannot be gotten. However, expression intensity-invariant networks, such as peak-piloted deep network (PPDN) [16] and deeper cascaded peak-piloted network (DCPN) [17], or Siamese network [5] can identify trends in expression changes. The former by identifying correlations between frames of different intensities in a sequence, and the latter by comparison, which introduced later in detail. 3D CNN proposed by Ji et al. [18] to capture motion information encoded in multiple adjacent frames for action recognition through 3D convolutions make it possible to recognize dynamic expression changes. In order to capture the differences in expressions over time, a model that can do the comparison between two temporally continuous or non-continuous face images is needed.

Recognition for Intermediate Expressions. In contrast to recognition for clear expressions, there are less research on intermediate expression recognition. A. Toisoul et al. [19] have used jointed face alignment networks and emotion classification networks to estimate the emotion categories of faces under naturalistic conditions as well as their continuous values of valence and arousal levels, and have obtained quite good results. Even though the datasets they used, AffectNet, SEWA and AFEW-VA, have lots of ambiguity in the rating of valence and arousal levels, they still give us inspiration to study the trends of expression intensity change.

And our group has developed comparison-based techniques to detect the changes in facial expression [5]. Kondo et al.successfully classified the smile level by applying Siamese-structure neural network on expression change sequences. The network received two preprocessed face images as inputs, and output two feature vectors. Then rather than calculated the distance between vectors in traditional Siamese network, they applied a sequence of fully connected layers to receive them, and output two likelihood values corresponding to ascension and descension labels relative to the degree of smiling. And this network achieved >95% accuracy in recognizing smiling changes under the reasonable attention regions that contributed to the predicted labels, such as mouth, cheeks, and the tail of eyes.

3 Siamese-structure Deep Neural Network

As introduced in Sect. 1, we apply the comparison-based network proposed by Kondo et al. [5] to solve our problem. Because Siamese network is used for measuring the similarity of two things, such as the application in face verification [25], semantic similarity analysis of vocabulary [26], etc., in order to make it can recognize the changes in expression, a image classification network is attached after. Figure 2 shows the structure, two Images I_0 and I_0 with ground truth (abbreviated as G.T.) y_0 and y_1 respectively, were fed into two identical CNNs (here, VGG16 is employed) with the same weight parameters. Following this, the extracted feature vectors were input into two fully connected layers with rectified linear units, finally output the predicted likelihood values $\hat{y}_0$ and $\hat{y}_1$. The ground truth likelihood values (y_0, y_1) can be considered as a two-dimensional one-hot vector, when I_0 show more negative expression, its corresponding element y_0 is set to be 0, the other element y_1 is set to 1, and vice versa. And, we define the predicted labels as follow:

- $\hat{y}_0 > \hat{y}_1$ means I_1 is predicted with more negative expression than I_0;
- conversely, $\hat{y}_0 < \hat{y}_1$ predicts that I_0 displays more negative expression.

For the sample in Fig. 2, $y_0 = 1, y_1 = 0$, because I_0 displayed more negative expression. If the network outputs the label $\hat{y}_0 > \hat{y}_1$, the prediction is correct, and vice versa.

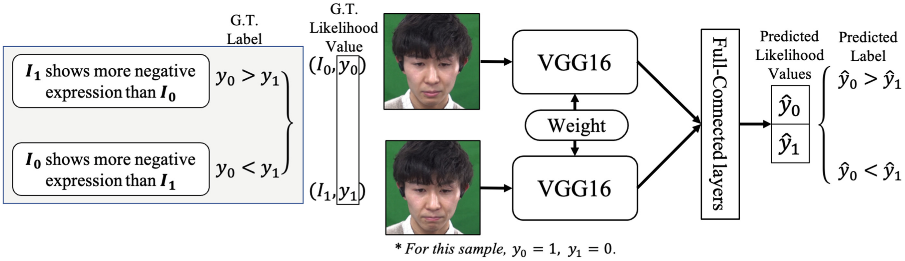
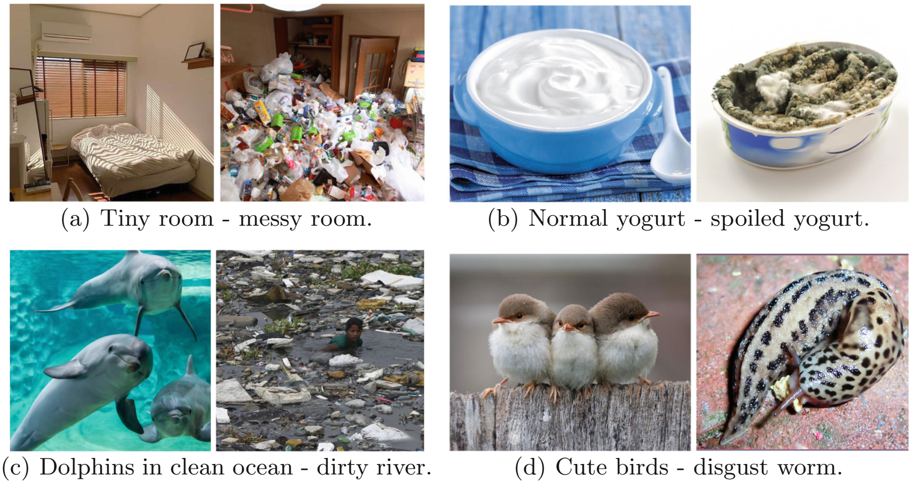


Fig. 2. Siamese-structure deep neural network.

4 Experimental Method and Construction of Data Sets

In the experiment, we invited a total of 11 subjects (male: female = 8: 3). Based on the assumption stated before, we collected the sequence data of facial expression changes by showing the visual stimuli to subjects. 76 groups of image pairs are used as visual stimuli, samples are shown in Fig. 3, and displayed on the liquid crystal display (LCD) monitor (D491UB | DOSHISHA, 4K, 3840 × 2160, 49inches).


(a) Tiny room - messy room.

(b) Normal yogurt - spoiled yogurt.

(c) Dolphins in clean ocean - dirty river.

(d) Cute birds - disgust worm.

Fig. 3. Four pairs of samples from visual stimuli used in the experiment.

Meanwhile, video camera (SONY HDR-CX630) was fixed directly front to recorded subject's facial expressions. During the experiment, subject sit on chair and faced the monitor, after a detailed description of this experiment, one pair of neutral image and negative image would be displayed on the screen for 10 s each in turn after a 3 s-countdown, this procedure is shown as Fig. 4. When the time was up, the next pair would be automatically played until the end of the experiment. The facial expression is expected to be changed from neutral/positive to negative under the influence of images. To prevent the surrounding sounds from interfering with them, a pair of wireless headsets playing white noise was worn.

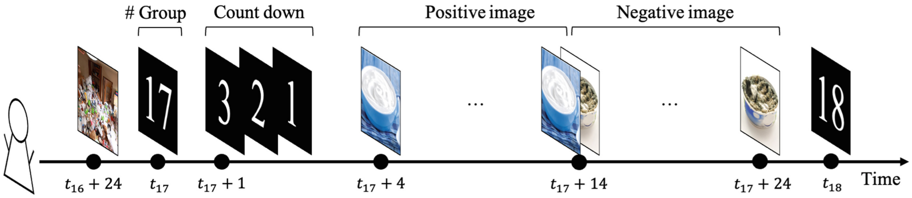


Fig. 4. The procedure of experiment.

We conducted the experiment and collected the dataset. The dataset includes changes in facial expressions from neutral to negative elicited by visual stimuli under laboratory conditions, with a higher level of labeling accuracy.

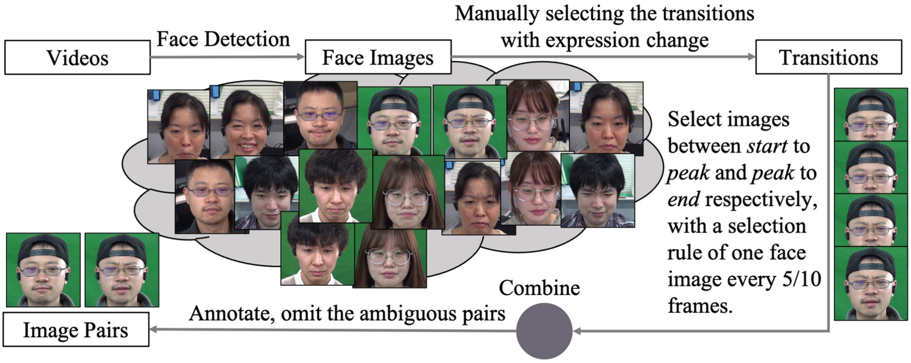

Videos
Face Detection
Face Images
Manually selecting the transitions
with expression change
Transitions
Select images between *start* to *peak* and *peak* to *end* respectively, with a selection rule of one face image every 5/10 frames.
Combine
Annotate, omit the ambiguous pairs
Image Pairs

Fig. 5. Preprocessing process.

The original data in dataset are video data, in order to obtain the input required for the deep neural network model, the video data need to be preprocessed as follows (see Fig. 5). Firstly, apply face detection, face alignment and face cutting to get face images. Secondly, manually select expression change sequences of no more than 20 s, and record the start frame (where the expression began), peak frame (where the expression intensity reaches its maximum) and end frame (where the expression ended) of each sequence, the frame rate is 30 frame per second. Thirdly, select 1 frame every 5 or 10 frames in the (start, peak) frame and (peak, end) frame interval respectively, then combine every two frames to get image pairs. After annotating and omitting the ambiguous pairs by human, clear image pairs with obvious difference expressions were finally remained.

Even though we expected expression change to happen when the image change from neutral to negative inside one image pair, when confronting with even the same image, different person will be stimulated with different internal emotions, thus affected in different external expressions. Sometimes we cannot collect expected number of sequences, because of expression change was not observed or appeared several times. The reason for the former one is sometimes emotional intensity is not strong enough to have changes in face visible to the naked eye. For the later one is the presence of intermittent facial expression changes. To illustrate in more detail, in response to a set of stimuli, subjects may exhibit two or more expression changes and there is a certain time interval between two changes. According to the follow-up small interview, this situation is most likely to occur due to the following reason. The first expression change was caused by the subconscious feeling of shock and disgust or displeasure when suddenly facing a negative stimulus on the monitor, while the subsequent expression changes were caused by the internal emotion fluctuation due to thinking. For example, in Fig. 3(c), from dolphins swimming in clean and clear ocean to a child walking in a dirty river polluted by garbage, the subject's first reaction was the disgusting emotion produced by the visual stimulus of clean (blue) to dirty (black) water, but after thinking or associating with the living conditions of the local people, the emotion of sadness would appear.

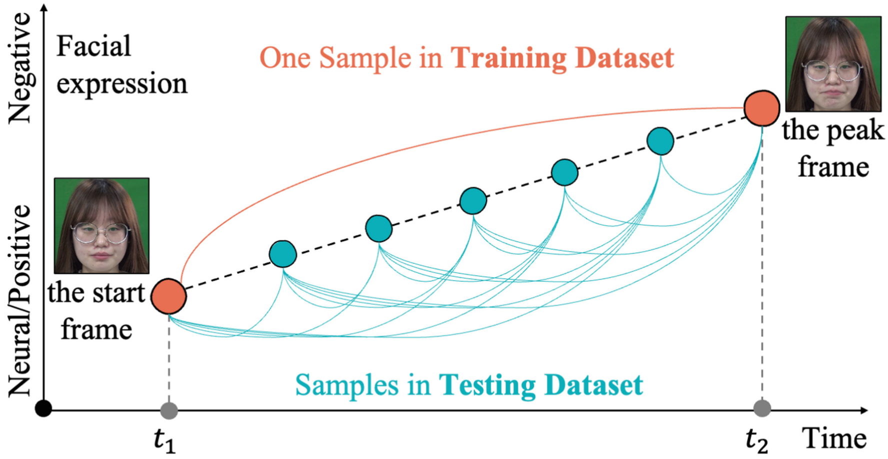


Fig. 6. Training dataset and testing dataset.

For training deep neural network, we need to separate the input image pairs into three parts; training set, validation set and testing set. Training dataset and validation dataset use the (start, peak) and (peak, end) frames in facial expression transitions, the start and peak frame image pair are connected with the orange line in Fig. 6; testing dataset uses the combination of frames inside expression transitions, select one frame every five frames, shown as the blue points in Fig. 6. As a result, we obtained 1360 training samples.

5 Results

15-Folds cross-validator is used in the training process. We obtained the training accuracy on basic data set

$$\frac{\#\text{Correct Classified Samples}}{\#\text{Whole Samples}} = \frac{1263}{1360} = 92.87\%.$$

To validate the region of interest (ROI) of the neural network, we visualize the images on the validation set using the Gradient-weighted class activation mapping (Grad-CAM) to see which pixels at which locations in the image have a strong influence on the output. Grad-CAM calculate the weights through global average of gradients.

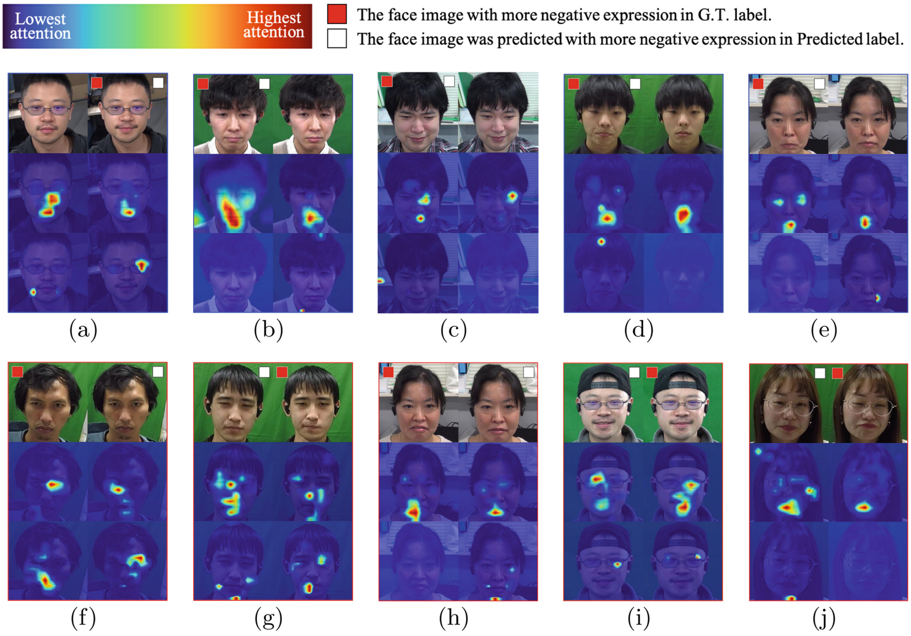

Lowest attention
Highest attention
🟥 The face image with more negative expression in G.T. label.
□ The face image was predicted with more negative expression in Predicted label.
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)

Fig. 7. Recognized samples from training dataset with Grad-CAM. In each sample, the top row displays two input images. The middle and the bottom rows indicate contribution maps corresponding to the predicted label and the discarded label in the estimation, respectively.

Figure 7 shows 10 samples. The upper left of the figure shows the heatmap colorbar, with different colors from left to right meaning from lowest attention to highest attention, used to visualize the importance of the regions on which the network relies in recognition. Two squares in red and white are shown on the upper right, indicating image with more negative expressions in ground truth and prediction, respectively. When the two squares appear on the same image, meaning the more negative one in this pair can be correctly recognized, instead, it is incorrectly recognized. For each sample, the top row shows two images I_0 and I_1 input to the network, and the red and white squares are displayed on the images, the middle row and the bottom row shows the attention regions that contributed to the predicted label and discarded label by contribution maps, respectively. Figure 7(a)∼(e) are correctly recognized samples, from the middle row, attention regions spread of mouth, nose and eyes; from the bottom row, attention to small and meaningless regions such as hair, brow, neck and mouth corners contributed to the discarded label. This results are convincing according to FACS. Figure 7(f)∼(j) are incorrectly recognized samples, which can be classified into two cases. One is two input images with a relatively high degree of similarity, such as Fig. 7(f), as with the middle row, the presence of meaningful ROI contributes to the discarded labels; another is the sample imbalance within the dataset, such as Fig. 7(g)∼(j), although the differences can be accurately identified, it is not possible to determine which one has a more negative expression.

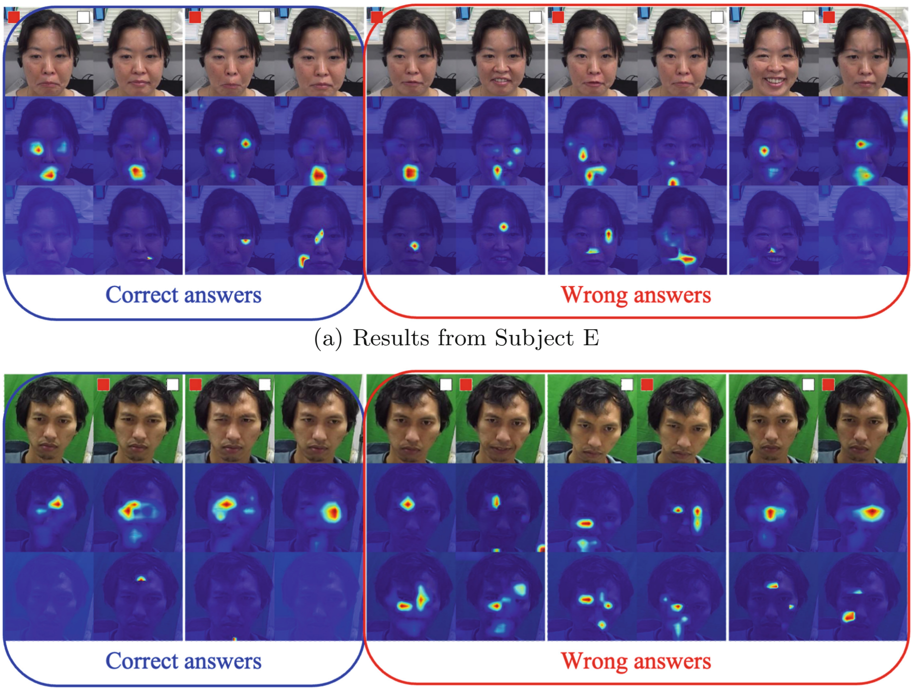
Due to the existence of individual differences, we calculated the classification accuracy for each subject separately, as shown in Table 1.

Table 1. Classification accuracy on each subject.

#Sub	#Sample	Accuracy
A	Fig. 7(a), Fig. 7(i)	311/332 ≈ 93.67%
B	Fig. 7(b)	176/178 ≈ 98.88%
C	Fig. 7(c), Fig. 7(g)	207/224 ≈ 92.41%
D	Fig. 7(d)	52/58 ≈ 89.66%
E	Fig. 7(e), Fig. 7(h)	38/58 ≈ 65.52%
F	Fig. 7(f)	71/80 ≈ 88.75%
G	Fig. 7(j)	153/162 ≈ 94.44%

What causes such low classification accuracy? For subjects E and F, we analyzed them separated as follows:

- E: The change in expression of Subject E was mainly focused on the lip and lip corner, the ambiguous change around lip and samples that facial expression change from positive and negative are less in training set, these two reasons lead to such low classification accuracy, see Fig. 8(a).
- F: Subject F's facial expression is frown and squint, his mouth movement is not obvious. Our training set has more samples of mouth region changes and fewer samples of frowns, and this imbalance of samples might be the cause of low recognition rate, and incorrect region focus, see Fig. 8(b).



(a) Results from Subject E

(b) Results from Subject F

Fig. 8. Presentation of the results for subjects E and F with low classification accuracy. Training data and results were visualized using Grad-CAM. The left two samples show the correct answers, the right three samples show the wrong answers.

We test the trained weigh model on the testing dataset, Fig. 9 shows one correct classification case, where whole image pairs in this sequence are correctly classified. and Fig. 10 shows two wrong cases, where some image pairs in the sequence are incorrectly classified.

Take Fig. 9 as an example, we will give a description. This figure can be divided into three parts:

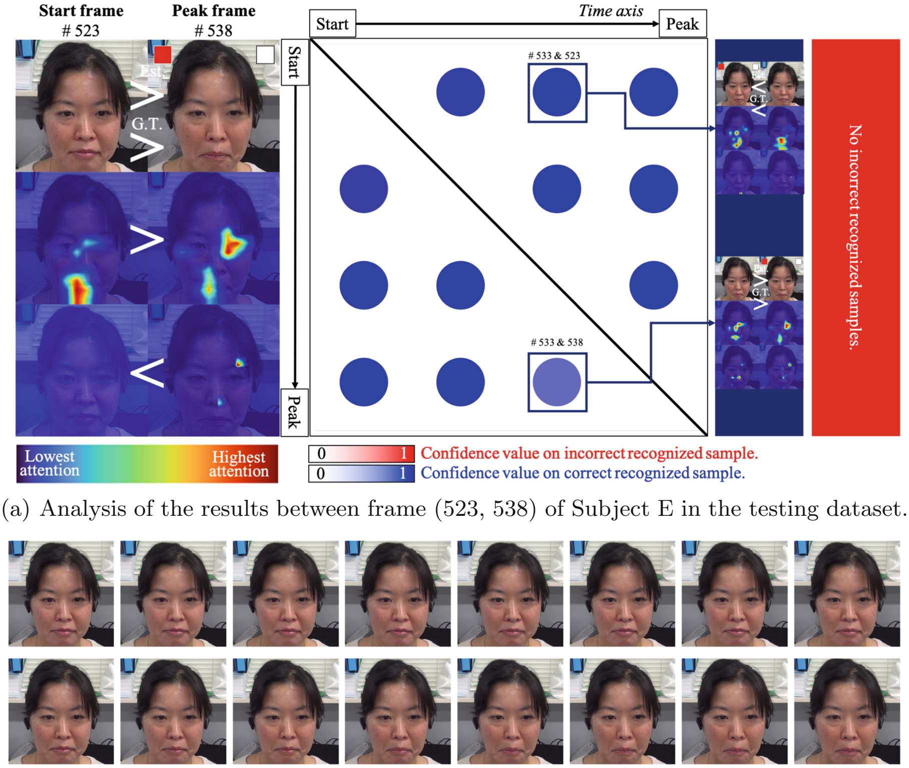
- The left image shows one pair of images ((start, peak) frames) in the training set;
- The middle image shows the confidence map generated by temporal alignment of image pairs in one transition from testing set. The horizontal axis from left to right and the vertical axis from top to bottom both represent all frames arranged in chronological order within an expression transition interval. Each dot represents the recognition result obtained after testing the corresponding two frames on the horizontal and vertical axes through the network. Where (1) blue dot means the correctly recognized image pair, no matter it is in the upper or lower part of the diagonal, which means the predicted label is the same as the ground truth label; (2) red dot means the incorrectly recognized

image pair, which means the predicted label is opposite of the ground truth label.

- For (start, peak) pairs: blue dot means $\hat{y}_0 > \hat{y}_1$ below the diagonal, and $\hat{y}_0 < \hat{y}_1$ above the diagonal; red dot means the opposite;
- For (peak, end) pairs: blue dot means $\hat{y}_0 < \hat{y}_1$ below the diagonal, and $\hat{y}_0 > \hat{y}_1$ above the diagonal; red dot means the opposite.

The shades of red/blue indicate the confidence value, the more red/blue the color, the higher the confidence value; the whiter the color, the lower the confidence value.

- The left image with blue and red rectangles displays some results with Grad-CAM visualization.

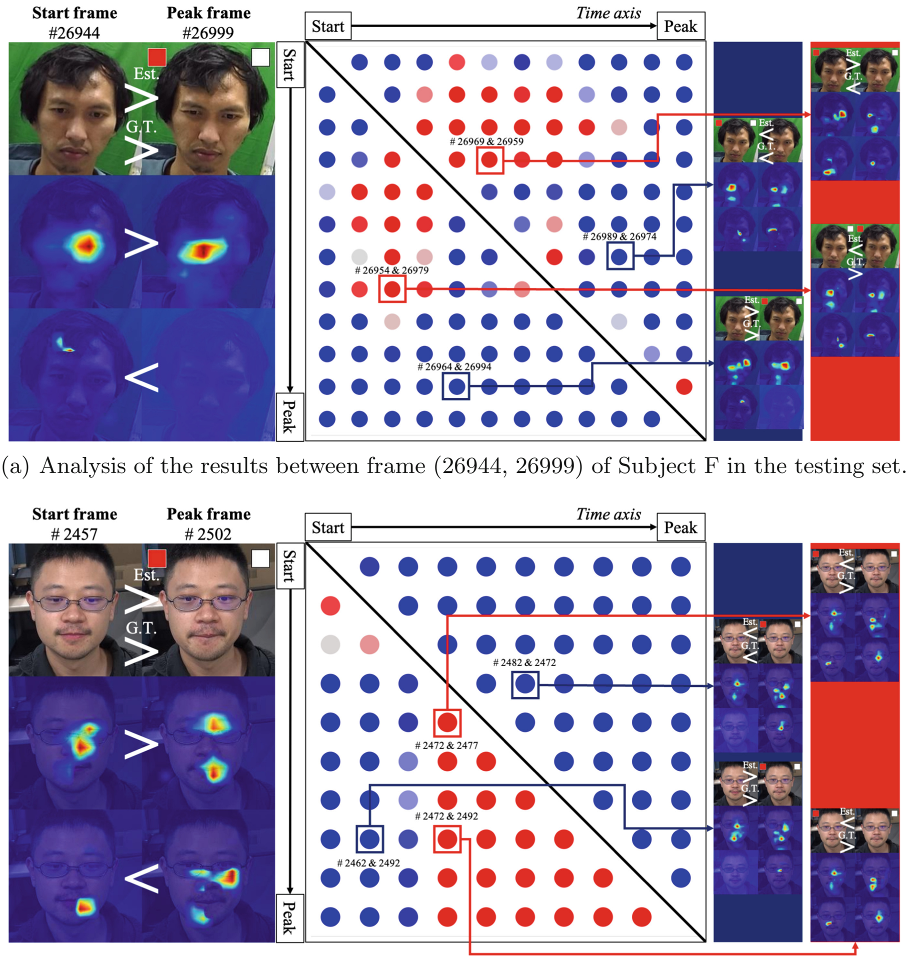


(a) Analysis of the results between frame (523, 538) of Subject E in the testing dataset.

(b) All face images from frame 523 to frame 538.

Fig. 9. Correct classified case.

For the sequence in Fig. 9(b), the expression change from neural (start) to negative (peak). By the fact that the color of all the dots in the confidence map is blue, all pairs of face images obtained at every 5 frames within the sequence can be correctly recognized. Therefore, we can illustrate that the expression within the sequence change linearly, since each face image expressed more negative expression than the previous one.



(a) Analysis of the results between frame (26944, 26999) of Subject F in the testing set.

(b) Analysis of the results between frame (2457, 2502) of Subject A in the testing dataset.

Fig. 10. Wrong classified cases with symmetrical misclassification and triangular misclassification.

Figure 10(a) shows the symmetrical misclassification, meaning the wrong answers appears on both sides of the diagonal, the data that fit closely to the diagonal are characterized by a high degree of similarity between the two face images, the reason for this wrong case is probably corresponding to the non-linear transition exists in this sequence. Although we assume the image sequences in our dataset are linear after processing, in Sect. 1, this processing is all based on manual work, and perhaps there are some non-linear changes hidden in these sequences that are not classified by humans. And, Fig. 10(b) shows the triangular misclassification, which means only one side of the data can be classified correctly, while most of the data on the other side of the diagonal cannot be classified correctly, which is probably due to the overfitting.

6 Conclusions

We collected a dataset on expression changes and achieved superior recognition results based on a comparative deep neural network approach on the recognition of expression changes that are on the neutral-negative expression axis. The network used the difference in features at lip corner and forehead of the person in the two images as a key to distinguish whether there is a negative change or not. However, more definite conclusions will be possible when we can take care of three existing problems:

– Inaccurate manual splitting of image sequences. According to visual stimuli, there are non-linear changes in the sequence of expression changes, so it needs to be manually segmented again to cut the non-linear changes into multiple linear changes. However, this manual segmentation will have certain errors.
– The samples in the dataset are unevenly distributed; for example, some subjects tend to make frowns when they are subjected to negative stimuli; some subjects tend to make frowning facial gestures when subjected to negative stimuli, while others tend to make facial gestures such as tugging at the corners of their mouths, and the inconsistent number of frowning and tugging facial pictures will lead to the model's inaccurate classification of a smaller number of sample categories.
– Overfitting is occurred due to the small size of the data set. This issue was addressed by data augmentation, but more data are still needed to achieve more accurate and generalizable conclusions.

Future work will specifically split the more ambiguous classification of negative expressions into clear sub-categories such as disgust, fear, etc., and add positive expressions to explore the variation of expressions on different expression combination axes with the help of Russell's circumplex model [27]. After the implementation into the monitoring device, the patients with dementia will be monitored in real time and suggestions will be given to improve HRQOL in response to their expression changes.

References

1. Ekman, P., Friesen, W., Hager, J.: Facial action coding system (FACS). A human face (2002)
2. Vasudeva, K., Chandran, S.: A comprehensive study on facial expression recognition techniques using convolutional neural network. In: International Conference on Communication and Signal Processing, pp. 1431–1436. IEEE, Chennai, India (2020)

3. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
4. Zhang, K., Huang, Y., Du, Y., et al.: Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans. Image Process. **26**(9), 4193–4203 (2017)
5. 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: 25th International Conference on Pattern Recognition (ICPR), pp. 4605–4612. IEEE, Milan, Italy (2021)
6. Tomkins, S.S., McCarter, R.: What and where are the primary affects? some evidence for a theory. Perceptual Motor Skills **18**(1), 119–158 (1964)
7. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Personality Soc. Psychol. **17**(2), 124–129 (1971). https://doi.org/10.1037/h0030377
8. Ekman, P., Friesen, W., Hager, J.: Facial action coding system (FACS). Am. Psychol. Assoc. PsysTests (1978). https://doi.org/10.1037/t27734-000
9. Zhao, K., Chu, W., Zhang, H.: Deep region and multi-label learning for facial action unit detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas, NV, USA (2016). https://doi.org/10.1109/CVPR.2016.369
10. He, J., Li, D., Yang, B., Cao, S., Sun, B., Yu, L.: Multi view facial action unit detection based on CNN and BLSTM-RNN. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, Washington, DC, USA (2017). https://doi.org/10.1109/FG.2017.108
11. Romero, A., Leon, J., Arbelaez, P.: Multi-view dynamic facial action unit detection. Image Vis. Comput. (2018). https://doi.org/10.1016/j.imavis.2018.09.014
12. Jacob, G.M., Stenger, B.: Facial action unit detection with transformers. In: 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Nashville, TN, USA (2021). https://doi.org/10.1109/CVPR46437.2021.00759
13. Vinay Kumar, B., Shreyas, B.S.: Face recognition using gabor wavelets. In: Fortieth Asilomar Conference on Signals, Systems and Computers. IEEE, Pacific Grove, CA (2007). https://doi.org/10.1109/ACSSC.2006.354817
14. Zhang, G., Huang, X., Li, S.Z., Wang, Y., Wu, X.: Boosting Local Binary Pattern (LBP)-Based Face Recognition. Advances in Biometric Person Authentication. SINOBIOMETRICS 2004. Lecture Notes in Computer Science **3338**, 179–186 (2004)
15. Ranftl, A., Alonso-Fernandez, F., Karlsson, S.: Face tracking using optical flow. In: 2015 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–5. IEEE, Darmstadt, Germany (2015). https://doi.org/10.1109/BIOSIG.2015.7314604
16. Zhao, X., Liang, X., Liu, L., Li, T.: Peak-piloted deep network for facial expression recognition. In: 14th European Conference on Computer Vision (ECCV), pp. 425–442. Amsterdam, The Netherlands (2016)
17. Yu, Z., Liu, Q., Liu, G.: Deeper cascaded peak-piloted network for weak expression recognition. Vis. Comput. **34**(12), 1691–1699 (2017). https://doi.org/10.1007/s00371-017-1443-0
18. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)

19. Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., Pantic, M.: Estimation of continuous valence and arousal levels from faces in naturalistic conditions. Nature Mach. Intell. **3**, 42–50 (2021)
20. Ramkumar, G., Logashanmugam, E.: An effectual facial expression recognition using HMM. In: 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), pp. 12–15. IEEE, Ramanathapuram, India (2016). https://doi.org/10.1109/ICACCCT.2016.7831590
21. Cohen, I., Sebe, N., Gozman, F.G., Cirelo, M.C., Huang, T.S.: Learning Bayesian network classifiers for facial expression recognition both labeled and unlabeled data. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. I-I. IEEE, Madison, WI, USA (2003). https://doi.org/10.1109/CVPR.2003.1211408
22. Khalifa, H., Babiker, B., Goebel, R., Cheng, I.: Facial expression recognition using SVM classification on mic-macro patterns. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 1757–1761. IEEE, Beijing, China (2017). https://doi.org/10.1109/ICIP.2017.8296583
23. Wang, X.H., Liu, A., Zhang, S.Q.: New facial expression recognition based on FSVM and KNN. Optik **126**(21), 3132–3234 (2015)
24. Li, J., Lam, E.Y.: Facial expression recognition using deep neural networks. In: 2015 IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1–6. IEEE, Macau, China (2015).https://doi.org/10.1109/IST.2015.7294547
25. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: the 27th International Conference on International Conference on Machine Learning (ICML), pp. 807–814 (2010). https://doi.org/10.5555/3104322.3104425
26. Kenter, T., Borisow, A., Rijke, M.D.: Siamese cbow: optimizing word embeddings for sentence representations. In: the 54th Annual Meeting of the Association for Computational Linguistics, pp. 941–951 (2016). https://doi.org/10.18653/v1/P16-1089
27. Russell, J.A.: A circumplex model of affect. J. Personality Soc. Psychol. **39**(6), 1161–1178 (1980)