2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897786

CONVOLUTIONAL NEURAL TREE FOR VIDEO-BASED FACIAL EXPRESSION RECOGNITION EMBEDDING EMOTION WHEEL AS INDUCTIVE BIAS

Ryo Miyoshi[†] *Shuichi Akizuki*[†] *Kensuke Tobitani*[⋆] *Noriko Nagata*[‡] *Manabu Hashimoto*[†]

[†] Chukyo University, Japan, [⋆] Kwansei Gakuin University, Japan, [‡] University of Nagasaki, Japan

ABSTRACT

In this study, we propose a video-based facial expression recognition method that utilizes the "emotion-wheel model." In research on emotions in the field of psychology, a model has been proposed in which basic emotions are arranged in a circular pattern, such as where "happiness" and "sadness" are opposites. Therefore, we utilized this knowledge as an inductive bias to improve accuracy by embedding features that are consistent with the emotion model in the process of class identification for recognizing facial expressions in videos. As a result of a performance evaluation using the CK+, MMI, and AFEW datasets, the recognition rates of the proposed method are 98.78%, 81.95%, and 55.31% for each dataset, which are 3.67%, 6.34%, and 4.9% higher than the baseline method, respectively. Furthermore, the proposed method outperforms the state-of-the-art method on MMI and AFEW.

Index Terms— Facial expression recognition, Emotion wheel, Deep learning, Tree structure, Emotion model

1. INTRODUCTION

Estimating emotions by recognizing facial expressions has attracted a lot of attention because it is expected to be done in a wide range of fields such as human-computer interaction and mobility from the viewpoint of naturalness. However, the task of facial expression recognition remains a challenging problem because of individual differences in facial expressions, such as variations in the ways and intensity of expressions.

In recent years, facial expression recognition using video has attracted much attention [1], and methods have been proposed to acquire features related to facial expressions independently of individuals. The method in [2] uses both facial landmarks and face images. The method in [3] disentangles facial features and expression features. These methods focus more on the movement of facial muscles to obtain general features of facial expressions.

In the field of psychology, two emotion models have been proposed. One is a model of people's basic emotions arranged in a circle, in which, for example, "happiness" and "sadness" emotions are located at opposite ends [4]. The other is a model of two-dimensional valence and arousal (VA) space (valence indicates how positive or negative an emotional state is, and arousal indicates how passive or active it is)[5]. In addition, each emotion has similarities (anger, disgust, fear, and sadness are negative emotions) and opposites (happiness and sadness). Moreover, it has been reported that it is effective to use emotion models for facial expression recognition [6, 7]. In [6], it was shown that the classification performance could be improved by using a deep learning model trained to estimate VA as a pre-training model for facial expression recognition models. In addition, in [7], it was reported that learning facial expression recognition and VA estimation simultaneously improves classification performance. From these studies, it is expected that the emotion model can be utilized for facial expression recognition to obtain generic features for facial expression classification. However, both studies require ground truth such as VA that quantifies the emotion model. However, only the expression class is given in many datasets for facial expression recognition, and a more detailed ground truth as VA is not given.

In this study, we aim to improve the accuracy of facial expression recognition by incorporating psychological knowledge called the "emotion-wheel model" into a facial expression recognition algorithm as an inductive bias, without using information other than the expression class such as VA. Specifically, we propose a tree structure deep learning model to achieve effective feature embedding based on the emotion-wheel model in the processing of feature embedding and class identification for video.

Our main contributions are summarized as follows.

- We propose a deep tree structure facial expression method that incorporates psychological knowledge, that is, the emotion wheel, as an inductive bias.
- The performance of the proposed method is compared on the public datasets CK+, MMI, and AFEW, and it outperforms the baseline method on all datasets. Furthermore, it outperforms the state-of-the-art method on MMI and AFEW.

This research was partially supported by the Center of Innovation Program from the Japan Science and Technology Agency (JST).

Fig. 1. Plutchik's wheel of emotions

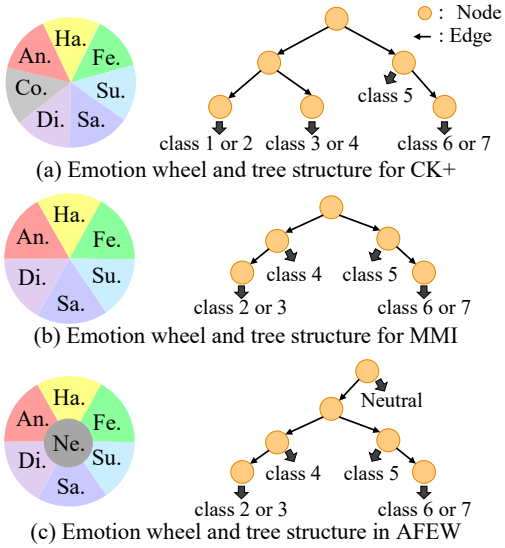

Fig. 2. Model of emotions and tree structure for each data set

2. FACIAL EXPRESSION RECOGNITION METHOD BASED ON EMOTION MODEL

2.1. Emotion model and tree structure

In this study, we use the model of Plutchik's wheel of emotions [4](Fig. 1), which consists of eight basic emotions, as a base. Here, the expression label assigned to each facial expression recognition dataset is slightly different. Therefore, we vary the emotion wheel model in accordance with the expression label given to each dataset. In this study, we use the public datasets CK+[8], MMI[9], and AFEW [10]. Figure 2 shows the emotion model and tree structure for each dataset. We vary Plutchik's emotion model in accordance with the facial expressions given in each dataset. The tree structure in this study has height $T = \lfloor (C+1)/2 \rfloor$, number of nodes $N = C - 1$, and number of edges $E = C - 2$, where C is the number of facial expression classes. Here, there is a neutral class in AFEW. Neutral is not an emotion, so it is always classified in the first branch of the tree structure. The class ID output by each node in the tree depends on how the emotion model is divided.

2.2. Proposed convlutional classification tree

An overview of the proposed method is shown in Fig. 3. This method consists of a feature extractor, which extracts features from video, and a convolutional classification tree (CCT), which embeds features and classifies facial expressions based on the above emotion model. First, the proposed method extracts a feature map by a feature extractor inputting a video with face regions. Then the feature map is input to the CCT, which recognizes feature embedding and facial expressions based on the emotion model.

The CCT learns the relationship between emotions based on the emotion model by learning a model that divides and hierarchizes the model, as shown in the upper left of Fig. 3. The convolutional classification node (CCN), a component of the CCT, consists of an attention module and a perception module, as shown in Fig. 4. The attention module is a module based on the attention branch of the attention branch network [11]. This module consists of a conv block, batch norm (BN), global-average-pooling (GAP), fully connected layer (FC), and sigmoid. The conv block is composed of two convolutional layers, ReLU and BN. The perception module is composed of a conv block, GAP, and FC. In the CCN, the feature map is first input to the attention module, and then the attention map and class likelihood are calculated. Then, the element product of the obtained attention map and the feature map is taken, and the attention map is reflected in the feature map. The feature map is then input to the perception module to calculate the class likelihood. If we denote the feature map as F and the attention map as F_a, the feature map passing from the parent node to the child nodes branching to the left and right is as follows.

$$F_l = F \circ F_a \tag{1}$$

$$F_r = F \circ (1 - F_a) \tag{2}$$

Here, $\circ$ represents Hadamard product, F_l is the feature map passed to the left child node, F_r is the feature map passed to the right child node. As shown in the above equation, when a child node branches to the left or right, a different feature map is passed to each node. In other words, the attention map plays the role of a router function. The CCT is trained to classify two or three classes at each node. Therefore, the loss function of the CCT is as follows.

$$L_1 = \sum_{i=1}^{M} \sum_{j=1}^{N} L_p^j(x_i) + \sum_{i=1}^{M} \sum_{j=1}^{N} L_a^j(x_i) \tag{3}$$

Here, M is the number of samples, N is the number of nodes, x_i is the ith input sample, $L_p^j(x_i)$ and $L_a^j(x_i)$ are the learning loss between the class label and probability score obtained from the perception module and attention module the class label corresponding to the jth node, respectively. To

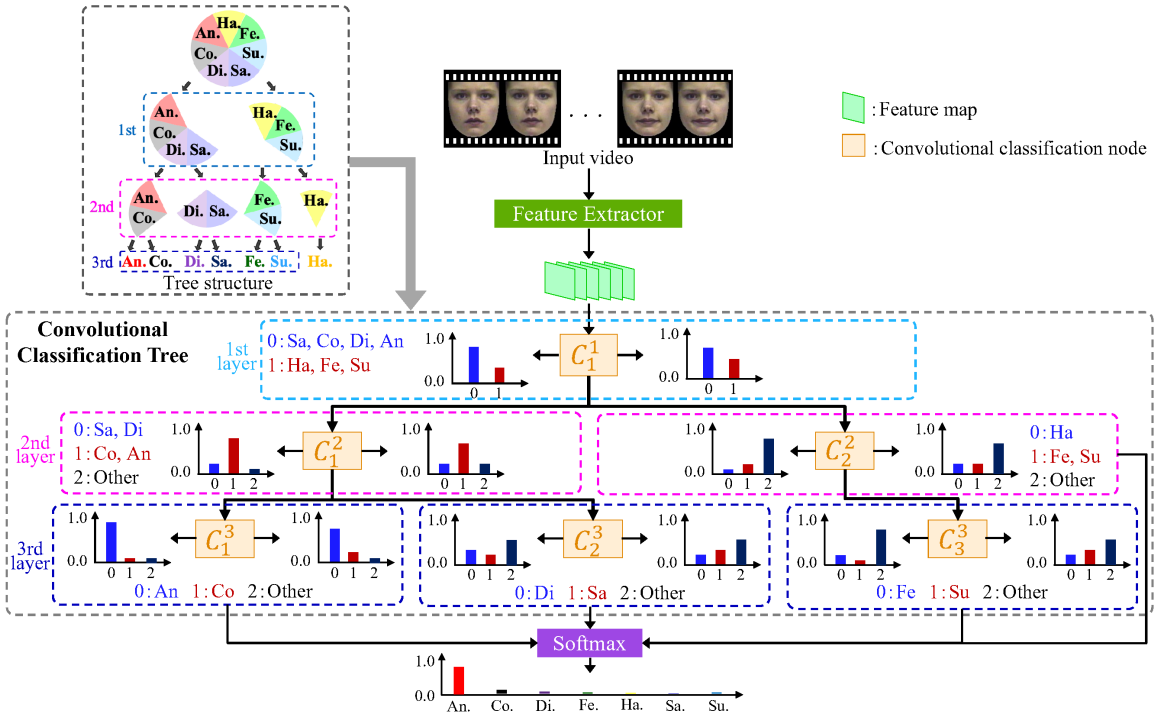

Fig. 3. Overview of proposed facial expression recognition method. This figure shows tree structure of proposed method for CK+. Tree structure is defined for each dataset.

reduce the variance of features due to the tree structure, ArcFace [12], a kind of metric learning, is introduced as a loss function for each node.

Furthermore, the likelihoods of the expression classes obtained from each module of each node are concatenated and the learning error is computed between the expression probability score calculated by the Softmax function and the class label.

$$L_2 = \sum_{i=1}^{M} L_{c_p}(x_i) + \sum_{i=1}^{M} L_{c_a}(x_i) \tag{4}$$

Here, L_{c_p} and L_{c_a} are the cross-entropy loss obtained from perception module and attention module, respectively.

The final learning error can then be expressed as follows.

$$L = L_1 + L_2 \tag{5}$$

3. EXPERIMENTS

3.1. Datasets used for the experiment

In this experiment, CK+ [8], MMI [9], and AFEW [10] were used. For CK+ and MMI, the performance was evaluated by subject-independent 10-fold cross-validation as in previous studies [1].

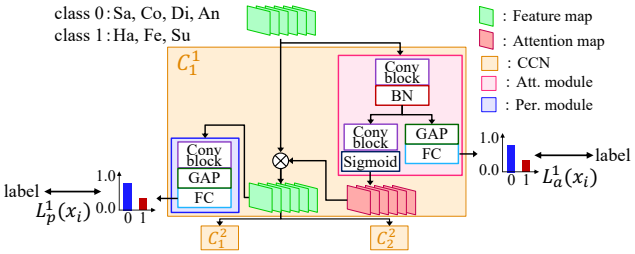

Fig. 4. Overview of convolutional classification node

CK+ [8] consists of 327 videos taken in the laboratory and collected from 118 subjects. The videos consist of a sequence that changes from a neutral expression to a scene with a peak expression. The expressions included in this dataset are anger, contempt, disgust, fear, happiness, sadness, and surprise.

MMI[9] consists of 326 videos featuring 32 participants. Of these, 213 videos are labeled anger, disgust, fear, happiness, sadness, and surprise. The videos start with a neutral expression. The expression peaks in the middle of the videos and returns to a neutral expression at the end.

AFEW[10] is the closest real-world dataset that contains video clips collected from various movies and TV dramas. It contains seven facial expressions: anger, disgust, fear, happiness, sadness, surprise, and neutral. This dataset is divided into training data (773 samples), validation data (383 sam-

Table 1. Experimental results on the CK+

Method	Accuracy
PHRNN-MSCNN（2017）[2]	98.50%
C3D-GRU（2019）[13]	97.25%
CTSLSTM（2019）[14]	93.90%
SC（2019）[15]	97.60%
LBVCNN（2019）[16]	97.38%
Enhanced ConvLSTM（2021）[17]	95.72%
MIC（2021）[3]	**99.71%**
R(2+1)D [18] + FC（Baseline）	95.11%
R(2+1)D [18] + CCT-Att. (Ours)	98.47%
R(2+1)D [18] + CCT-Pec. (Ours)	98.78%
R(2+1)D [18] + CCT-Ens. (Ours)	98.78%

Table 2. Experimental results on the MMI

Method	Accuracy
3D CNN-DAP（2014）[19]	63.40%
PHRNN-MSCNN（2017）[2]	81.18%
CNN+LSTM（2019）[20]	78.61%
CTSLSTM（2019）[14]	78.40%
MIC（2021）[3]	81.29%
R(2+1)D [18] + FC（Baseline）	75.61%
R(2+1)D [18] + CCT-Att. (Ours)	80.98%
R(2+1)D [18] + CCT-Pec. (Ours)	**81.95%**
R(2+1)D [18] + CCT-Ens. (Ours)	**81.46%**

ples), and test data (653 samples). The data in the three sets are guaranteed to be from mutually exclusive movies and actors. Since the labels of the test data are not publicly available, we train the model on the training data and evaluate its performance on the validation data. Note that the validation data is not used in the training phase for tuning parameters and hyperparameters.

3.2. Experimental setting

In this study, the baseline method consists of a feature extractor with R(2+1)D [18] and a discriminator with three fully connected layers. R(2+1)D was pretrained by Kinetics-400 [23]. Face images aligned by OpenFace [24] were used as input to the model. The size of the aligned face images was 112×112. The length of the video input to the model was 16 frames. In the evaluation of the proposed method, experiments were conducted for all ways of dividing the emotion model for each dataset. For each dataset, there were 14 ways to divide the emotion model for CK+, 6 ways for MMI, and 12 ways for AFEW. The recognition rate of the proposed method was evaluated based of the scores obtained from the attention module and perception module (CTT-Att., CTT-Per., respectively) and the sum of the two scores (CTT-Ens.).

Table 3. Experimental results on the AFEW

Method	Accuracy
Unidirectional LSTM（2017）[21]	48.60%
CTSLSTM（2019）[14]	51.20%
C3D-GRU（2019）[13]	49.78%
Former-DFER（2021）[22]	50.92%
MIC（2021）[3]	53.72%
R(2+1)D [18] + FC（Baseline）	50.41%
R(2+1)D [18] + CCT-Att. (Ours)	**54.77%**
R(2+1)D [18] + CCT-Pec. (Ours)	**55.04%**
R(2+1)D [18] + CCT-Ens. (Ours)	**55.31%**

3.3. Experimental results

We compared our method with state-of-the-art methods that use video as input and no additional training data to ensure a fair experiment. We also report the most accurate results for each dataset when the emotion model is split in all the splitting patterns.

Results on CK+. The recognition rate of the proposed method for CCT-Per. and CCT-Ens. was 98.78%, which was 3.67% higher accuracy than the baseline method. **Results on MMI**. The recognition rate of the proposed method for CCT-Per. was 81.95%, which was 6.34% more accurate than the baseline method. Furthermore, the proposed method outperformed the state-of-the-art method by 0.66%, which is a slight improvement. **Results on AFEW**. The recognition rate of the proposed method was 53.31% for CCT-Ens. which was 4.9% more accurate than the baseline method. Furthermore, the proposed method outperformed the state-of-the-art method by 1.59%.

As a result of these experiments, we confirmed that the accuracy of the proposed method is significantly better than that of the baseline method on all datasets. Furthermore, the proposed method outperforms the state-of-the-art method on MMI and AFEW, confirming that using emotion models for feature embedding as inductive bias is effective for obtaining generic features in facial expression recognition.

4. CONCLUSION

In this study, we proposed a tree-structure facial expression recognition method for feature embedding based on the emotion model proposed in the field of psychology. In experiments, we evaluated the performance of the proposed method on the public datasets CK+, MMI, and AFEW, and we confirmed that the proposed method significantly improves the accuracy over the baseline method on all datasets. Furthermore, the accuracy of our method was better than the state-of-the-art method on MMI and AFEW. This confirms that using emotion models as inductive bias in facial expression recognition effectively acquires generic features.

5. REFERENCES

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[2] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.

[3] X. Liu, L. Jin, X. Han, and J. You, "Mutual information regularized identity-aware facial expression recognition in compressed video," *Pattern Recognition*, vol. 119, pp. 108105, 2021.

[4] R. Plutchik, *The emotions*, University Press of America, 1991.

[5] J. A. Russell, "Evidence of convergent validity on the dimensions of affect.," *Journal of personality and social psychology*, vol. 36, no. 10, pp. 1152, 1978.

[6] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 907–929, 2019.

[7] D. Kollias and S. Zafeiriou, "Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework," *arXiv preprint arXiv:2103.15792*, 2021.

[8] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and image understanding*, vol. 91, no. 1-2, pp. 160–187, 2003.

[9] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE International Conference on Multimedia and Expo*, July 2005, pp. 5 pp.–.

[10] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 03, pp. 34–41, 2012.

[11] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10705–10714.

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[13] M. K. Lee, D. Y. Choi, D. H. Kim, and B. C. Song, "Visual scene-aware hybrid neural network architecture for video-based facial expression recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

[14] M. Hu, H. Wang, X. Wang, J. Yang, and R. Wang, "Video facial emotion recognition based on local enhanced motion history image and cnn-ctslstm networks," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 176–185, 2019.

[15] M. Verma, H. Kobori, Y. Nakashima, N. Takemura, and H. Nagahara, "Facial expression recognition with skip-connection to leverage low-level features," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 51–55.

[16] S. Kumawat, M. Verma, and S. Raman, "Lbvcnn: Local binary volume convolutional neural network for facial expression recognition from image sequences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[17] R. Miyoshi, N. Nagata, and M. Hashimoto, "Enhanced convolutional lstm with spatial and temporal skip connections and temporal gates for facial expression recognition from video," *Neural Computing and Applications*, pp. 1–12, 2021.

[18] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[19] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian conference on computer vision*. Springer, 2014, pp. 143–157.

[20] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 223–236, 2017.

[21] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 569–576.

[22] Z. Zhao and Q. Liu, "Former-dfer: Dynamic facial expression recognition transformer," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1553–1561.

[23] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[24] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.