Some Facial Landmarks Should Be Left Alone

Cheng Qiu
Vanderbilt University

Abstract

Understanding emotions play a pivotal role in human communication, with facial expressions providing key insights into one's emotional state. This study investigates the impact of facial landmarks on emotion classification by introducing perturbations, such as physical masks, to a facial emotion dataset (MaskFer). We trained VGG16 models on FER2013 and evaluated their performance on both FER2013 and MaskFer, observing significant performance drops, particularly for emotions like happy and surprise, while accuracy for disgust improved. The analysis revealed that models trained on FER2013 did not always capture the most salient features for all emotions, suggesting that focusing on individual facial features might enhance emotion recognition. We propose a Perturb Scheme comprising three phases: attention-based classification, clustering of important pixels, and training a new classifier. Our experiments, conducted on the FER2013 dataset using various models, demonstrate that the Perturb Scheme can effectively improve classification accuracy by emphasizing regional features over holistic ones. This approach highlights the potential benefits of isolating and leveraging individual facial features in emotion recognition tasks.

1 Introduction

Understanding emotions plays a pivotal role in how we perceive and interact with one another. When interpreting facial emotions, key landmarks such as the eyes and mouth provide significant insights into a person's emotional state [2]. Additionally, faces can be divided into two halves, focusing on the eyes and eyebrows on one side and the mouth on the other [7]. To enhance our comprehension of facial emotions, it is essential to understand how these facial landmarks contribute to emotion prediction.

To investigate the impact of facial landmarks on emotion classification, we introduced perturbations, such as a physical mask, to a facial emotion dataset, referred to as MaskFer [3]. We then trained VGG16 models on Fer2013 and tested them on both FER2013 and MaskFer. The performance results are presented in Table 1.

Model	**Angry**	**Happy**	**Sad**	**Neutral**	**Surprise**	**Disgust**	**Fear**
DenseNet [5]	-39.1%	-48.5%	12.2%	-51%	-18.8%	-49.4%	-20%
DPN [1]	-34.8%	-62.8%	23.9%	-59.6%	-26.4%	-58.72%	-22.8%
ResMasking [8]	-31.6%	-42.2%	-4.3%	-52.7%	-34.5%	-39.1%	-12.6%
ResNet [4]	-19.9%	-54.6%	16%	-59.7%	-19.3%	-49.2%	-18%
VGG16 [11]	-22.8%	-35%	-7.9%	-43.2%	-20.1%	-49.3%	-17.2%
Ensemble	-31.6%	-42.2%	-4.3%	-52.7%	-34.5%	-39.1%	-12.6%

Table 1: Accuracy gain/loss when facial mask is applied to Fer2013

The accuracy and F1 scores for each emotion class decreased when models trained on Fer2013 were tested on MaskFer. Notably, the rate of accuracy and F1-score decline varied across different classes. For emotions like happiness and surprise, the model's accuracy dropped by up to 85%, whereas for emotion sad, there was a slight increase in accuracy in some cases, as shown in Table 1. In this case, masking the mouth allowed the model to extract better information relevant to the emotion of sad such as the eyebrow.

By examining the saliency map in Figure 1, it is evident that for emotions such as disgust, anger, and sadness, the model trained on Fer2013 may not have learned the optimal salient features. In contrast, the

model trained on MaskFer was able to capture the features in the eyebrows when identifying an angry face. Therefore, current models trained on Fer2013 do not effectively leverage all the information present in the vicinity of major facial landmarks like the eyebrows. While classifying emotions based on overall facial landmarks, considering them as individual features has potential advantages. By extracting and leveraging information from regional features, new classifiers can gain valuable insights without relying on the presence of specific features.

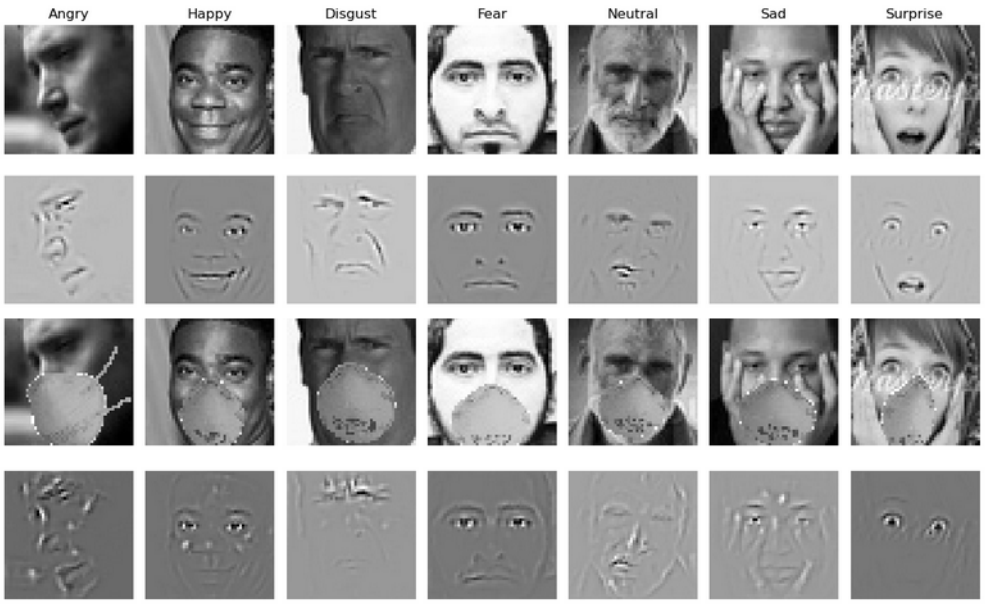

Figure 1: Rows 1 and 3 are the unmasked image and masked image respectively and Rows 2 and 4 are the saliency map of the trained VGG16 model on Fer2013 and MaskFer respectively

This procedure was applied to other models in facial emotion recognition tasks, and a similar trend was observed, as shown in Table 1. This further suggests that some facial landmarks may distract models from identifying the most important landmarks for specific emotions.

2 Related Works

Neural networks are a natural choice for tasks like facial emotion recognition due to their recent advancements. One notable example is ResNet [4], which enabled deeper networks by using skip connections to propagate information between layers. Although deeper networks can increase accuracy, they also add computational complexity. Alternatively, DenseNet [5] introduced direct connections between layers to mitigate computational complexity while maintaining comparable performance. More recently, Dual Path Network (DPN) [1] leveraged the benefits of both ResNet and DenseNet with a dual path configuration, improving classification accuracy over its predecessors.

Within the FER2013 benchmark, a facial recognition dataset, recent work has built on top of convolutional neural networks (CNNs) by integrating attention mechanisms that focus on the main features of the face [6]. Another model based on the VGG architecture achieved a single-network classification accuracy of 76% on the FER2013 dataset by tuning hyperparameters such as kernel size and learning rate using genetic algorithms [10]. Another approach applied spatial transformations to enhance the concept of differentiable attention, addressing the impact of rotations, translations, and other transformations on accuracy [2]. Additionally, ensemble methods combining multiple models have shown to improve the accuracy of facial emotion detection by a small margin [9].

Emotion classification in images with perturbations, such as partial face blockage, presents distinct challenges from standard facial emotion recognition. One dataset exploring this area is MaskFer, which includes

images where the bottom part of the face is masked [3]. Research in this area often employs CNNs to identify masks and then estimate emotions based on the visible facial features. This approach acknowledges the significant impact masks have on emotion recognition by covering key expressive areas of the face [7].

Recent studies have advanced the field of facial emotion recognition for individuals wearing masks by employing innovative deep learning approaches and focusing on facial landmarks. One study proposed a method for recognizing emotions on masked faces by enhancing low-light images and analyzing upper facial features using CNNs. This approach utilized the AffectNet dataset, containing over 420,000 images spanning eight types of facial expressions. By covering the lower part of the face with a synthetic mask, the study used boundary and regional representation methods to highlight the head and upper facial features. Feature extraction was performed based on the detected facial landmarks of the partially covered face, integrating landmark coordinates and histograms of oriented gradients into the classification process using a CNN [7].

3 Methodology

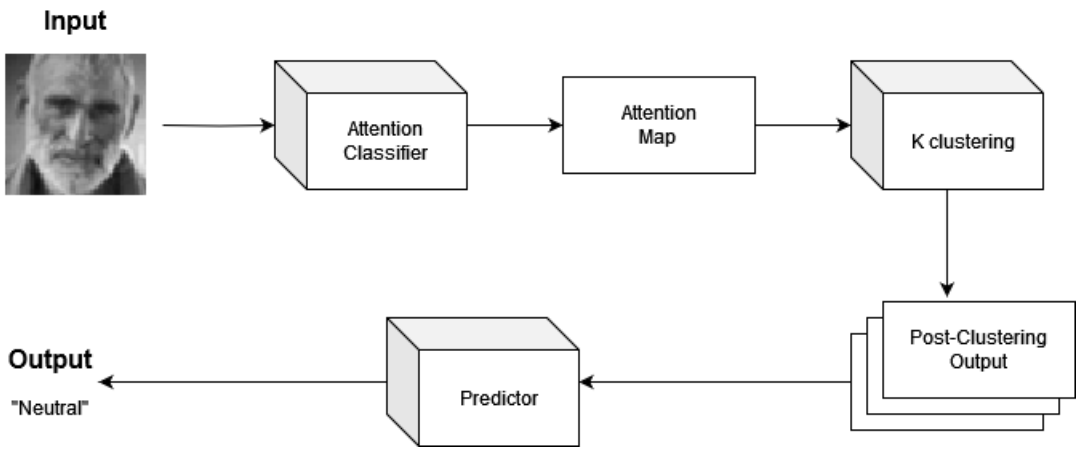

Figure 2: Framework for Perturb Scheme

To fully leverage the regional characteristic in facial recognition, we propose a new training scheme, the Perturb Scheme, shown in Figure 2 that includes threes phases. The first phase focuses on training an attention classifier to better pin point the important pixels. The second phase involve clustering those pixels into classes base on importance. Thirdly, training a new classifier to perform the emotion classification task with the inputs from the clustering. Refer to Figure.

Attention classifier: The first phase is training a neural network focused on understanding the spatial attention in order to better understand the influence of each pixel toward the prediction of all emotion. Since different facial landmarks may have different effects on the performance of emotion predictions. To achieve this, spatial attention modules were added between layers of the neural networks [12].

Clustering: The second phase is to isolate local patches of pixels that have the highest attention. To identify those patches, we will use K-mean clustering with n clusters to represent n meaningful facial landmarks. By leveraging the attention, pixels is clustered into different classes to be masked out based on the euclidean distance shown in Equation 1. Due to the computation required during clustering, it's unfeasible to perform clustering for each image, instead the approach that we took is updating the clusters every epoch.

Given an $I \in \mathbb{R}^{BXCXWXH}$ and its spatial attention map $S \in \mathbb{R}^{BXWXH}$, the distance between any two points on the 2D grid P_{ij} and P'_{kl}, where $i, k \in W$ and $j, l \in H$, can be compute using the following:

$$\text{Dist} = \sqrt{(\lambda(i-k))^2 + (\lambda(j-l))^2 + (\alpha(I_{ij} - I_{jl}))^2} \tag{1}$$

λ and α are constants used to control the weights of the pixel distance on the grid and the intensity distance.

Predictor: The third phase is training another neural network to learn from the clustered data in order to optimize the emotion detection task.

4 Experiment

We empirically demonstrate the effectiveness of the Perturb Scheme in enhancing the capabilities of deep learning models for emotion recognition. Our performance comparisons include prominent deep learning architectures such as DenseNet, ResNet, and others.

4.1 Dataset

FER2013 dataset is a publicly available well-known dataset for facial emotion recognition. The dataset consists of 35887 gray scaled images of emotions The FER2013 dataset, a well-known publicly available resource for facial emotion recognition, consists of 35,887 grayscale images across seven different emotion classes. Of these, 28,708 images are used for training, while 3,589 images are designated for validation and testing, making the dataset suitable for deep learning applications. However, FER2013 contains incorrectly labeled faces, irrelevant images, and an imbalanced distribution of image data, with some classes containing up to 7,000 images and others as few as 500 (see Figure 3).

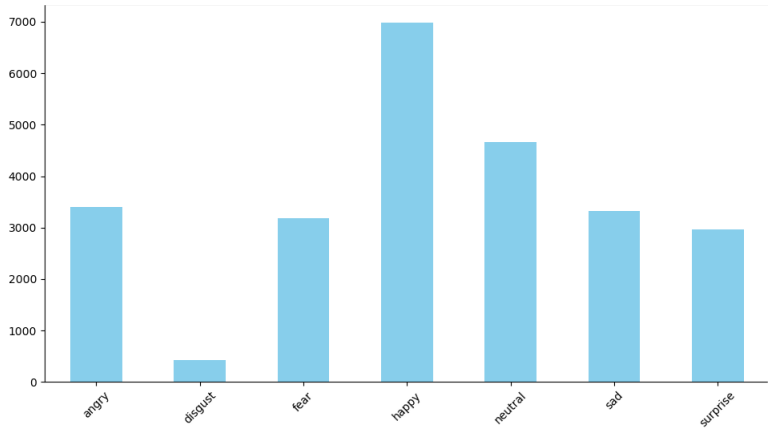

Figure 3: Distribution of classes in Fer2013 dataset

4.2 Training

All models were trained on an NVIDIA RTX 3090 for 200 epochs using cross-entropy loss as the primary loss function. Stochastic Gradient Descent (SGD) with Nesterov momentum (0.9) and weight decay was employed. The initial learning rate was set to 0.01 and adjusted during training using ReduceLROnPlateau (RLRP) [9]. For data preparation, we adopted the image augmentation techniques recommended by Zhong et al., including random flipping, cropping, and erasing, to prevent over-fitting, especially given the scarcity of samples in some classes [13]. For the Perturb Scheme, the parameters α and λ in Equation 1 were set to 1.2 and 1.5 respectively with the number of clusters for the k-mean clustering set to 3.

5 Evaluation and Analysis

For evaluation, all models were trained on FER2013 and tested using the testing set from the same dataset. We selected several models commonly used for emotion classification, including ResMasking [8] and ResNet

[4], as well as baseline models like VGG16 [11], to assess the performance of the proposed Perturb Scheme. Given its lightweight nature, the Perturb Scheme can be easily integrated into any backbone network. Our results demonstrate that the proposed model yields positive outcomes. **Note:** Models trained with the Perturb Scheme will be prefixed with "P" to distinguish them from the baseline models.

5.1 Attention-Based Clustering

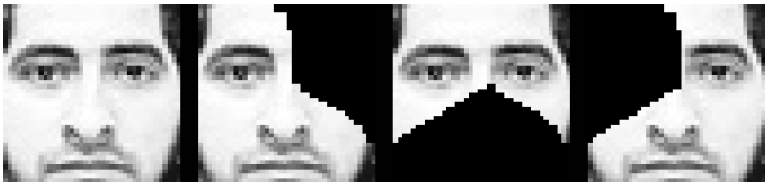

Figure 4: Facial image after applying clustering $n = 3$ based on the attention

In the context of spatial attention for emotion classification tasks, our models predominantly utilize local regions around the eyes and mouth. The attention classifier, as illustrated in Figure 4, underscores the significance of these areas, with two clusters focusing on the eyes and one on the mouth and nose region. Regarding the clustering parameters depicted in Figure 5, we conducted experiments varying the number of clusters. The highest performing models aligned with three clusters, corresponding to the most meaningful features in this task. Increasing the number of clusters, which also heightens computational demands, did not consistently enhance accuracy and, in some instances, led to a decrease in performance.

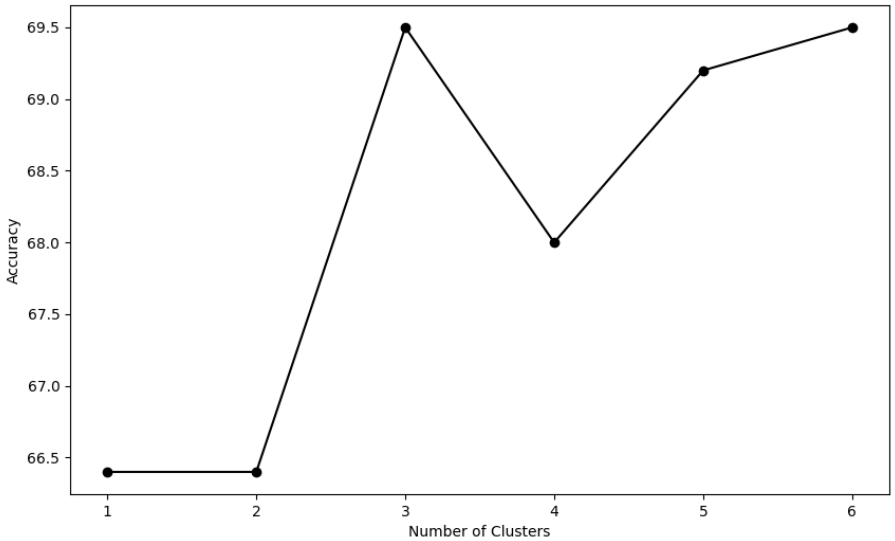

Figure 5: Performance for different number of clusters

5.2 Performance

Across all emotion class, except for disgust, the PVGG experienced a considerable improvement in accuracy of up to 7%. As for performance for class-wise emotion classification, there's generally a positive change in the accuracy with the exception of disgust. The decrease in the emotion disgust could be due to the fact the model was not able to emphasize the mouth enough as shown in Figure 6. However for other emotions,

2.1 Summary (Baseline)

Model	Accuracy	F1
DenseNet [5]	72.6%	0.70
DPN [1]	71.2%	0.68
ResMasking [8]	66%	0.63
ResNet [4]	69.2%	0.685
VGG16 [11]	66.5%	0.653
Ensemble	73.4%	0.71

2.2 Summary (Perturb)

Model	Accuracy	F1
DenseNet [5]	–	–
PDPN	71.9%	0.69
ResMasking [8]	–	–
PResNet	72.3%	0.70
PVGG	69.5%	0.66
PEnsemble	40%	0.35

Table 2: Table 2.1 and Table 2.2 illustrate the performance comparison between classifier trained without and with Perturb Scheme

removing facial landmarks such as mouth and eyes did result in significant increase in accuracy. Comparing the saliency map for the baseline VGG16 model and the proposed model PVGG shown in Figure 6, it can be observed for emotions such as fear, PVGG is able to put more emphasis on the important features such as the mouth and less emphasis on less important features like the nose as compared to VGG16. Other models such as ResNet and DPN also experienced slight increase across the board in accuracy for most emotion classes.

Model	Angry	Happy	Sad	Neutral	Surprise	Disgust	Fear
DenseNet [5]	–	–	–	–	–	–	–
DPN [1]	0.8%	2%	1.5%	1.4%	-0.5%	-3.6%	3%
ResMasking [8]	–	–	–	–	–	–	–
ResNet [4]	3.5%	1%	1.5%	-0.6%	2.1%	5.6%	8.4%
VGG16 [11]	12.2%	-1.2%	-17.3%	2.2%	1.6%	5%	18%
Ensemble	–	–	–	–	–	–	–

Table 3: Changes in class-wise performance of perturb scheme over the baseline on Fer2013 dataset

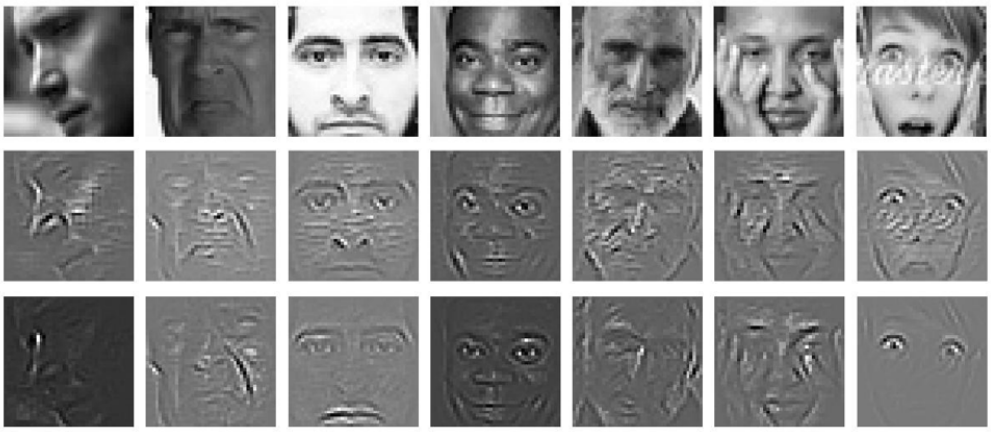

Figure 6: Comparison of the saliency map between VGG16 and PVGG. From the top row to bottom are the original facial image, the salient map for VGG, and the salient map for PVGG respectively

6 Conclusion

Our findings demonstrated that incorporating regional feature extraction through the Perturb Scheme can improve the performance of emotion recognition models, particularly when dealing with occluded facial

regions. The application of attention-based clustering and subsequent training on clustered data enables the models to focus on the most informative facial landmarks, enhancing their predictive accuracy. While the overall performance gains were evident across most emotion classes, certain emotions like disgust showed a decrease in accuracy, indicating the need for further refinement in handling specific facial expressions. This research highlights the potential of regional feature emphasis in advancing facial emotion recognition technologies.

References

[1] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks.

[2] Yassine El Boudouri and Amine Bohi. EmoNeXt: an adapted ConvNeXt for facial emotion recognition. In *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. ISSN: 2473-3628.

[3] Bin Han, Hanseob Kim, Gerard Jounghyun Kim, and Jae-In Hwang. Masked FER-2013: Augmented dataset for facial expression recognition. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 747–748. IEEE.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition.

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks.

[6] Shervin Minaee and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network.

[7] Mukhriddin Mukhiddinov, Oybek Djuraev, Farkhod Akhmedov, Abdinabi Mukhamadiyev, and Jinsoo Cho. Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people. 23(3):1080. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[8] Luan Pham, The Huynh Vu, and Tuan Anh Tran. Facial expression recognition using residual masking network. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4513–4519. IEEE.

[9] Christopher Pramerdorfer and Martin Kampel. Facial expression recognition using convolutional neural networks: State of the art.

[10] Muhammad Sam'an, Safuan Safuan, and Muhammad Munsarif. Convolutional neural network hyperparameters for face emotion recognition using genetic algorithm. 33(1):442–449. Number: 1.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition.

[12] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module.

[13] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation.