Develop Method to Efficiently Apply Image-Based Facial Emotion Classification Models to Video Data

Hee Min Yang[1], Joo Hyun Lee[1], and Yu Rang Park[1,2(✉)]

[1] Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, South Korea
yurangpark@yuhs.ac

[2] Department of Artificial Intelligence, Yonsei University, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea

Abstract. The ability to recognize emotions through facial cues, in childhood, is helpful for social interactions. Image-based facial emotion recognition models need low computing power, but cannot accept sequential information from video data. Conversely, video-based facial emotion recognition models require high computational power, so it cannot be easily applied in a low computing environment. In this paper, we propose a method that classifies the emotion from facial expression video data by applying threshold using an image-based model. The proposed method improves the accuracy of 3.67%, 24.74%, and 15.13% for each video dataset by reducing the non-emotion in the video and responding more sensitively to the expressed emotion than other methods that simply select the most frequent emotion in the video. The results of the study showed the threshold method can improve the performance of emotion classification without modifying the facial emotion classification model.

Keywords: Facial Emotion Recognition · Deep Learning · Computer Vision · Child

1 Introduction

The ability to recognize and express one's own and others' emotions based on facial cues is directly linked to an individual's ability to interact with others. This skill is even more important in childhood, when the first social interactions occur, before speech is fully developed [1, 2]. The inability to recognize facial emotions is closely linked to child development problems [3]. It can be one of the reasons for developmental delay in basic social skills needed to adapt to social life. Low emotional knowledge in children is associated with negative outcomes, including poor social functioning, low academic achievement, and internalizing/externalizing behavior problems [1] [4, 5].

Facial emotion recognition has been reported to be helpful for children with developmental behavioral conditions that make it difficult to recognize emotions, such as

A. Cunha et al. (Eds.): MobiHealth 2023, LNICST 578, pp. 353–360, 2024.
https://doi.org/10.1007/978-3-031-60665-6_26

autism spectrum disorder (ASD) [6–8]. However, most emotion recognition studies focus on adult data, so when applied to children's faces, performance decreases. Park et al. (2022) showed that emotion classification after splitting of child and adult facial image data achieved an accuracy improvement of 22.4% compared to before splitting [9]. As a result of the emotion classification of children's facial expressions by machine learning, children tend to be highly expressive in terms of positive emotions and ambiguous in terms of negative emotions [10, 11]. For example, when children are ambiguously angry in terms of negative emotions, they sometimes show neutral expressions, making it difficult to know what emotions the child is experiencing.

Video-based facial emotion recognition models require high computational power because they use the sequential information of video data [12, 13]. Conversely, image-based models can also be implemented at low computational power [14], but have the problem of not accepting sequential information of a video data. In an environment where computational power is limited, such as mobile devices, a method is needed to efficiently apply image-based models to video data.

Therefore, we aimed to (1) develop a method based on the classification of human emotions in the video to select the representative emotion of a video using an image-based facial emotion classification model for children, and (2) evaluate the applicability and effectiveness of the developed method by applying it to a real public video dataset.

2 Materials and Methods

2.1 Dataset

We used the two child facial expression video datasets, the DuckEES databases [15] and the LIRIS Children Spontaneous Facial Expression Video [16]. The DuckEES dataset contains facial expressions of emotion created by children and teenagers between the ages of 8 and 18. The video dataset contains 251 videos with six facial expressions of emotion (happy, sad, fear, disgust, pride, embarrassment) and 'neutral', and the emotion labels were evaluated by 36 human cross-validators. We use 121 videos with an accuracy of 0.7 or higher based on cross-validation, which is the cutoff for the final dataset presented by the DuckEES researchers, excluding embarrassment and pride videos for which the image-based classification model did not learn about these labels. Of the 121 videos, there are 36 happy, 17 sad, 18 fear, 20 disgust and 30 neutral videos. The video was recorded in 25 frames. The LIRIS-CSE dataset contains 180 videos that have six facial emotion expression labels (happy, sad, angry, fear, surprise and disgust) and the participants' mean age is 7.3 years. The 180 videos include 61 happy, 26 sad, 1 angry, 32 fear, 51 surprise and 9 disgust. The video was also recorded in 25 frames. And we created the combined dataset using DuckEES and LIRIS-CSE as described above. There is no additional processing and they were simply used together. For this study, we chose the following seven labels for the two datasets: happiness, sadness, anger, fear, surprise, disgust, and neutral.

2.2 Proposed Method: Threshold

The proposed method, threshold, selects the representative emotion of a video using an image-based facial emotion classification model, which consists of two parts (Fig. 1). In

part A: the part of classifying the emotion of each frame in a video by the classification model**,** the method used facial video data with the true label about the facial emotion, which was labeled by the labeler with cross-validation. The facial video data is divided into frames, and the frames are used as a kind of image data as input to the image-based emotion classification model. Using the frame images of the video, the facial emotion classification model predicts the emotion for each frame. This procedure generates tabular data composed of the predicted emotion results for each frame, which is used in the next part. In part B: the part of determining the representative emotion for a video, the number of emotions for each frame of the facial video is aggregated and sorted in the order of the most emotions. Threshold is the ratio of sensitivity to non-neutral emotions. When Eq. (1) is satisfied,

$$\max emotion(\text{without neutral}) \geq (threshold \times numberofframetotal) \tag{1}$$

the representative emotion of the video becomes the emotion of the maximum number. If Eq. (1) is not satisfied, the most frequent emotion or neutral is selected as the representative emotion. The accuracy of the current threshold is calculated by comparing the selected representative emotion to the true label on each video. The architecture for this study is summarized and visualized in Fig. 1.

2.3 Evaluation Metrics

We use accuracy and F1-score as the evaluation metrics. Accuracy measures how close the predicted value is to a true or accepted value of being true. The F1-score is an integrated indicator of how accurately the model predicted and whether the model actually captured all important outcomes.

2.4 Experimental Setting

We conducted our experiments on a desktop computer equipped with an Intel Core i7–9700 CPU, 16 GB of RAM, and a 256 GB SSD. GPU processing was not used for this project, only CPU power was used.

We used the image-based model [17] pre-trained on the FER2013 dataset (Facial Expression Recognition 2013 Dataset) [18] using Mini-Xception as the model architecture, a miniature version of Xception [19]. And we set the threshold to increase from 0.05 to 1.00 in 0.05 increments to find the optimal value.

There are two settings for categorizing emotions, the broad emotion setting and the specific emotion setting. In the broad emotion setting, the emotion is categorized by three labels: negative, positive, and neutral, to clearly compare non-emotion and emotion [10, 11]. On the other hand, the specific emotion setting is for evaluating the robustness of our method in more complicated classification tasks. The emotion has seven labels: happiness, sadness, anger, fear, surprise, disgust, and neutral [20]. In each setting, we compare our threshold method to the baseline Top 1 method using evaluation metrics. Top 1 is a method that selects the most frequent emotion in each frame of a facial expression video as the representative emotion.

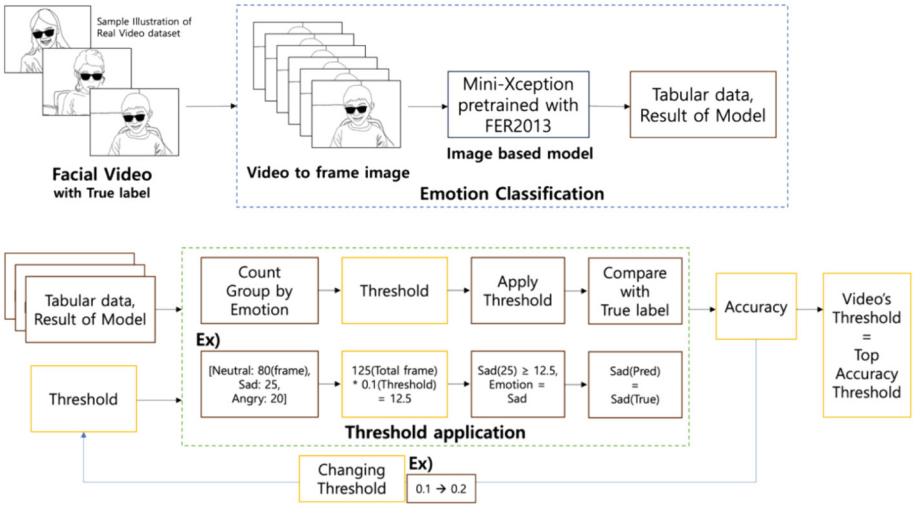

Fig. 1. The figure shows the architecture of the threshold method. The architecture has Part A, where an image-based model classifies the emotion for each frame (i.e., image) of the facial video, and Part B, where the classified emotion results are used to determine the threshold for selecting the representative emotion for that facial video.

2.5 Method Validation

We used 5-fold cross-validation to validate the proposed method. In this method, the original data is divided into 5 subsets of equal size. Then, one of these subsets is used as validation data, and the rest of the subsets are used as training data to learn the model. This process is repeated 5 times, with different subsets selected as validation data each time. As a result, we estimate the performance of the final model by averaging the model performance of 5 times.

3 Result

3.1 Broad Emotion Test (3 Label)

We expanded the range of emotions and tested three broad ranges of emotions. This is to evaluate negative emotions by uniting them. Experimental performance is measured by the average accuracy from 5-fold cross-validation. For the DuckEES data set, the optimal threshold was 0.15, 0.05 for LIRIS-CSE, and 0.05 for DuckEES+LIRIS-CSE. The accuracy of the method applying the optimal threshold to each data set is 0.8545, 0.7468, and 0.7813. Top 1 accuracies were 0.8087, 0.4944, and 0.6211. That is, the threshold showed improvements in accuracy of 0.0458, 0.2474, and 0.1602 compared to the baseline top 1. It can be seen that the threshold method outperforms top1 in DuckEES and DuckEES+LIRIS-CSE (Table 1).

We compare the threshold (with the optimal value of 0.05) and the top 1 with the combined dataset (DuckEES+LIRIS-CSE) for each emotion with the performance of

Table 1. Evaluation result of Top1 and Threshold based on 3 labels (Neutral, Positive, Negative).

Dataset	Threshold		Top 1 Accuracy	Difference*
	Accuracy	Optimal threshold		
DuckEES [15]	0.8545	0.15	0.8087	0.0458
LIRIS-CSE [16]	0.7468	0.05	0.4944	0.2474
DuckEES+LIRIS-CSE [15, 16]	0.7813	0.05	0.6211	0.1602

* Difference is the difference between the accuracy of Threshold and the accuracy of Top1.

accuracy and F1-score. As a result, the threshold has higher accuracy and F1-score values than the top 1 in all three emotions. In particular, in F1-score, the threshold is 0.1558, 0.0554, 0.0854 and 0.1598 higher in the order of negative, positive, neutral and average of three emotions (Table 2).

Table 2. Comparison results of top 1, threshold (with optimal threshold value 0.05) and threshold with median value for the combined dataset of DuckEES [15] and LIRIS-CSE [16]. The higher results are highlighted in bold.

	Emotion							
	Negative		Positive		Neutral		Average of 3 emotions	
	Threshold(0.05*)	Top 1	Threshold	Top 1	Threshold	Top 1	Threshold	Top 1
Accuracy	**0.7879**	0.6844	**0.9226**	0.8970	**0.8519**	0.6611	**0.8540**	0.7475
F1-score	**0.8000**	0.6442	**0.8878**	0.8324	**0.4054**	0.3200	**0.7811**	0.6213

* The optimal value of the threshold from the combined dataset, each threshold value is 0.05 in this table.

3.2 Specific Emotion Test (7 Label)

We experimented with seven labels from two video dataset. Experimental performance is measured by the average accuracy from 5-fold cross-validation. Threshold was tested by raising it to 0.05 units from 0.05 to 1.00 to find the optimal value. For the DuckEES data set, the optimal threshold was 0.20, LIRIS-CSE was 0.05, and DuckEES+LIRIS-CSE was 0.20. The accuracy of the method applying the optimal threshold to each data set is 0.6635, 0.4216, and 0.5153, which is higher than the top1 accuracy of 0.6440, 0.3222, and 0.4519. That is, the threshold showed accuracy improvements of 0.0195, 0.0994, and 0.0634 for the Top 1 baseline. The experimental results can be seen in Table 3.

4 Discussion

In this study, we proposed to use thresholds in an image-based classification model to select representative emotions from child facial videos. In particular, we applied the method to a real public video dataset to evaluate its feasibility and efficacy. As a result of

Table 3. Evaluation result of Top1 and Threshold based on 7 labels (Neutral, happy, sad, angry, fear, surprise and disgust).

Dataset	Threshold		Top 1 Accuracy	Difference*
	Accuracy	Optimal threshold		
DuckEES [15]	0.6635	0.30	0.6440	0.0195
LIRIS-CSE [16]	0.4216	0.05	0.3222	0.0994
DuckEES+LIRIS-CSE [15, 16]	0.5153	0.20	0.4519	0.0634

* Difference is the difference between the accuracy of Threshold and the accuracy of Top1.

the evaluation by the two datasets and the combined dataset, our method shows higher accuracy than the so-called Top 1, which is a method of selecting the most frequent value among the emotions in the video. The same image-based emotion classification model was used, but the accuracy performance of our method was outperformed.

Performance comparison between the threshold (with the optimal threshold value) and the top 1, the F1-score difference of negative emotion between the two methods is larger than that of positive emotion, which shows that when the Top 1 method selects the most frequent emotion as the representative emotion for the facial emotion video, the actual negative emotion tends to be buried in Neutral due to the many Neutral frames in the video [10, 11]. Threshold, in contrast, finds the emotions that have been buried in Neutral and also shows the result of preserving Neutral as itself in the Neutral video.

Our threshold method captures the instantly appearing facial emotions and selects them as the representative emotions of the video more accurately than before. This is because studies on human emotion recognition have shown that people recognize emotions by estimating the context of the situation [20, 21], and studies on the importance of different parts of the face for emotion recognition have found that sufficient information is needed in the important areas (the study identified the eyes and mouth) for emotion recognition [22, 23]. So, like humans, our method captures emotional expressions that have strong facial information and context (laughing, frowning, or crying) that appear for a while, rather than the basic neutral state in the video, to determine what emotion the video has. And the Threshold method improves performance by efficiently interpreting the results of the classification model without having to develop or replace additional models. So far, regardless of each facial emotion recognition model's type, the proposed method can be easily applied, and can be expected to achieve higher performance in the research results.

This proposed method has a limitation that requires a process of finding an appropriate threshold for the recording environment and the participants' faces. Therefore, prior data for the threshold search is required to be used in the actual environment. In addition, since this study used facial data from children, further research with adult data is needed to verify whether adults have the same substantial performance improvement.

5 Conclusions

The results of the threshold-based emotion classification method proposed in this study suggest that emotion classification performance can be improved without modifying the classification model itself, especially in children's data where emotion classification is difficult.

Acknowledgement. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP 2022-0-00064) grant funded by the Korea government(MSIT).

References

1. Izard, C.E.: Emotional intelligence or adaptive emotions? Emotion **1**, 249–257 (2001)
2. Barth, J.M., Andrea, B.: A longitudinal study of emotion recognition and preschool children's social behavior. Merrill-Palmer Q. (1982), 107–128 (1997)
3. Harms, M.B., Martin, A., Wallace, G.L.: Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies. Neuropsychol. Rev. **20**, 290–322 (2010)
4. Trentacosta, C.J., Fine, S.E.: Emotion knowledge, social competence, and behavior problems in childhood and adolescence: a meta-analytic review. Soc. Dev. **19**(1), 1–29 (2010)
5. Ensor, R., Spencer, D., Hughes, C.: You feel sad? Emotion understanding mediates effects of verbal ability and mother–child mutuality on prosocial behaviors: findings from 2 years to 4 years. Soc. Dev. **20**(1), 93–110 (2011)
6. Happé, F., Frith, U.: Annual research review: towards a developmental neuroscience of atypical social cognition. J. Child Psychol. Psychiatry **55**(6), 553–577 (2014)
7. Hobson, R.P., Ouston, J., Lee, A.: Emotion recognition in autism: coordinating faces and voices. Psychol. Med. **18**(4), 911–923 (1988)
8. Carolien, R., et al.: Emotion regulation and internalizing symptoms in children with autism spectrum disorders. Autism **15**(6), 655–670 (2011)
9. Park, H., et al.: Facial emotion recognition analysis based on age-biased data. Appl. Sci. **12**(16), 7992 (2022)
10. Washington, P., et al.: Improved digital therapy for developmental pediatrics using domain-specific artificial intelligence: machine learning study. JMIR Pediatrics Parent. **5**(2), e26760 (2022)
11. Anwar, S., Milanova, M.: Real time face expression recognition of children with autism. Int. Acad. Eng. Med. Res **1**(1), 1–8 (2016)
12. Xia, X., Zhao, Y., Jiang, D.: Multimodal interaction enhanced representation learning for video emotion recognition. Front. Neurosci. **16**, 1086380 (2022)
13. Wei, Q., Huang, X., Zhang, Y.: FV2ES: a fully End2End multimodal system for fast yet effective video emotion recognition inference. IEEE Trans. Broadcast. **69**, 10–20 (2022)
14. Pandey, S., Sonakshi, H.: Facial emotion recognition using deep learning. In: 2022 International Mobile and Embedded Technology Conference (MECON). IEEE (2022)
15. Giuliani, N.R., et al.: Presentation and validation of the DuckEES child and adolescent dynamic facial expressions stimulus set. Int. J. Methods Psychiatric Res. **26**(1), e1553 (2017)
16. Khan, Rizwan Ahmed, et al.: A novel database of children's spontaneous facial expressions (LIRIS-CSE). Image Vis. Comput. **83**, 61–69 (2019)

17. Arriaga, O., Valdenegro-Toro, M., Plöger, P.:Real-time convolutional neural networks for emotion and gender classification. arXiv preprint arXiv:1710.07557 (2017)
18. Goodfellow, I.J., et al.: Challenges in representation learning: a report on three machine learning contests. In: Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3–7. Proceedings, Part III 20. Springer, Heidelberg (2013)
19. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
20. Ekman, P.: Universals and cultural differences in facial expressions of emotion. In: Nebraska Symposium on Motivation. University of Nebraska Press (1971)
21. Frijda, N.H.: Recognition of Emotion. Advances in Experimental Social Psychology, vol. 4, pp. 167–223. Academic Press (1969)
22. Birmingham, E., et al.: The moving window technique: a window into developmental changes in attention during facial emotion recognition. Child Dev. **84**(4), 1407–1424 (2013)
23. Kim, M., Cho, Y., Kim, S.-Y.: Effects of diagnostic regions on facial emotion recognition: the moving window technique. Front. Psychol. **13**, 966623 (2022)