Facial Emotion Recognition with Contextual Information and Model Scalability

Savina Jassica Colaco, Dong Seog Han*

School of Electronic and Electrical Engineering

Kyungpook National University, Daegu, Republic of Korea

savinacolaco@knu.ac.kr, *dshan@knu.ac.kr

Abstract

Gaining a deep understanding of human emotions from their point of view is essential in everyday social exchanges. Machines possessing this competence have the potential to interact more efficiently with humans. However, the incorporation of contextual information is crucial for understanding emotions, and by considering context, a wider range of emotional states can be deduced. This paper proposes an emotion recognition system incorporating facial landmarks and context-aware analysis. The context-aware model achieves an accuracy of up to 78.41%, surpassing models that exclusively rely on facial expressions. The significance of context in perceiving emotions is emphasized, as it allows for the identification of a broader range of emotional states and improves interactions between machines and humans.

Ⅰ. Introduction

Automatic or real-time emotion detection is utilized in several applications, such as human identification, healthcare, virtual reality, and cognitive research [1]. Deep learning methods have created numerous systems for recognising human emotions. However, the majority of identification models concentrate solely on facial data in order to identify distinct emotions. This paper employs a deep learning model to accurately detect emotions within the contextual information of images. Contextual information enhances the recognition of intricate emotions.

Ⅱ. Experiment & Discussion

1. Model

We developed a scalable model for recognizing emotions. This model uses facial landmarks and context-aware detection, further improved by

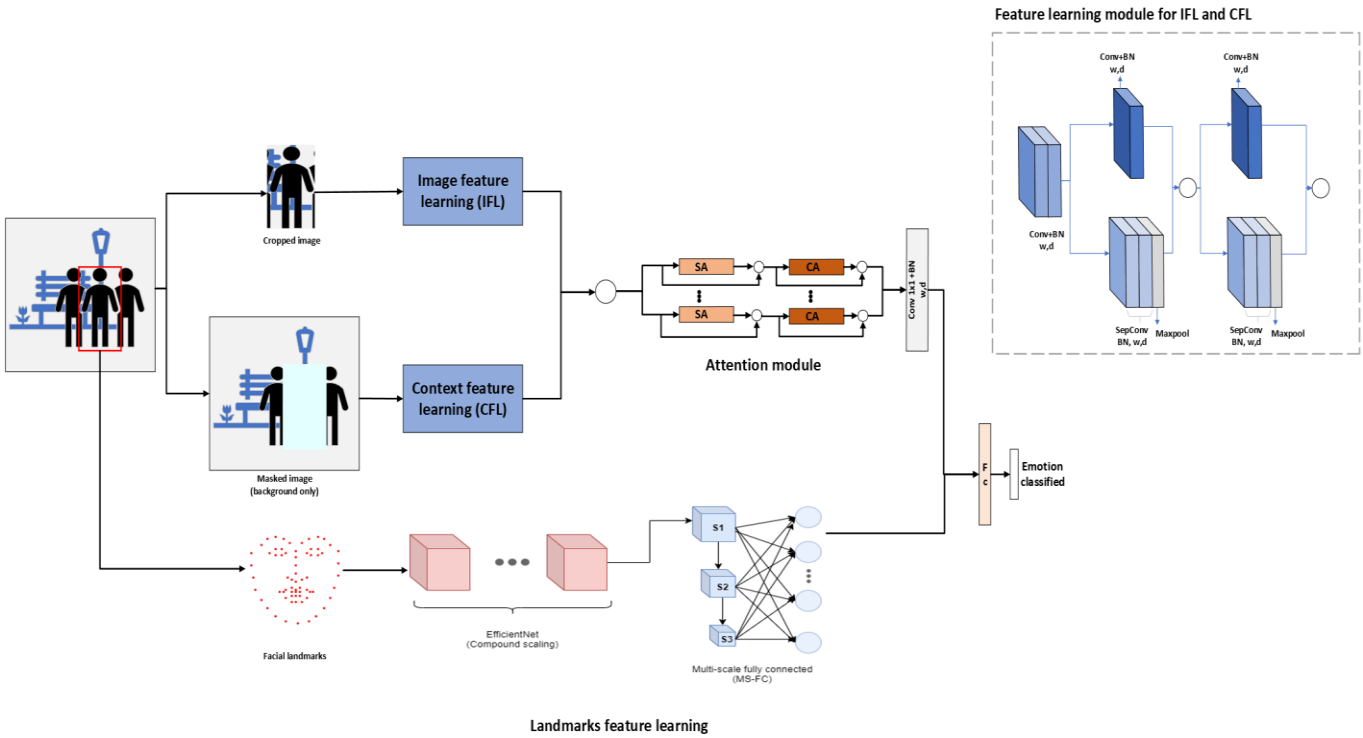

Fig 1: Context-aware emotion recognition framework

attention and multi-branch processes. This method utilizes both facial features and contextual information to enhance accuracy. The process commences by partitioning images into two distinct components: one that emphasizes the subject and another that pertains to the background. CNNs are employed to examine each component individually. The approach utilizes efficient mini xception layers to extract various aspects from subject- and background-focused images, reducing processing requirements. Furthermore, a customized version of the EfficientNet model [2] is employed for facial landmarks, enhancing the management of features using compound scaling. The system integrates characteristics extracted from the image, context, and landmarks and applies an attention module to prioritize significant sections for emotion analysis. This module employs spatial and channel attention techniques. In addition, the landmarks module has a multi-scale fully linked layer to consolidate and examine input at different scales. The model categorizes emotions by taking into account both facial expressions and context.

2. Experiment and Results

The EMOTIC dataset [3], specifically created to identify emotions in real-life situations, has a total of 23,571 images that have been meticulously annotated with detailed information about 34,320 persons. Every individual is classified into 26 distinct emotional categories, encompassing a variety of sentiments such as peace, happiness, anger, and sadness. In addition, emotions are assessed using continuous metrics of valence (degree of pleasantness), arousal (tendency to take action), and dominance (amount of control). The collection includes manually annotated body portions in photos scaled to 112 × 112 pixels.

Table 1. Comparison analysis with and without context models with respect to scaling factors

Model	Scaling	Parameters (millions)	Test Accuracy (%)
Without context	W=D=0.25	3.4M	75
	W=D=0.5	5M	67
	W=D=1	8M	69.97
With context	W=D=0.25	3.5M	77.72
	W=D=0.5	5M	77.3
	W=D=1	8M	78.41

The table 1 presents a comprehensive evaluation of emotion recognition models on EMOTIC dataset, investigating the influence of contextual information on their effectiveness. Models are evaluated using various scaling factors, such as width (w) and depth (d), with a particular emphasis on the number of parameters and the resulting test accuracy. The models without contextual data exhibit a decline in accuracy as the number of parameters grows. In contrast, models that incorporate contextual data not only have a little greater number of parameters but also exhibit a substantial enhancement in accuracy across all scaling levels. The highest performance is observed at the biggest scale (W=D=1), achieving a test accuracy of 78.41%. The presence of context is advantageous for improving the precision of emotion detection models, especially as the model becomes more intricate.

Ⅲ. Conclusion

This paper proposed emotion recognition model, consisting of contextual information and facial landmarks which improves the classification accuracy of the complex models. The deep learning model, consists of feature learning and attention modules to identify important features. The proposed model can be fine-tuned for further improvement with the help of inception modules. Since all features are not important for emotion recognition, concentrating on specific regions to detect basic, as well as complex emotions, can be achieved in the future.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2021R1A6A1A03043144).

REFERENCES

[1]B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," Pattern recognition, vol. 36, pp. 259-275, 2003.

[2]S. J. Colaco and D. S. Han, “Deep Learning-Based Facial Landmarks Localization Using Compound Scaling,” IEEE Access, vol. 10, pp. 7653- 7663, 2022.

[3]R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, “Context Based Emotion Recognition Using Emotic Dataset,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 11, pp. 2755- 2766, 2019.