PERFORMANCE EVALUATION OF MICRO-EXPRESSION RECOGNITION WITH VARIOUS NETWORK DESIGN VARIATIONS IN DEEP LEARNING

KOTMI HATA[1], TOMOAKI HIROSE[1], HIRONOBU TAKANO[1]

[1]Graduate School of Engineering, Toyama Prefectural University, Toyama, Japan
E-MAIL: u454018@st.pu-toyama.ac.jp, takano@pu-toyama.ac.jp

Abstract:

Human emotions are conveyed not only through verbal elements but also through non-verbal cues such as facial expressions, posture, and tone of voice. Among these cues, micro-expressions serve as important indicators of genuine emotions, as they can appear unconsciously even when a person tries to suppress their feelings. This study investigates the effectiveness of deep learning models for estimating internal emotional states based on micro-expressions.

The experiment was conducted by comparing five conditions: (1) facial region (full face vs. ROI), (2) number of input frames (apex frame vs. three-frame sequence), (3) presence or absence of fine-tuning using a facial expression dataset, (4) use of the Attention Branch Network (ABN), and (5) use of the ArcFace. The results showed that the use of ROIs and the implementation of ABN improved recognition accuracy, suggesting the importance of capturing local features and emphasizing attention regions in micro-expression recognition. In contrast, no significant improvement was observed from the number of input frames, fine-tuning, or the use of ArcFace. These findings indicate that the choice of input region and the use of attention mechanisms are key factors in enhancing the accuracy of emotion recognition based on micro-expressions.

Keywords:

Micro-expression; Emotion Estimation; ArcFace; Fine-tuning; ABN; ResNet

1. Introduction

Communication can be broadly categorized into two types[1]. One is verbal communication, which involves using language through speech, writing, or printed materials. The other is non-verbal communication, which includes facial expressions, tone of voice, gestures, and other forms that do not rely on words. Non-verbal elements, such as facial expressions, posture, gestures, and hand signs, play an essential role in conveying emotions effectively and facilitating smooth communication alongside verbal expressions. It is said that when people are uncertain about interpreting others' intentions, they rely on visual cues (e.g., facial expressions and gestures) for 55%, verbal content for 7%, and auditory information for 38% in their emotional judgment[2]. However, facial expressions may sometimes convey deceptive emotions. For example, people may show a smile to be polite, even if it does not reflect their true feelings. When the facial expression does not match the internal emotional state, it becomes more difficult to accurately interpret the person's true emotions. In such cases, humans can still make reasonable emotional inferences by interpreting the conversational context and the other person's overall demeanor. However, facial expression recognition systems using computers make it difficult to achieve such a nuanced understanding. On the other hand, micro-expressions are facial expressions that occur even when a person attempts to suppress or hide their emotions, and they are believed to reflect true feelings. Therefore, micro-expressions allow for more accurate emotional estimation, even when the outward expression may be deceptive. In this study, we aim to develop an emotion recognition method that estimates internal emotional states from facial features. To achieve this, we investigate which types of deep learning models are effective for estimating emotions from micro-expressions.

2. Related work

2.1. Datasets

Table1 lists the datasets of facial expressions and micro-expressions. CK+ is a facial expression dataset, while the others are micro-expression datasets. The CK+ dataset includes the seven basic emotional labels: anger, disgust, fear, happi-

ness, sadness, surprise, and contempt[3]. The SAMM dataset is the most culturally diverse, featuring participants from various ethnic backgrounds[4]. It is labeled with eight emotional labels: the same seven basic emotions as in CK+, plus an additional "others" label. The SMIC[5] dataset was created using three different types of cameras: a high-speed camera (HS), a standard visible light camera (VIS), and a near-infrared camera (NIS). It is labeled with three emotional labels: positive, negative, and surprise. The CASME II dataset is the largest and most widely used micro-expression dataset to date, and is labeled with five emotional labels: happiness, surprise, disgust, repression, and others[6].

TABLE 1. Dataset Overview.

Dataset	Subject	Samples	Emotion label
CK+	123	593	7
SAMM	32	159	7
CASME II	26	247	5
SMIC(HS)	16	157	3
SMIC(VIS)	8	71	3
SMIC(NIS)	8	71	3

2.2. Previous research

Wang et al. demonstrated that using transfer learning with a TLCNN (Transferring Long-term Convolutional Neural Network) for emotion recognition based on micro-expressions improves accuracy[7]. The results reported by Wang et al. are shown in Table2.

Kato et al. investigated the emotional classification performance by micro-expression using two types of features: LBP-TOP and CBP-TOP[8]. This study evaluated the accuracy of emotional estimation for the score-level fusion of the LBP-TOP and CBP-TOP. In addition, the feature selection using the ratio of the inter-class variance to the within-class variance was employed. The three types of ROIs, partial facial regions, were used for feature extraction. The experimental results showed that the feature selection for LBP-TOP and CBP-TOP was effective for emotional classification by micro-expression. In contrast, the score-level fusion did not improve the emotional estimation performance.

TABLE 2. Experimental results of Wang et al. []

Dataset used for fine-tuning	Accuracy
CK+ $\rightarrow$ MMEW(micro)	65.6
MMEW(Macro) $\rightarrow$ MMEW(micro)	69.4
CK+ $\rightarrow$ SAMM	73.5

3. Proposed method

3.1 Facial landmark detection and frame extraction

First, the frames in which micro-expressions occur are extracted from the video. Although each sequence in the datasets records the process of facial expression changes for each emotion, there are frames where no visible expression change occurs. Therefore, the onset frame, where the expression begins to change, is taken as the first frame, the apex frame, where the expression reaches its peak, is set as the third frame, and the intermediate frame is set as the second frame. These three frames are not necessarily consecutive, but are temporally spaced to capture the dynamic progression of facial expressions. Next, facial landmarks are detected for each extracted frame. For landmark detection, the face landmark detector in the Dlib library[9] was utilized. The detected landmarks are then used to crop the face or ROIs. As examples, the extracted frames from the CK+ and SAMM datasets are shown in Figures 1 and 2, respectively. The three extracted frames are converted to grayscale and then combined to form a single RGB-like image, which is used as the input. Furthermore, the training image data was augmented by a factor of ten through rotations, flips, and contrast adjustments. This study conducted experiments using the apex frame alone and the set of three temporally sampled frames.

FIGURE 1. Three frames capturing the process of facial expression change.CK+

3.2. Selection of ROI

ROI stands for "Region Of Interest", and in this study, it refers to facial regions associated with expressions. In this experiment, the face was divided into 15 regions based on the ROI

FIGURE 2. Three frames capturing the process of micro-expression change.SAMM

segmentation method used in Kato's research. The selected ROIs are shown in the Figure3. When a person experiences an emotion, particular facial regions move accordingly, and these movements appear as micro-expressions. The selected ROIs correspond to the facial areas where such expression-related changes are likely to occur.

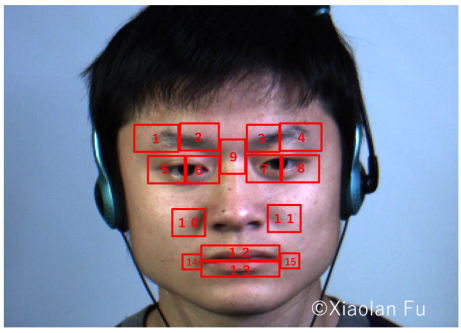

FIGURE 3. Selection of ROI.

3.3. Network Structures

In this experiment, various models were investigated. Here, we present one example: a network based on ResNet that, incorporatinges an Attention Branch Network (ABN) [10] and ArcFace[11]. The architecture of the constructed network is shown in the Figure4, and the hyperparameters are shown in the Table3. ABN is a technique that generates an attention map from the extracted feature maps to highlight important regions. This attention map is then applied to the original feature maps, and the resulting weighted features are used for classification. This study considered two approaches: one that uses only the apex frame, and another that uses a sequence of three frames. The difference between the two lies in the first layer of the network. For the apex frame approach, since the input consists of a single frame, 2D convolutional kernels are used. In contrast, the three-frame approach employs 3D convolutional kernels to capture temporal dependencies across frames. In this case, no padding is applied along the temporal axis. This design aims to enable the network to extract features that represent temporal variation in facial expressions.

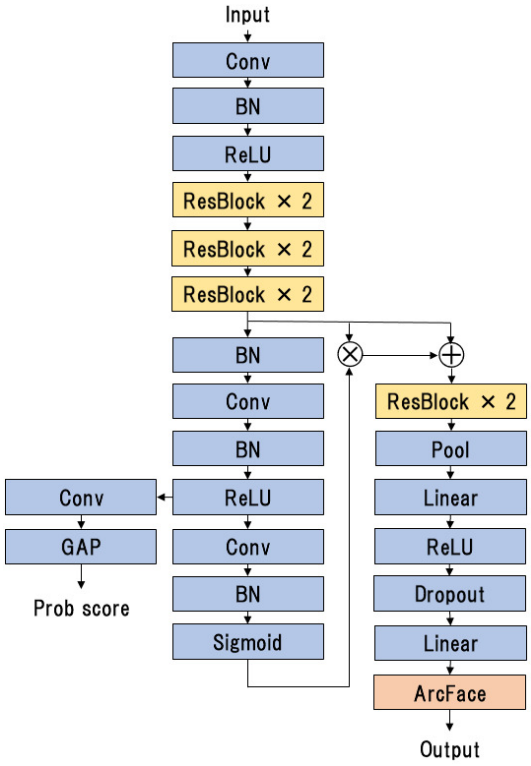

FIGURE 4. Network based on ResNet with ABN and ArcFace implemented.

TABLE 3. Hyperparameters.

Batch size	16
Number of epochs	100
Optimization function	SGD(1e-3)

4. Evalution method

In the experiment, each dataset was divided into five subsets, and five separate training sessions were conducted, using each subset as the test data in turn. In this cross-validation setting, the emotional labels in the test data were distributed as evenly as possible across the folds. For example, if there were 100 samples labeled as "anger," they were split into five groups of

20; similarly, 50 samples labeled as “happiness” were divided into five groups of 10. In this experiment, recognition accuracy was defined as the proportion of samples for which the predicted class matched the ground-truth label.

5. Evaluation Results and Discussion

The items compared in this experiment are shown in Table4. As an example of the results, the performance using ROI-divided facial images from the SAMM dataset is shown in Table5, and the performance using full-face images is shown in Table6.

TABLE 4. Comparison Items.

1	Facial region used Full face or ROI
2	Input frames Apex frame or 3frame
3	Fine-tuning with CK+ With or without
4	Use of ABN With or without
5	Use of ArcFace With or without

TABLE 5. Comparison of results using ROI. []

Use of ABN	Network	2D (Apex)	3D3F
w/o ABN	ResNet	62.83	62.83
	CK+ Fine-tuning	61.17	62.90
	w/ ArcFace	62.00	64.50
w/ ABN	ResNet	64.53	**66.17**
	CK+ Fine-tuning	62.83	**65.40**
	w/ ArcFace	64.53	**65.33**

TABLE 6. Comparison of results using the full face. []

Use of ABN	Network	2D (Apex)	3D (3F)
w/o ABN	ResNet	62.83	60.37
w/ ABN	ResNet	63.70	62.83

First, a comparison of the results using ResNet in Table 5 and Table 6 reveals that using the ROI as input yields higher accuracy than using the full face. This may be because micro-expressions involve subtle movements, and focusing on specific regions such as ROIs makes it easier to extract relevant features than when using the entire facial image. In the comparison based on input frames, no clear advantage was observed when using the full face. However, when using ROIs, the approach that utilized three frames resulted in higher accuracy. This may be because ROIs, being cropped facial regions, emphasize inter-frame changes more effectively than the full-face input. Next, according to Table 5, no improvement in accuracy was observed when fine-tuning was conducted using the CK+ dataset. However, since previous studies have reported accuracy improvements through fine-tuning with facial expression datasets, it is possible that the network used in this study was not well-suited for such fine-tuning.

In the comparison based on the presence or absence of ABN, all networks showed improved accuracy when ABN was implemented. This improvement is likely attributable to ABN assigning greater attention to salient facial regions, thereby facilitating the detection of subtle changes characteristic of micro-expressions.

Finally, in the comparison involving ArcFace, no clear advantage of ArcFace was observed. Although ArcFace is a type of metric learning and was used as a loss function in this study, it had little impact on performance. Therefore, we consider that in micro-expression-based emotion recognition using deep learning, the feature extraction process plays a more crucial role.

6. Conclusion

In this study, we investigated the effectiveness of deep learning models for emotion recognition based on micro-expressions. The experiment involved five comparative conditions: facial region (full face vs. ROI), input frames (apex frame vs. three frames), fine-tuning using a facial expression dataset (CK+), the use of Attention Branch Network (ABN), and the use of ArcFace.

As a result, no significant improvement in accuracy was observed from the number of input frames, fine-tuning with the facial expression dataset, or the use of ArcFace. However, the use of ROI as input and the implementation of ABN contributed to an improvement in recognition accuracy.

While previous studies demonstrated the effectiveness of ROI in emotion recognition using handcrafted features, our results indicate that ROI-based division of facial images is also effective in deep learning models. Furthermore, the use of ABN, which assigns greater weight to regions of interest, improved accuracy, demonstrating its effectiveness in micro-expression recognition as well.

In future work, we plan to improve the emotion classification performance by combining handcrafted features with those extracted using deep learning.

References

[1] Mammadzada Sevinj Salim, "Verbal and Non-Verbal Communication in Linguistics", International Journal of Innovative Technologies in Social Science, Vol. 38, No. 2, 2023, doi:10.31435/rsglobal_ijitss/30062023/8003.

[2] A. Mehrabian and J. A. Russell, "An approach to environmental psychology", The MIT Press, Cambridge, MA, USA, 1974.

[3] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression", Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR2010), pp.94–101, 2010.

[4] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset", IEEE Trans. Affect. Comput., Vol. 9, No. 1, pp.116–129, 2018.

[5] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement collection and baseline", Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognitpp. 1–62013

[6] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation", Plos One, Vol.9, No.1, pp.1–8, 2014.

[7] S. -J. Wang, B. -J. Li, Y. -J. Liu, W. -J. Yan, X. Ou, X. Huang, F. Xu, and X. Fu, "Micro-expression recognition with small sample size by transferring long-term convolutional neural network", Neurocomputing, Vol. 312, pp. 251–262, 2018.

[8] K. Kato, H. Takano, M. Saiko, M. Kubo, and H. Imaoka, "Comparative study of feature extraction method for emotional classification by micro-expression," Proc. 2021 APSIPA ASC, pp. 1781–1785, 2021.

[9] D. E. King, "Dlib-ml: A machine learning toolkit", J. Mach. Learn. Res., Vol. 10, pp. 1755–1758, 2009.

[10] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of a attention mechanism for visual explanation," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 10705-–10714, 2019.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR2019), 2019, doi: 10.1109/CVPR.2019.00482.