Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/21/20231109

Facial expression recognition based on Feature Pyramid Network

Yongcheng Huang

Faculty of Electrical Engineering, Mathematics & Computer Science, Delft University of Technology, Mekelweg 5, 2628CD Delft, Netherlands

Y.Huang-51@student.tudelft.nl

Abstract. Facial expression recognition with significant implications across fields such as psychology, computer science, and artificial intelligence. This paper proposes a combination of a Feature Pyramid Network (FPN) and a Residual Network (ResNet) to construct a recognition model. The main objective of the proposed model is to refine the multi-level feature representation of facial expressions. This approach aims to provide a more holistic understanding of the diverse and complex nature of facial expressions, recognizing the intricate interplay between macro and micro-expressions. Experimental results underscore the model's considerable superiority over traditional methods, particularly in terms of accuracy and adaptability to objects of varying sizes and complexities. This comprehensive approach to facial expression recognition showcases the potential of integrating different neural network architectures, furthering our understanding of the subtleties of facial expressions. The research, therefore, presents a significant contribution to the field of facial expression recognition, demonstrating the efficacy of integrating multi-scale feature extraction techniques to improve model performance. It sets the stage for future research directions in this domain, paving the way for more sophisticated emotion recognition systems that can be deployed in real-world applications.

Keywords: facial expression recognition, Feature Pyramid Network, Residual Network.

1. Introduction

The fascinating field of facial expression recognition can be traced back to the early Aristotelian era (4th century BC) [1]. Over the years, advances in science and technology have played a crucial role in automating the process of facial recognition [2]. Before the widespread adoption of Neural Networks, facial expression recognition was achieved through various feature extraction methods, such as optical flow [3]. Often, the processes used to draw out characteristics were routinely combined with machine learning approaches to improve the interpretation and handling of the obtained traits. By employing approaches like Support Vector Machines (SVM), researches were able to classify different facial expressions using the extracted features [4].

As deep learning gained prominence, neural networks became the go-to approach for classification tasks. Several deep learning techniques, including Multi-layer perceptron (MLP) [5], started to be employed in facial expression recognition tasks to enhance accuracy and efficiency. Also in 2016, Bargal, Sarah Adel, et al. tried to use Visual Geometry Group Network of 16 layers (VGG16), which is another State Of The Art (SOTA) model, to recognize facial expression [6]. In the above-mentioned

DOI: 10.54254/2755-2721/21/20231109

works based on Convolutional Neural Networks, researchers have used a bottom-up approach, using the highest level of semantic layers to construct the output layer for prediction. Nevertheless, facial expressions demand not only the evaluation of high-level semantic attributes, but also the analysis of low-level texture details. Moreover, another main drawback of the above-mentioned models is that when the number of layers is large, it is prone to gradient explosion or vanishing, which affects training. Therefore, it is necessary to fuse multi-level feature information to further enrich the feature representation of the network while solving the problem of gradient explosion and vanishing in deep learning. Feature Pyramid Network (FPN), as a new architecture for a universal feature extractor, has made significant improvements in various scenarios since IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR2017) [7]. special residual architecture can learn the mapping relationship between input and output more effectively, which has been used to further improve the performance of analyzing facial expressions [8]. Also, for its outstanding performance, FPN has been extensively used in a range of object recognition projects, such as the task on dataset Microsoft Common Objects in Context 2017 dataset (MS-COCO 2017), which was featured in CVPR2021 [9]. All these features above are showing that FPN, a general feature extractor, has been used to get huge improvements in various situations.

To further enhance the feature representation of facial expressions and improve network performance, this study adopted the FPN architecture and its integration with ResNet in facial expression recognition tasks. In this research, the Facial Expression Recognition 2013 Dataset (FER-2013) open-source dataset was employed for both training and assessment purposes, which is available at Kaggle [10]. The Resnet deep residual network is utilized as the primary network to derive feature maps of various scales. This is followed by the FPN which in turn enhances the feature representation capabilities to improve network performance. This process is mainly divided into five steps. The initial step involves feature extraction, where feature maps of varying scales are gathered via a primary network. The subsequent step is the bottom-up pathway, where a collection of feature pyramids is constructed by amplifying the feature maps from lower levels and integrating them with those from higher levels. Then is top-down pathway, which creates a set of feature maps that have both high-level semantic information and fine-grained details by upsampling the feature maps at each level and merging them with the corresponding feature maps from the higher level. Next, to improve information flow between the feature maps from different levels, we need to use lateral connections. Finally, the feature maps are used from the top-down pathway to output facial expression recognition results. This research creates a more accurate and efficient facial expression recognition model by combining the capabilities of FPN and ResNet. Compared with traditional methods such as ResNet50, the proposed model has significant improvements in terms of accuracy. The experiments demonstrate that this research can effectively analyze changes in facial emotions and make reasonable judgments.

2. Methodology

2.1. Dataset description and preprocessing

The dataset used in this study, called FER-2013, is sourced from Kaggle [11]. Comprised of grayscale images of faces, each image is 48x48 pixels in size. The images have been auto-adjusted to guarantee that the face is essentially centered and takes up approximately the same space in every image. The dataset contains two main features: the pixel values of the images and the emotion label.

The objective here is to categorize the faces from the dataset according to the emotion conveyed through the facial expression. There are seven categories for this classification, and they are given labels from 0-6: Angry (0), Disgust (1), Fear (2), Happy (3), Sad (4), Surprise (5), and Neutral (6). The dataset is split into a training set, which contains 28,709 examples, and a public test set, comprising 3,589 examples. An example of a data point in this dataset would be a 48x48 pixel grayscale image of a face, associated with an emotion label. Figure 1 showcases some instances from the dataset.

DOI: 10.54254/2755-2721/21/20231109

Figure 1. Images from the FER-2013 dataset.

2.2. Proposed approach

The focus of this proposed method for emotion classification revolves around the innovative fusion of ResNet50 and Feature Pyramid Network (FPN). This method is built upon the solid basis of ResNet50, a well-known convolutional neural network, combined with the multi-level feature learning potential of the FPN. These technologies, when combined, will be able to extract and utilize spatial and contextual information from facial images effectively, ultimately leading to better performance in the emotion classification task. Figure 2 below illustrates the structure of the system.

Figure 2. The pipeline of the model.

2.2.1. *Resnet50.* The convolutional base of this model is established by employing ResNet50, which is renowned for its efficient handling of the vanishing and exploding gradient problem through the incorporation of skip connections. ResNet50 is used as a feature extractor in the process, with the input size set at 48x48, modified from the original input dimension for ResNet network, which allows us to use smaller grayscale images. Prior to being input into the network, the images are subjected to a sequence of transformations. These include resizing to 512x512, normalizing pixel values, and applying data augmentation methods like random horizontal flipping and rotation, all with the intention of enhancing the network's robustness. The architecture of the model, although consistent with the standard ResNet50 network, is modified in the final fully connected layers to be tailored to this emotion classification task. These layers output a score for each of the 7 emotion categories instead of the original 1000 classes.

2.2.2. *Feature Pyramid Network (FPN).* The feature extraction capability of this model is enhanced by employing the FPN. Initially, the output from the last three layers of the ResNet50 network, labeled as C5, C4, and C3, is obtained. Then the layer C5_reduced is created with C5 as input, which reduces the number of channels in the C5 tensor to 256. After that, the layer C5_upsampled is created, where the first input (here, C5_reduced) is resized to have the same width and height as the second input (here, C4). This is part of the process of constructing the feature pyramid in the FPN, where higher-level features are upsampled and merged with lower-level features. Finally, a kernel of size 3x3 is created and applied to the upsampled tensor from the previous step. The output is stored into the P5 variable. This operation can help to smooth the features after the upsampling operation. Figure 3 below shows the process of creating the new P5 layer.

DOI: 10.54254/2755-2721/21/20231109

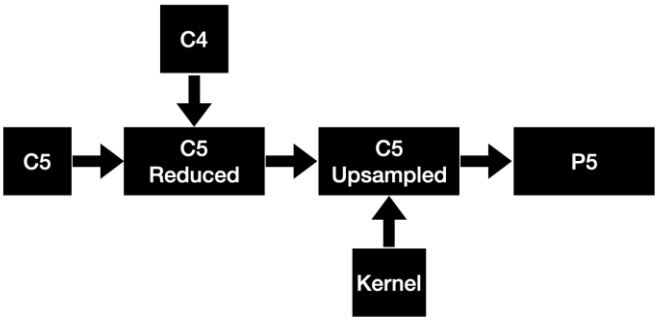


Figure 3. Creating P5 layer.

By doing so, the P5 layer is created and other FPN layers could be created by applying the same method. After creating the P3, P4, P5 layers, P6 is obtained through a 3x3 convolution with a stride of 2 on C5. Also, P7 is computed by applying Rectified Linear Unit (ReLU) followed by a 3x3 convolution with a stride of 2 on P6. Each layer created in the previous step is subjected to a ReLU activation function. This introduces non-linearity, thereby increasing the model's ability to learn intricate patterns. Following the application of the ReLU function, prediction layers are generated from the activated outputs. These prediction layers are then concatenated to form a cohesive feature representation. Lastly, a softmax function is applied to the concatenated layer, providing a probability distribution over possible classes and facilitating the selection of the most probable class as the final prediction. The Figure 4 below showing the whole structure of FPN described above.

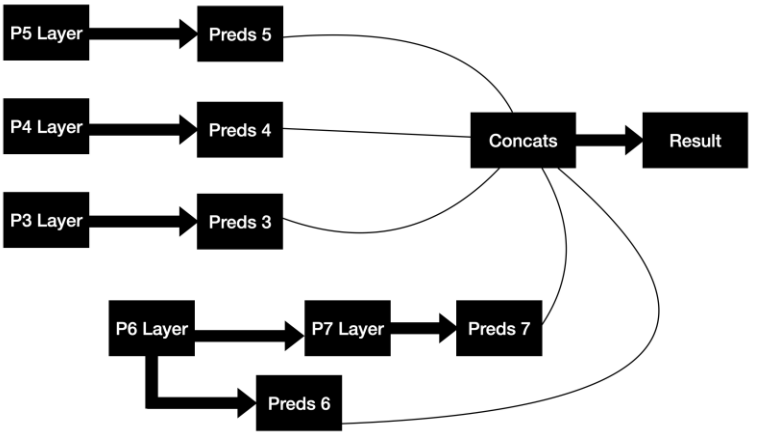


Figure 4. The structure of FPN.

2.2.3. *Loss function.* Choosing the right loss function plays a pivotal role in the training of deep learning models. For this emotion classification task, the Categorical Crossentropy loss function is optimal due to its effectiveness in multi-class classification problems. Categorical Crossentropy measures the dissimilarity between the model's predictions and the actual labels, thus, pushing the model to assign higher probability to the correct class during the training phase. The loss is calculated for each image, where the model's predicted probabilities for each emotion class are compared against the one-hot encoded actual label.

$$H(p,q) = -\sum_{i} p(i) \log q(i) \tag{1}$$

DOI: 10.54254/2755-2721/21/20231109

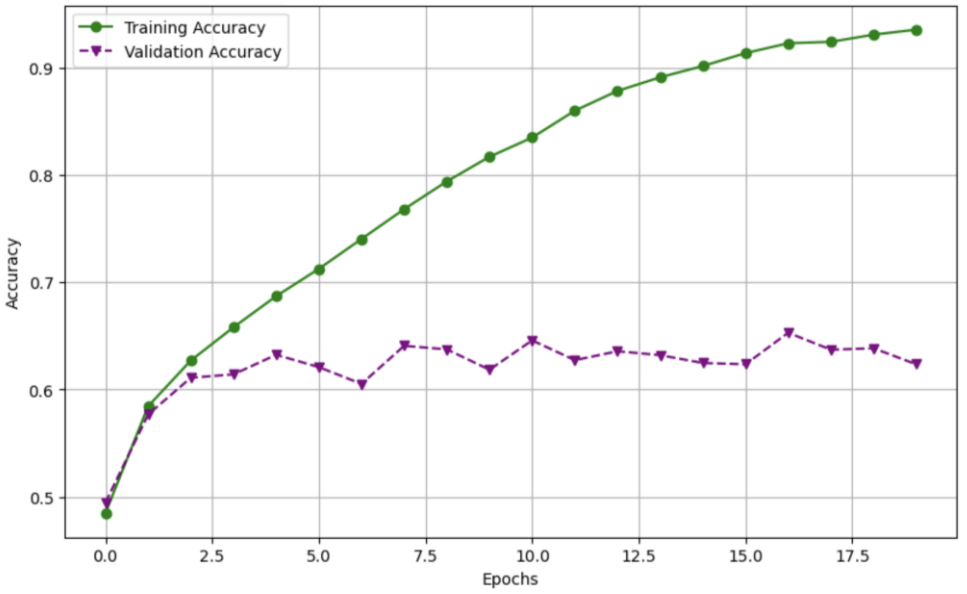
The above formula denotes the Categorical Crossentropy loss, where $p(i)$ signifies the actual probability of class i and $q(i)$ illustrates the predicted probability of class i. Then, the gradients are backpropagated through the model to update the weights. To prevent overfitting, a regularization term iseincorporated in the form of L2 regularization, which adds a penalty proportional to the square of the magnitude of the model parameters. The parameters of the loss function, including the weight decay for L2 regularization, were determined through a process of hyperparameter tuning, ensuring optimal performance of our model in the emotion classification task.

2.3. Implementation details

In the execution of the suggested model, several important aspects are underscored. Firstly, pertaining to hyperparameters: the learning rate is established at 0.001 and it decreases by a tenth whenever a standstill in the validation loss is noticed. A batch size of 16 is used, and the model trains for a total of 100 epochs. The choice of optimizer is the Adam optimizer due to its efficient handling of gradient descent in high-dimensional spaces. Data augmentation techniques are applied to the dataset to increase its size and diversity, and to reduce overfitting. These techniques include random rotation, horizontal flipping, and random scaling of the input images. This ensures that the model is exposed to a variety of perspectives and scales of the facial images during training, which improves its robustness and ability to generalize to unseen data. Considering the images' background, the dataset is composed of grayscale facial images where the faces are predominantly centred and consume nearly the same space in every image. As a result, any background present is relatively uniform across images and doesn't necessitate specific treatment. It is assumed that the model learns to focus on the facial features for the emotion classification task.

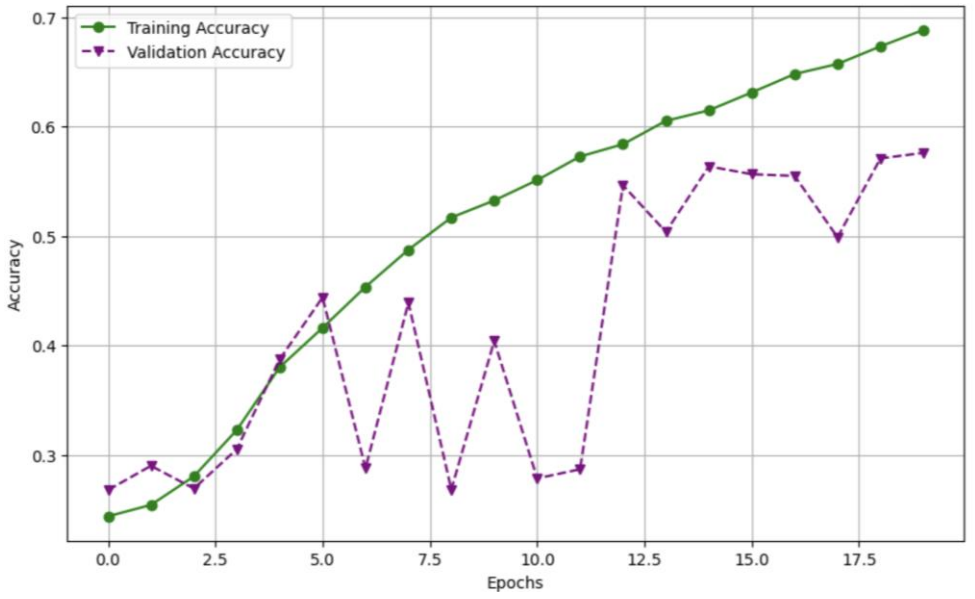
3. Results and discussion

In the conducted study, a hybrid model comprising the ResNet50 and FPN was utilized to perform facial emotion recognition from a collection of over 20,000 images, each labelled with a specific emotion. A comparative analysis of the accuracy of the hybrid model against that of the standalone ResNet50 model is provided in Figure 5 (a) and Figure 5 (b).



(a) The result of Resnet50 + FPN model.

DOI: 10.54254/2755-2721/21/20231109

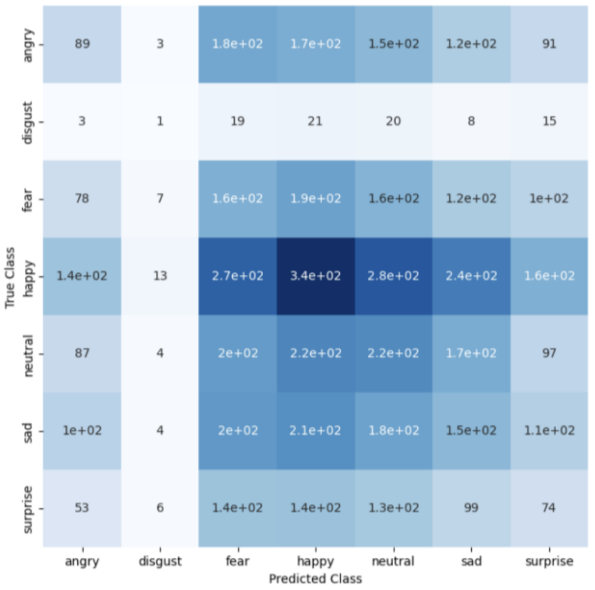


(b) The result of Resnet50 model

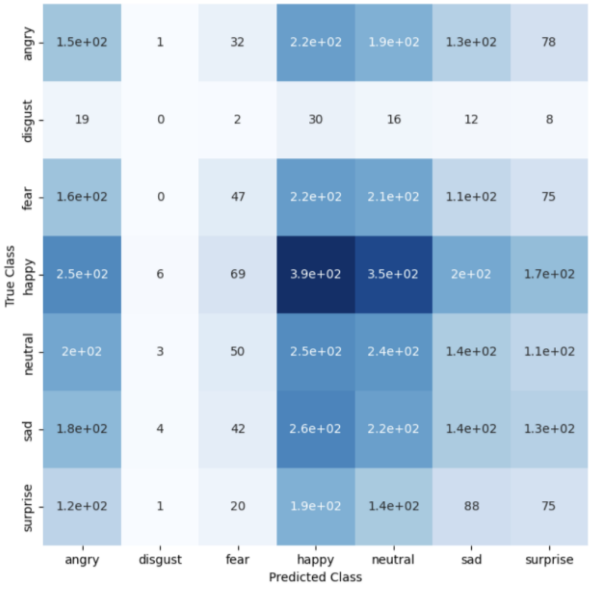
Figure 5. The result curves of 2 models.

As can be seen from the Figure 5, the ResNet50+FPN model achieves 93% validation accuracy after only a few calendar hours of training, while the accuracy of the standalone ResNet50 model exhibits greater volatility before stabilizing at a lower final value. In addition, the ResNet50+FPN model exhibited higher initial accuracy, suggesting that it provides a more efficient solution for cold-start scenarios. The superior performance of the ResNet50-FPN model is attributed to the hierarchical feature learning facilitated by the FPN structure. Merging the FPN with ResNet50 allows for the combination of high level semantic information and low level detail to produce a diverse and rich feature set that is more suitable for complex facial emotion recognition tasks. pre-training of the ResNet50 backbone is a key step in improving the overall performance of the model through learning rate optimisation, extended training time and the application of enhancements such as TrivialAugment, Random Erasing, MixUp and CutMix are implemented. These strategies greatly enhance the accuracy of the ResNet50 model. Relative to previous studies, this model showed an improvement in accuracy of 25% over the standalone ResNet50 model. This highlights the significance of merging different neural network architectures to enhance the performance of emotion recognition models.

DOI: 10.54254/2755-2721/21/20231109



(a) The confusion matrix of Resnet50 + FPN model



(b) The confusion matrix of Resnet50 model

Figure 6. The confusion matrix of the 2 models.

Figure 6 (a) and Figure 6 (b) provides an illustrative summary of the model's performance using two confusion matrices, one presenting the overall performance, and the other detailing the performance per emotion class. These matrices visually represent the distribution of true and predicted labels, providing insights into the model's strengths and weaknesses. It is notable that the model exhibited a minor challenge in accurately classifying the 'disgust' emotion due to its sparse representation in the dataset, as evident from the confusion matrices and corroborated. This observation underscores the need for more balanced datasets to achieve a more uniform emotion classification accuracy across all classes.

DOI: 10.54254/2755-2721/21/20231109

In summary, the integration of the FPN into the ResNet50 model resulted in substantial improvements in facial emotion recognition performance. The confusion matrices visually articulate this enhancement while also spotlighting areas for further refinement. Future research could consider the exploration of alternative combined network architectures or advanced training strategies to further optimize emotion recognition accuracy.

4. Conclusion

The integration of a FPN with ResNet50 stands as the key contribution of this study. Experiments conducted on the FER-2013 dataset show that this combination outperforms traditional methods, both in terms of accuracy and adaptability to objects of diverse sizes and complexities. The model, which is a combination of ResNet50 and FPN, effectively recognizes emotional expressions in images, achieving an accuracy of 93%. This outcome signifies a remarkable 25% enhancement in comparison to using the standalone ResNet50 model. The enhanced performance of the hybrid ResNet50-FPN model is mainly due to the hierarchical feature learning facilitated by the FPN architecture. Merging the FPN with ResNet50 creates a blend of high-level semantic characteristics and low-level detailed attributes. This amalgamation produces a diversified and rich set of features that are more suited to the complex task of facial emotion recognition. Future research will take into consideration potential limitations and explore further the performance of the model across diverse facial features and emotional states. More advanced deep learning architectures will also be examined to enrich the model's interpretive ability. ‘Words-emotion' has been identified as the next research objective, with emphasis placed on an in-depth analysis of the ‘semantic meaning of emotion' aspect. This will hopefully contribute to a more comprehensive understanding of the complexities in facial expression recognition.

References

[1] Bettadapura V 2012 Face expression recognition and analysis: the state of the art arXiv preprint arXiv:1203.6722
[2] Happy S Routray A 2014 Automatic facial expression recognition using features of salient facial patches IEEE transactions on Affective Computing 6(1): pp 1-12
[3] De S Liyanage C Suen C 2003 Real-time facial feature extraction and emotion recognition Fourth international conference on information, communications and signal processing IEEE pp 1310-1314
[4] Balasubramanian B 2019 Analysis of facial emotion recognition 3rd International Conference on Trends in Electronics and Informatics (ICOEI) IEEE 2019 pp 945-949
[5] Kartali A 2018 Real-time algorithms for facial emotion recognition: A comparison of different approaches 14th Symposium on Neural Networks and Applications (NEUREL) IEEE pp 1-4
[6] Bargal S Barsoum E Ferrer C 2016 Emotion recognition in the wild from videos using images 18th ACM International Conference on Multimodal Interaction ACM pp 433-436
[7] Lin T 2017 Feature pyramid networks for object detection IEEE conference on computer vision and pattern recognition IEEE pp 2117-2125
[8] Li B Lima D 2021 Facial expression recognition via ResNet-50 International Journal of Cognitive Computing in Engineering 2: pp 57-64
[9] Zhao G Ge W Yu Y 2021 GraphFPN: Graph feature pyramid network for object detection IEEE international conference on computer vision (CVPR) IEEE pp 2763-2772
[10] Giannopoulos P Perikos I Hatzilygeroudis I 2018 Deep learning approaches for facial emotion recognition: A case study on FER-2013 Advances in Hybridization of Intelligent Methods: Models, Systems and Applications pp 1-16
[11] Goodfellow I Erhan D Carrier P et al 2013 Challenges in representation learning: A report on three machine learning contests Neural Information Processing: 20th International Conference (ICONIP) Springer berlin Heidelberg pp 117-124