Multi-class Facial Emotion Recognition Using Deep Learning Techniques

Yoon-Teck Bau1, 2, a), Kai-Ru Toh1, b) and Kian-Chin Lee1, 2, c)

1Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia.

2Centre for Advanced Analytics, CoE for Artificial Intelligence, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia.

b) Corresponding author: kairu.toh@gmail.com  
a) ytbau@mmu.edu.my

c) kclee@mmu.edu.my

**Abstract.** Emotion recognition from facial expressions has attracted extensive research interest in the field and has key applications in human-computer interaction, healthcare, customer service, sales, and education. In this research, three deep learning-based emotion recognition techniques are developed to recognize human facial emotions into seven different emotion classes that include anger, disgust, fear, happiness, neutral, sadness, and surprise. The emotions dataset used in the study is publicly available from Roboflow website, and the dataset is then preprocessed by resizing images to 48×48, normalizing pixels between 0 and 1, and categorizing the labels into binary encoded vectors. An exploratory data analysis is conducted to count the number of images for each emotion and plot 10 sample images for each emotion before applying deep learning techniques for emotion recognition. The implemented three deep learning techniques are VGG16, EfficientNetV2S and ResNet50. Two metrics such as accuracy and classification report are used to evaluate the performance of these deep learning techniques. From the experimental results, VGG16 performed the best among the three deep learning techniques in terms of accuracy. The classification report for multi-class emotion classification reveals the performance of the technique across different emotion categories. Among the various classes, the emotion happy was the most accurately recognized by the deep learning technique. Conversely, the emotion disgust was the least accurately recognized.

# INTRODUCTION

Human emotions are essential to interpersonal relationships, communication, and decision making. Emotions are key indicators of an individual's mental state. Therefore, techniques that can help to identify human emotions are crucial to creating an emotion recognition system that can effectively interact, communicate and empathize with humans. An accurate emotion recognition system can be used in a wide range of applications, including in healthcare, such as helping patients with stress detection and mental health monitoring to analyze and propose effective solutions. Besides being used in healthcare, emotion recognition systems can also be used in education to customize learning programs that best suit students based on their emotional states.

Facial emotion is one of the most common and important ways for people to express feelings. Nevertheless, the existing techniques lack features that help humans to interpret face emotions correctly. The individual differences in human facial structure, cultural differences in emotions and emotions nuances have contributed to a huge obstacle in front of emotion recognition systems.

This research proposes deep learning techniques to solve the issue of emotion recognition by facial expressions. The deep learning technique can recognize intricate patterns and features within facial images corresponding to different emotions. In this research, a dataset of diverse and thousand images of faces will be used to test the deep learning techniques to provide high accuracy in emotion recognition and classification. This research will use three reliable and effective emotion recognition techniques for identifying multiple classes of emotions and compare the performance results between the techniques.

# LITERATURE REVIEW

Deep learning has significantly enhanced facial emotion recognition performance over the past several years. Deep learning techniques and convolution neural networks can directly learn hierarchical features from the facial data without feature engineering. These techniques are now a very common and standard method of facial emotion recognition due to their ability to scale easily to new data and their performance improvements over traditional methods.

Various research studies examined various convolutional neutral network architectures for facial emotion recognition. Kalpana Chowdary et al. used pre-trained networks like VGG, ResNet, and MobileNet with the help of transfer learning techniques by fine tuning fully connected layers [1]. Although MobileNet showed computational efficiency, its light weight may prevent the network from capturing slight facial features necessary for the correct recognition of emotions.

Similarly, Singh & Nand also proposed a method of convolutional neutral network technique to train the facial expression dataset [2]. The research focused on enhancing technique performance by employing preprocessing methods such as resizing and histogram equalization to counter low resolution of images. Then, Timothy et al. employed the InceptionV3 technique, leveraging from its parallel convolutional architecture in learning global and local features in parallel [3]. Nevertheless, its parallel architecture added training memory usage and reduced interpretability of the approach. Additionally, it required careful data preprocessing and fine-tuning of inputs, which can make the technique challenging to transfer to other different facial emotion recognition datasets.

Other research used attention mechanisms for performance improvement. Pham et al. suggested the residual masking network that combined the attention-based masking with residual learning to focus attention on relevant areas of the face [4]. However, its depth structure and additional computational cost made it unacceptable for real time or lightweight deployment contexts. Also, the need for fine grained optimization and the possibility of overfitting with small datasets makes it even challenging. Moreover, Lavanya et al. proposed a residual neural network with convolutional block attention module for selective attention to the salient spatial and channel features [5]. However, the addition of attention modules increased the overall technique complexity and computational requirements.

Besides, transfer learning is also a prevalent practice in facial emotion recognition. Rane et al. proposed a hybrid approach of combining a pre-trained VGG and a specifically trained convolutional neural network for a balance between real time performance and recognition accuracy [6]. A strategy like this can successfully enhance real time response ability, thus being viable for applications where acceptable performance and speed. In the same way, Bentoumi et al. used feature extraction from VGG and ResNet for better generalization [7].

Recent studies also contemplated the use of EfficientNet based models for facial emotion recognition using transfer learning. Utami et al. employed different variants of EfficientNet on the facial expression recognition database through transfer learning. The pre-trained convolutional base in their research is first frozen and then fine tuned later to the dense layers after standard pre-processing of the face images [8]. The results concluded that the smaller EfficientNet version yielded an acceptable accuracy and computational cost trade-off, whereas larger EfficientNet version yielded better accuracy but at the cost of very much increased computational cost.

Overall, deep learning techniques such as VGG and ResNet have proven efficient for facial emotion recognition by deriving useful features from faces. Studies on architecture such as the smaller EfficientNet model have also demonstrated the potential to achieve a balance between model complexity and computation cost. Despite differences in methodology and technique architecture, these techniques share the mutual goal of increasing the accuracy and robustness of emotion classification from facial expression images.

Among the most pressing topics is the method generalizability across diverse setups. High accuracy was realized by [1] and [7] with controlled environment sets such as Cohn-Kanade dataset (CK+) and Karolinska directed emotional faces (KDEF) but fared badly with more difficult, real word sets such as facial expression recognition 2023 dataset (FER2013) or face expression recognition plus (FER+) dataset. Lower accuracy was realized by [2] and [5] with face expression recognition plus (FER+), a problem that suggests that many current techniques are not robust enough to transfer to variations in lighting, pose, and spontaneous facial expressions that are encountered in free settings.

There is also another gap concerning the underutilization of feasible, real-time techniques. Most research still depends upon deep and computation intensive architectures such as InceptionV3. Despite the existence of more efficient alternatives such as the smaller EfficientNet model. This calls upon more research to make facial emotion recognition techniques both feasible in real time and with limited devices.

Another gap is the long-discussed ambiguity between visually similar feelings. Some studies like [4] and [6], were limited by ambiguity between fear and surprise or anger and disgust expressions. Such problems are generally amplified by a class imbalance in large population datasets, where some like happy and neutral are overrepresented. Even though this problem is well known, few studies leveraged special techniques like class weighing or oversampling to offset the imbalance and provide fairness to classification. Little attention is also paid to subtle expressions and multimodal inputs. All considered methods only make use of static face images and classify feelings into discrete classes. In practice, however, the states of feelings are generally subtle and continuous.

In addition, some studies only compare their methodology to one dataset with minimal generalization to new, unseen data. Cross dataset testing where a method is trained and assessed over a different dataset sparingly reported in the literature, although this is a requirement for robustness.

To consider interpretability or visual explanation strategies, attention-based networks like residual neural networks with convolutional block attention module [5] and residual masking network [4] that are transparent, can be applied. Most techniques are black boxes and do not give insight into what part of the face or feature is responsible for classification of emotions. This lack of transparency will hinder trust and use, especially in sensitive contexts like education, healthcare, and human-computer interaction.

Overall, though recent research has expanded the boundaries of facial emotion recognition systems, there are some areas that require further research. Closing such gaps will play a crucial role in enabling more accurate, more efficient, and more interpretable emotion recognition methods that can be successfully translated to real word uses.

# FACIAL RECOGNITION TECHNIQUES

## Dataset

This study uses a dataset containing facial emotion images labeled with seven different emotion classes for emotion recognition. The emotion dataset comes from a public dataset released on the Roboflow website. The dataset provides labeled images that can be used to train and test emotion recognition deep learning techniques [9]. There are 23,807 images in the training set and 3,395 images in the test set, where the images are distributed across the seven categories: angry, disgust, fear, happy, neutral, sad, and surprise. The distribution of the number of images for each emotion class varies. The emotional appearances of the surprise and disgust classes are less frequent in both the training and test datasets and are thus underrepresented. Specifically, there are 2,282 samples of the surprise class and 337 samples of the disgust class in the training set. In the test set, the surprise class has 314 images, while the disgust class has just 38 images. Each image in the dataset is resized to 48×48 pixels, represented in grayscale format to ensure uniformity across all input data. The dataset was created in quite diverse environments where facial expressions might appear from different sides and with other appearances like eyeglasses.

## Training and Testing

In this research, the dataset of images was categorized into separate directories for training and testing. The TRAIN\_DIR and TEST\_DIR variables were used to specify the file path for the training and testing sets, respectively. To ensure uniform input dimensions to the deep learning techniques, all images were resized to 48×48 pixels, as specified by the IMG\_SIZE variable. This made the preprocessing consistent and compatible with the chosen technique architectures in the development process.

## Exploratory Data Analysis

Exploratory data analysis is an important step to understand the structure of the data. This study will use a graphical view of images to conduct the EDA of the collected datasets for each emotion class and inspect the spread of facial expressions as well as the balance of the dataset. The total number of images for each emotion class is 10 images for each emotion class. Then, a title is added above each group of images to show the label of the emotion class in the generated visualization. This visualization is intended to provide an overview of the dataset. Finally, ten sample images from each emotion class in the training set as shown in Figure 1.

A collage of different faces

AI-generated content may be incorrect.

**Figure 1.** Ten sample images from each emotion class in the training set

## Data Preprocessing

The data preprocessing step is used to pre-train the dataset for emotion recognition techniques. This includes several common examples, including resizing images, normalizing pixel values, and one-hot encoding labels. Before doing this preprocessing step, the researcher needs to ensure that the labels are all set up correctly for the classification task. In this research, the integer labels of the seven emotion classes are one-hot encoded and were used for the to\_categorical() function, which converts the integer labels of y\_train and y\_test into binary vectors. The value of "1" denotes active classes, and the value of "0" indicates inactive classes in both datasets.

## Visual Geometry Group Technique

Visual Geometry Group or VGG, is a deep learning technique proposed by [10]. VGG technique is popular in research due to VGG's simplicity and consistent architecture, which uses only small 3×3 convolutional filters in all layers. Therefore, these small filters along with the 2×2 maximum pooling layer, enable the network to learn the fine scale of spatial details in the source image while gradually reducing the spatial dimensionality of the feature map. There are many versions of VGG deep learning technique, and one of the most popular is a version called VGG16, which attempts to implement 13 convolutional and 3 dense layers for a total of 16 layers of deep architecture [10], which can refer to Figure 2. For the last layer of each block, maximum pooling is used. Then, a rectified linear unit (ReLU) activation function is applied to introduce non-linearity into the technique for the first dense layer. The last dense layer, the softmax activation function is used to compute probabilities for multi-class classification tasks. In transfer learning, pre-trained weights from the ImageNet dataset, a large-scale image collection with millions of labeled images, are used to initialize models like VGG, helping to significantly reduce training time and computational cost [11]. This adaptability makes VGG well-suited for facial emotion recognition tasks since this technique can take advantage of excellent feature extraction capabilities [12].

A screenshot of a cell phone

AI-generated content may be incorrect.

**Figure 2.** Structured breakdown of the VGG16 architecture

## EfficientNet Technique

Efficient Networks or EfficientNet is a family of convolutional neutral networks that seeks to obtain better performance with the best use of computing resources. Unlike the traditional deep learning techniques, EfficientNet introduces a compound scaling technique that scales the depth, width, and resolution of the network equally with a set of predetermined scaling factors. The balanced scaling technique enables better performance at reduced training time and memory usage. EfficientNetV2S is a variant of the original EfficientNet architecture that enables faster training and better performance while reducing computational complexity [13]. The character S in EfficientNetV2S technique stands for small, which means it’s a lightweight version technique. As shown in Figure 3, the base architecture begins with an input layer followed by a rescaling layer and a stem convolutional layer implemented using a 3×3 convolution and 24 filters. The first block is preceded by batch normalization and a rectified linear unit (ReLU) activation function. The backbone of the network consists of a series of mobile inverted bottleneck convolution (MBConv) blocks arranged in six stages that include downsampling the spatial resolution and boosting the depth of the feature maps. These six block stages transform the input to 48×48×3 to 24×24×24, 12×12×48, 6×6×64, 3×3×128, 3×3×160 and finally 2×2×256. The three numbers represent the dimensions of the feature maps at each stage of the network, where the first two values indicate the height and width of the feature map, and the third value indicates the number of channel depths, showing how the network gradually reduces spatial size while increasing feature complexity. Each of these blocks includes expansion layers, projection layers, depthwise convolutions, squeeze modules, excitation modules, batch normalizations, and non-linearity activations. A top final ReLU activation and a batch normalization layer complete the base architecture.

A screenshot of a phone

AI-generated content may be incorrect.

**Figure 3.** Structured breakdown of the EfficientNet architecture

## Residual Neural Networks Technique

Residual Neural Networks or ResNet are a deep learning technique that is built to solve the vanishing gradient problem in deep neural networks. ResNet is widely used for image classification tasks due to ability to capture complex hierarchical features [14]. ResNet50 is widely utilized in computer vision tasks because it has the capability of learning abstract and hierarchical features of images. The 50 in ResNet50 means that there are a total number of learnable layers in the network, including 49 convolutional layers and 1 fully connected layer at the end for multi-class classification. As shown in Figure 4, the base architecture of ResNet50 begins with an input layer, followed by a 7×7 convolutional layer that has 64 filters, then normalized using batch normalization and activated using the ReLU function. This is succeeded by a maximum pooling layer with size 3×3 that reduces the spatial dimension of the input image. The architecture then progresses through a sequence of residual blocks, with a total of four stages, where each stage is composed of residual blocks consisting of three layers of stacked convolutional in the order 1×1, 3×3, and 1×1 convolutions, which perform dimension reduction, feature extraction, and recovery of dimension, respectively. Next, a global average pooling (GAP) layer is applied to compress each feature map into a single value by averaging its spatial elements. The key feature of each block is the use of shortcut connections, which bypass the inner layers of the block by adding the input to the output, allowing gradients to propagate more easily through the network. After that, the compressed features are passed into a fully connected dense layer that output the final classification output across the emotion classes.

A screenshot of a phone

AI-generated content may be incorrect.

**Figure 4.** Structured breakdown of ResNet50 architecture

# RESULTS AND DISCUSSION

The chapter demonstrates the results and analysis of facial emotion recognition using the three deep learning techniques: VGG16, EfficientNetV2S, and ResNet50. A comparison between all the techniques is given from the perspective of classification accuracy and classification report of the testing data. The methods are compared among themselves to emphasize all the techniques’ strengths and weaknesses, and a discussion is given to comment on the resultant outcomes. Table 1 shows the classification accuracy evaluation for the three deep learning techniques used: Visual Geometry Group (VGG16), EfficientNetV2S, and Residual Neural Networks (ResNet50). VGG16 technique achieved the highest testing accuracy of 60%, while EfficientNetV2S achieved 54% and ResNet50 achieved 53%. The result of VGG16 indicates that its comparatively simpler design was better in identifying important characteristics from facial images compared to EfficientNetV2S and ResNet50 techniques. This validates that the depth of the technique is not always an assurance of improved performance, particularly for small and well-proportioned datasets.

|  |  |
| --- | --- |
| **TABLE 1.** Comparison result table of three deep learning techniques, the best result is in bold | |
| **Technique** | **Accuracy** | |
| Visual Geometry Group | **0.60** | |
| EfficientNet | 0.54 | |
| Residual Neutral Networks | 0.53 | |

Based on the classification report among the three deep learning techniques, VGG16 performing best in the happy emotion class (Precision: 0.80, Recall: 0.85, F1-score: 0.83) but worst on fear (Recall: 0.22, F1-score: 0.32). EfficientNetV2S performing best on happy (Precision: 0.75, Recall: 0.74, F1-score: 0.75) and worst on disgust (Recall: 0.26, F1-score: 0.34). ResNet50 performing well on happy (Precision: 0.80, Recall: 0.75, F1-score: 0.77) but poorly on disgust (Recall: 0.26, F1-score: 0.36).

# CONCLUSION

This research implemented and compared the performance of three deep learning techniques: VGG16, EfficientNetV2S, and ResNet50 in recognizing facial emotions from facial images. Among the techniques in this study, VGG16 achieved the highest accuracy of 60%, beating EfficientNetV2S and ResNet50, both of which achieved 54% and 53% accuracy. This result suggests that the simple and effective architecture of VGG16 is well suited for emotion detection as it can capture useful spatial features and still generalize better with lower resolution and grayscale training data. This research also tested the benefit of initializing pre-trained weights using transfer learning, which simplified training while retaining competitive performance. Furthermore, for multi-class emotion recognition, happy expression performed the best with the highest accuracy overall across all emotion classes, indicating that its facial characteristics were better distinguishable. In contrast, the disgust expression performed the worst due to both the low amount of training data and the subtlety of its facial features, which often resemble those of other negative emotions. A potential extension for future research is exploration of input imagery at higher resolution levels like from 48x48 to 96×96 pixels images. Higher resolution images might provide the technique with a richer set of fine-grained visual details, which may improve the test accuracy, especially for subtler emotional expressions. In addition, it would demonstrate the flexibility of deep learning architectures like VGG16 to generalize well to higher level input data.

# References

1. M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, “Deep learning-based facial emotion recognition for human–computer interaction applications,” Neural Comput & Applic **35**(32), 23311–23328 (2023).
2. E. Singh, and P. Nand, “Enhancing Facial Emotion Recognition through Deep Learning and Image Pre-processing,” in *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, (IEEE, Dehradun, India, 2023), pp. 1–4.
3. J. Timothy, R. Prasetio, D. Tandi, H. Lucky, and I. A. Iswanto, “Facial Emotion Recognition with InceptionV3 CNN Architecture,” in *2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS)*, (IEEE, Jakarta Selatan, Indonesia, 2023), pp. 1–5.
4. L. Pham, T. H. Vu, and T. A. Tran, “Facial Expression Recognition Using Residual Masking Network,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, (IEEE, Milan, Italy, 2021), pp. 4513–4519.
5. M. S. Lavanya, V. Arun, M. Tapkire, and K. P. Suhaas, “Transfer Learning Based Facial Emotion Recognition,” SN COMPUT. SCI. **6**(1), 35 (2024).
6. M. Rane, S. Shahare, S. Daware, Y. Shedge, S. Deshmukh, and G. Sarak, “Human Facial Emotion Recognition using Deep Learning Techniques,” in *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, (IEEE, Bhilai, India, 2023), pp. 1–6.
7. M. Bentoumi, M. Daoud, M. Benaouali, and A. Taleb Ahmed, “Improvement of emotion recognition from facial images using deep learning and early stopping cross validation,” Multimed Tools Appl **81**(21), 29887–29917 (2022).
8. P. Utami, R. Hartanto, and I. Soesanti, “The EfficientNet Performance for Facial Expressions Recognition,” in *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, (IEEE, Yogyakarta, Indonesia, 2022), pp. 756–762.
9. “Emotion Classification Dataset by shoes,” Roboflow, (2024).
10. K. Simonyan, and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” (2015).
11. A. Dhillon, and G. K. Verma, “Convolutional neural network: a review of models, methodologies and applications to object detection,” Prog Artif Intell **9**(2), 85–112 (2020).
12. Z. Liu, H. Mao, C. -Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE, New Orleans, LA, USA, 2022), pp. 11966–11976.
13. M. Tan, and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” (2020).
14. F. A. Demmese, S. Shajarian, and S. Khorsandroo, “Transfer learning with ResNet50 for malicious domains classification using image visualization,” Discov Artif Intell **4**(1), 52 (2024).