Employee Emotion Recognition Method Based on Improved MobileNetV3

Xi Chen[1,a*], Miaoyun Hu[1,b], Xinle Zou[2,c], Yate Tan[2,d]

[1]Shunde Polytechnic, Foshan, China
[2]Guangdong Xi'an Jiaotong University Research Institute, Foshan, China

[a*]13825571131@139.com, [b]21457357@qq.com,
[c]zouxinyue202401@126.com, [d]3629416073@qq.com

Abstract. This paper has made improvements to the MobileNetV3 model, incorporating deep separable convolutions, inverted residual structures, and optimized time-consuming layer structures. Additionally, an improved attention mechanism has been proposed, utilizing a serial spatial channel attention mechanism. After multiple experiments, the improved model achieved an accuracy rate of 94.95% on the KDEF dataset, demonstrating that the enhancements have increased the accuracy of facial expression recognition.

Keywords: Employee Emotion Recognition Method, Improved MobileNetV3, facial expression recognition.

1 Introduction

Emotion recognition plays a pivotal role in assessing and maintaining the health and productivity of employees within an organization. It is a dynamic process that involves identifying and interpreting the emotional states of individuals, which can significantly influence their performance and well-being at work. The ability to recognize and respond appropriately to emotions is crucial for fostering a positive work environment, enhancing team collaboration, and ensuring that employees are engaged and motivated(Prasad, et al.,2023)[1]. When employees feel understood and their emotional needs are met, they are more likely to be productive, have lower levels of stress, and experience higher job satisfaction. Moreover, emotion recognition can be instrumental in early detection of burnout or mental health issues, allowing for timely interventions that can prevent more severe consequences(Jiang,et al., 2023)[2].

Despite its importance, the existing methods of emotion recognition have several limitations. Traditional approaches often rely on manual observation or self-reporting, which can be subjective and prone to inaccuracies. Furthermore, these methods may not be scalable or practical in large organizations where continuous monitoring of employee emotions is necessary(Bisht et al., 2022)[3]. Technological solutions, such as facial expression analysis and voice tone detection, have shown promise but are still in their infancy and have limitations in terms of accuracy and reliability. They may also

K. Zhang et al. (eds.), *Proceedings of the 5th International Conference on Economic Management and Big Data Application (ICEMBDA 2024)*, Advances in Economics, Business and Management Research 313,
https://doi.org/10.2991/978-94-6463-638-3_18

raise privacy concerns, as they require the collection and analysis of sensitive personal data. The field of emotion recognition is ripe for innovation, and there is a need for more sophisticated and ethically sound methods that can accurately capture and interpret the emotional landscape of the workplace(Zhang et al., 2023)[4].

The potential of improved MobileNetV3, a lightweight deep neural network architecture, offers a promising avenue for advancing emotion recognition technology. MobileNetV3 is designed to be efficient and effective, making it suitable for deployment on mobile devices and in real-time applications(Selvi A S, et al., 2023)[5]. Its ability to handle high-dimensional data with reduced computational resources could revolutionize the way emotions are detected and analyzed in the workplace. By leveraging the power of deep learning, MobileNetV3 can potentially recognize subtle emotional cues from facial expressions, voice patterns, and even physiological signals, providing a more nuanced and comprehensive understanding of employee emotions. The integration of such technology could lead to the development of systems that not only identify emotions but also predict and respond to emotional changes, thereby creating a more adaptive and supportive work environment.

The objectives of this paper are to explore the significance of emotion recognition in enhancing employee health and productivity, to examine the limitations of current methods, and to propose the use of an improved MobileNetV3 as a potential solution. The structure of the paper will begin with an overview of the importance of emotion recognition in the workplace, followed by a critical analysis of existing methods and their shortcomings(Liang S, et al., 2023)[6]. Subsequently, the paper will introduce MobileNetV3 and discuss its capabilities and potential applications in emotion recognition. Finally, the paper will conclude with a discussion on the implications of adopting advanced emotion recognition technology for organizational health and productivity, as well as the ethical considerations and future directions for research and development in this field(Zhao G, et al., 2020)[7].

2 Related Work

The journey of emotion recognition technology is a testament to the intersection of human psychology and computational intelligence. From its nascent stages, where researchers relied on basic algorithms to detect emotional cues, the field has burgeoned into a sophisticated domain leveraging advanced machine learning techniques. Early methods involved rule-based systems that identified emotions through predefined patterns in facial expressions, speech, and text. However, these were limited by their inability to generalize across diverse human behaviors. The advent of machine learning, particularly deep learning, revolutionized emotion recognition by enabling models to learn from large datasets and improve their accuracy in detecting nuanced emotional states(Wang M, et al,. 2023)[8].

Emotional psychology, the study of how emotions influence behavior, cognition, and decision-making, provides a foundational framework for emotion recognition technology. Theories such as the James-Lange theory, which posits that physiological changes precede emotional experiences, and the cognitive theory of emotion, which emphasizes the role of cognitive processes in emotion, offer insights into the complex nature of emotions. Understanding these theories is crucial for developing algorithms that can accurately interpret and respond to emotional cues. Emotional psychology also informs the ethical considerations in emotion recognition, ensuring that technology respects human autonomy and privacy(Xu X,et al., 2022)[9].

The landscape of emotion recognition methods is diverse, encompassing techniques that analyze facial expressions, speech, physiological signals, and textual data. Facial expression recognition, for instance, uses computer vision to track facial movements and map them onto affective states. Speech analysis, on the other hand, examines prosodic features like pitch, intensity, and rhythm to infer emotions. Textual analysis, often referred to as sentiment analysis, utilizes natural language processing to discern emotions from written language. Each method has its strengths and limitations, and the choice of method often depends on the specific application and the type of emotional cues available(Banerjee, A., et al., 2023)[10].

MobileNetV3, an efficient and lightweight neural network architecture, has made significant strides in image recognition, particularly in mobile and edge computing environments where computational resources are limited. Designed with a focus on mobile-first experiences, MobileNetV3 incorporates advanced techniques such as inverted residuals and linear bottlenecks to reduce model size and improve speed without compromising accuracy. Its application in emotion recognition has been promising, particularly in real-time facial expression analysis, where quick and efficient emotion detection is paramount. MobileNetV3's ability to handle high-resolution inputs makes it an excellent candidate for fine-grained emotion detection tasks.

3 Improved MobileNetV3 Architecture

MobileNetV3 is a lightweight neural network model designed for efficient image recognition and understanding on mobile devices with limited computational resources. Compared to its predecessor, MobileNetV3 has made significant improvements and optimizations in network structure and computational performance, achieving the goal of efficient and accurate image recognition and understanding on mobile devices.

Firstly, the original MobileNetV3 introduced a new type of network block called the "inverted residual structure," which combines residual connections and linear bottleneck structures to effectively reduce the computational burden and number of parameters of the network. This structure not only enhances the model's feature extraction capabilities but also makes the model more lightweight, suitable for deployment on mobile devices. Secondly, the model further optimized the design of time-consuming layers, using lightweight operations such as depthwise separable convolutions to reduce computational complexity and memory consumption. This design allows MobileNetV3 to achieve higher accuracy while maintaining faster inference speeds and lower power consumption, meeting the requirements of mobile environments(Si L, et al.,2023)[11]. The original MobileNetV3 also redesigned the time-consuming layers in the network structure, adopting a new structure of depthwise separable convolutions and lightweight attention mechanisms to further enhance the model's performance and generalization capabilities. In addition, this study changed the SE attention mechanism in MobileNetV3 to a serial spatial channel attention mechanism, further improving the model's perception of feature regions and effectively reducing power consumption on mobile devices. The improved structure of MobileNetV3 is shown in Table 1, where Input represents the size of the current layer, Operator represents the operation, exp size is the first 11-dimensional size, #out is the depth of the output matrix, Serial-SA-CA indicates whether attention mechanisms are used, NL represents the type of activation function, and s refers to the stride size.

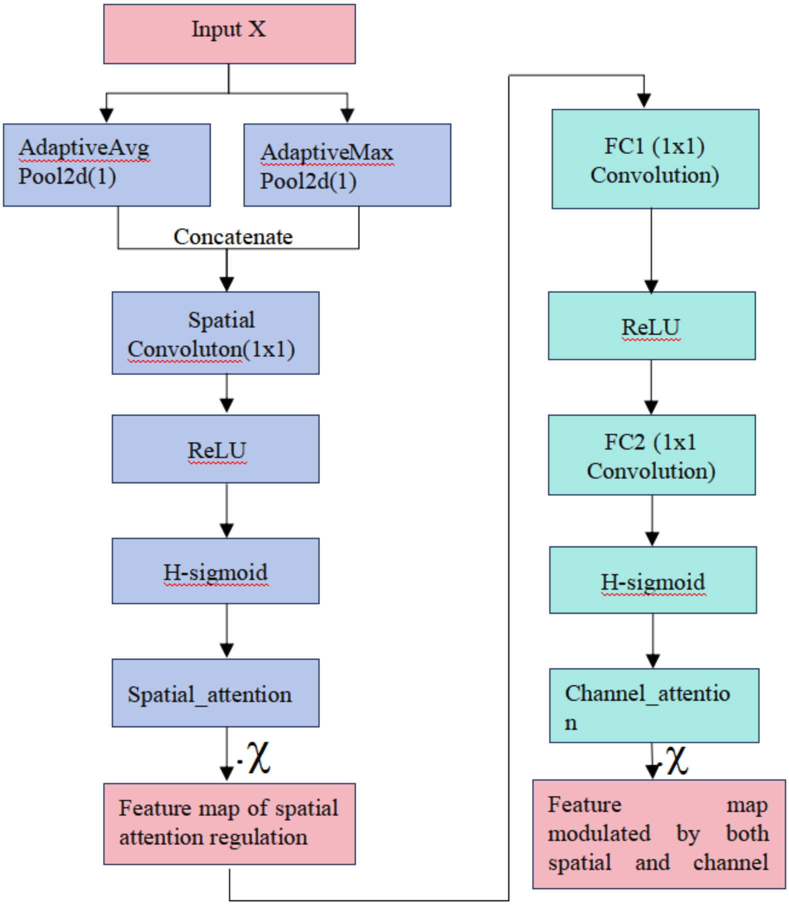

Fig. 1. Improved MobileNetV3 Architecture

Pluggable attention mechanisms and feature fusion methods are key strategies for boosting network performance on complex tasks. Attention mechanisms enable networks to concentrate on the most relevant parts of the input data, which can lead to more efficient and accurate processing. For example, Partial Channel Pooling Attention (PPA) selectively focuses on specific channels within feature maps, helping the network to capture important information without the need for extra parameters. Feature fusion methods integrate features from different sources or levels, providing a more holistic view of the data. Modules like the Three-Dimensional Weight Attention Module (WAM) consider both spatial and channel information when calculating attention weights, facilitating a more sophisticated integration of features.

The combination of these techniques allows networks to better identify and prioritize critical information, which can result in improved performance on complex tasks. These methods also enhance the network's ability to generalize, making it more adaptable to new data. As research continues, the use of attention mechanisms and feature fusion is expected to play a significant role in improving network capabilities across a variety of complex applications.

Figure 1 shows the improved MobileNetV3 architecture.

4 Employee Emotion Recognition Method

The training process of the improved MobileNetV3 model involves several steps. Initially, the model is initialized with random weights. The chosen architecture, which includes the inverted residual structure and depthwise separable convolutions, is set up according to the design specifications.

Training begins with forward propagation, where the input images are passed through the network to generate predictions. These predictions are then compared to the actual labels using a loss function, such as cross-entropy, which measures the difference between the predicted and true values.

The loss value is backpropagated through the network, allowing the calculation of gradients for each parameter. These gradients indicate how much and in which direction the weights should be adjusted to minimize the loss. An optimizer, such as Adam or SGD, is used to update the weights based on the gradients.

The training process is iterative, with each iteration (epoch) improving the model's accuracy. During training, it's essential to monitor for overfitting, where the model performs well on the training data but poorly on unseen data. Techniques such as dropout, regularization, and early stopping are employed to mitigate this issue.

The emotion recognition process using the improved MobileNetV3 model starts with feeding preprocessed input images into the trained model. The model processes the images through its layers, extracting features that are indicative of the emotional content.

The classification strategy involves mapping these features to the predefined emotion categories. This is typically done using a fully connected layer at the end of the network, which outputs a probability distribution over the emotion classes(Munsif, M, et al., 2024)[12].

Moreover, attention mechanisms, such as the serial spatial channel attention mechanism mentioned earlier, can be used to focus on the most relevant parts of the image for emotion recognition. This can lead to better interpretability and performance, especially when dealing with complex emotional expressions.

Finally, the classification strategy should include a post-processing step, where the predicted probabilities are converted into discrete class labels, and any additional business logic or decision-making criteria are applied.

In conclusion, the improved MobileNetV3 model, with its efficient architecture and advanced training techniques, is well-suited for emotion recognition tasks on mobile devices. The combination of data preprocessing, careful training, and strategic classification ensures high performance and accuracy in real-world applications.

5 Experimental Design and Result Analysis

The dataset chosen for this study is the KDEF (Karolinska Directed Emotional Faces) dataset, as depicted in Figure 2, which is a standard dataset extensively utilized in the field of emotion recognition research. Developed by the Karolinska Institute in Sweden, the KDEF dataset includes a variety of facial expression photographs from 70 distinct individuals, totaling around 4,900 images. These images encompass seven fundamental emotions, namely happiness, sadness, disgust, anger, surprise, fear, and neutrality. Each photograph has been subjected to a stringent standardization process, including uniform photographic conditions, facial expression instructions, and lighting conditions, to ensure the data's reliability and consistency.

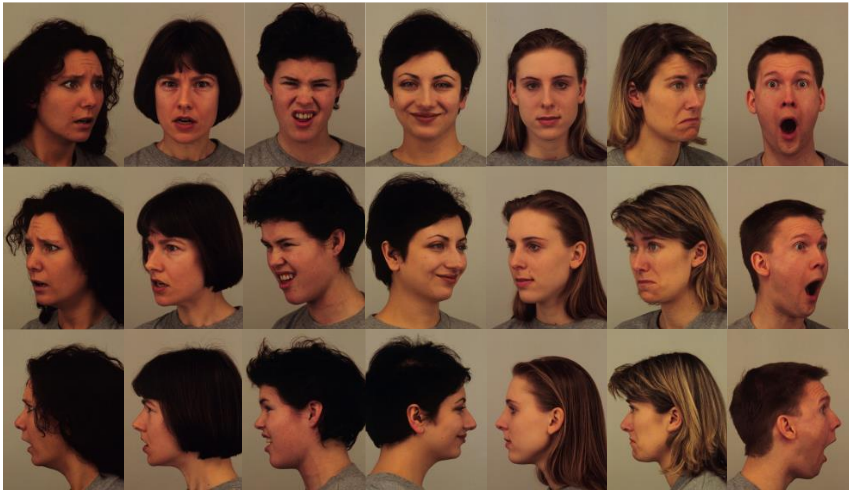

Fig. 2. The KDEF dataset

The KDEF dataset is particularly valuable for emotion recognition studies due to its comprehensive depiction of human emotions and the controlled conditions under which the photographs were taken. The standardization process ensures that the facial expressions are as natural and genuine as possible, unimpeded by extraneous factors such as varying lighting or camera angles that could potentially distort the outcomes.

In addition to the seven primary emotions, the KDEF dataset also provides annotations for the intensity of the emotions, which can be instrumental for studies aiming to explore the nuances of emotional expression. The dataset's annotations and standardized format render it an excellent resource for training and validating machine learning models, such as the improved MobileNetV3 model, within the realm of emotion recognition.

The KDEF dataset's widespread adoption in affective computing is also attributed to its diversity. It includes images of individuals from various ethnic backgrounds, facilitating the creation of a more inclusive model that can generalize effectively across different populations. This diversity is essential for developing models that are not only precise but also equitable and unbiased.

When employing the KDEF dataset for training the improved MobileNetV3 model, it is crucial to adhere to ethical guidelines for data usage, including securing necessary permissions and safeguarding the privacy of the individuals whose images are featured in the dataset. Furthermore, researchers should contemplate the ethical implications of emotion recognition technology and its applications, ensuring that the technology is applied responsibly and does not compromise personal freedoms or privacy.

We divide the original dataset into a training set and a test set in a 7:3 ratio to ensure the model's effectiveness and generalization ability. By adopting the KDEF dataset, this study is able to utilize its rich emotional expression data to conduct in-depth research and evaluation of emotion recognition algorithms.

To ensure the performance and stability of the model, parameter tuning and model optimization were conducted. A multitude of tests were performed by modifying key parameters such as batch size, optimizer type, number of training epochs, and the application of transfer learning. The optimal combination was selected. Specifically, a batch size of 32, training epochs of 150, and a learning rate of 0.0001 were determined, and transfer learning methods were adopted to fully leverage the features of pre-trained models.

To thoroughly investigate the influence of attention mechanisms on the performance of the MobileNetV3 model, this section has designed three sets of comparative experiments, detailed as follows: (1) A model incorporating the "Squeeze-and-Excitation Attention Mechanism" module, referred to as SE-MobileNetV3; (2) A model with both channel attention module and spatial attention module, connected in parallel, known as Parallel-SA-CA-MobileNetV3; (3) A model with both channel attention module and spatial attention module, connected in series, known as Serial-SA-CA-MobileNetV3. The final results of the experiments have been compiled in Table 1. By comparing the outcomes of each experiment, a more comprehensive understanding of the specific impact of attention mechanisms on the performance of the MobileNetV3 model can be achieved.

Table 1. Experimental Results

Methods	Accuracy(%)
SE- MobileNetV3	94.81
Parallel- SA-CA -MobileNetV3	93.31
Serial- SA-CA - MobileNetV3	94.95
POSTER++	94.44

MANet	91.75
Improved DenseNet	90.71
DAN	88.77
ResNet-50	86.81

From the comparative experimental results in Table 1, it can be seen that the use of serial spatial channel attention mechanism achieves the highest accuracy on the KDEF dataset. The Serial-SCA-MobileNetV3 module operates on both spatial and channel dimensions, making this attention mechanism more targeted for extracting facial expression features. It can capture more expressive features, thereby better achieving the goal of facial expression recognition.

To thoroughly investigate the specific performance of the Serial-SA-CA-MobileNetV3 model in recognizing seven different facial expressions, a pre-trained Serial-SA-CA-MobileNetV3 model was used to conduct a facial expression recognition experiment on the KDEF dataset. The recognition results for all samples in this dataset were recorded, with the specific data shown in Table 2.

Table 2. Experimetal Results

Expression	Accuray(%)	Sensitivity(%)	Specificity(%)
Fears	89.25	91.39	98.17
Anger	95.31	96.67	99.20
Disgust	94.67	93.75	99.13
Joy	98.57	98.10	99.76
Neutral	99.51	96.17	99.92
Sadness	90.28	92.86	98.32
Surpris	95.12	93.30	99.20

6 Conclusions

This paper proposes an employee emotion recognition method based on an improved MobileNetV3 model, achieving satisfactory recognition results on the KDEF dataset. Compared to other state-of-the-art methods, the algorithm presented in this paper shows significant performance improvements.

References

1. Prasad, S. B. R., & Chandana, B. S. (2023). Mobilenetv3: a deep learning technique for human face expressions identification. *International journal of information technology*, *15*(6), 3229-3243.
2. Jiang, B., Li, N., Cui, X., Zhang, Q., Zhang, H., Li, Z., & Liu, W. (2023). Research on facial expression recognition algorithm based on improved MobileNetV3.

3. Bisht, S., Singhal, A., & Kaushik, C. (2022, September). Face Recognition Using Deep Neural Network with MobileNetV3-Large. In *International Conference on Advances and Applications of Artificial Intelligence and Machine Learning* (pp. 115-123). Singapore: Springer Nature Singapore.
4. Zhang, Z., Yang, X., Luo, N., Chen, F., Yu, H., & Sun, C. (2023). A novel method for Pu-erh tea face traceability identification based on improved MobileNetV3 and triplet loss. *Scientific Reports*, *13*(1), 6986.
5. Selvi, A. S., Aakash, S., Hariharan, M., & Abijith, P. (2023, November). EmoTune: Deep Emotion Detection and Music Recommendation System using MobileNetV3. In *2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)* (pp. 1-6). IEEE.
6. Liang, S., Tan, T., & Jonathan, J. (2023). MobileNetV3-based Handwritten Chinese Recognition Towards the Effectiveness of Learning Hanzi. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *7*(6), 1394-1402.
7. Zhao, G., Yang, H., & Yu, M. (2020). Expression recognition method based on a lightweight convolutional neural network. *IEEE Access*, *8*, 38528-38537.
8. Wang, M., Mei, Q., Song, X., Liu, X., Kan, R., Yao, F., ... & Qiu, H. (2023). A Machine Anomalous Sound Detection Method Using the lMS Spectrogram and ES-MobileNetV3 Network. *Applied Sciences*, *13*(23), 12912.
9. Xu, X., Tao, R., Feng, X., & Zhu, M. (2022, July). A lightweight facial expression recognition network based on dense connections. In *International Conference on Knowledge Management in Organizations* (pp. 347-359). Cham: Springer International Publishing.
10. Banerjee, A., Mutlu, O. C., Kline, A., Surabhi, S., Washington, P., & Wall, D. P. (2023). Training and profiling a pediatric facial expression classifier for children on mobile devices: machine learning study. *JMIR formative research*, *7*, e39917.
11. Si, L., Li, J., Wang, Z., Wei, D., Gu, J., Li, X., & Meng, L. (2023). A Novel coal-gangue recognition method for top coal caving face based on IALO-VMD and improved MobileNetV2 network. *IEEE Transactions on Instrumentation and Measurement*.
12. Munsif, M., Sajjad, M., Ullah, M., Tarekegn, A. N., Cheikh, F. A., Tsakanikas, P., & Muhammad, K. (2024). Optimized efficient attention-based network for facial expressions analysis in neurological health care. *Computers in Biology and Medicine*, *179*, 108822.