Title: Enhancing Face Emotion Recognition with Advanced ML Methods

Introduction

The significance of facial emotion recognition (FER) has grown significantly in recent times, given its wide-ranging utility in fields such as human-computer interaction, healthcare, social robotics, and security systems. FER is a challenging problem that involves recognizing and interpreting human emotional states expressed through facial expressions. Accurately identifying the emotional states associated with various facial expressions, such as anger, disgust, fear, happiness, sadness, and surprise, is the ultimate objective of FER.

Despite the advances in FER research, there are still several challenges that need to be overcome to enhance its precision and dependability. Notably, the absence of dependable expression data is a crucial issue. Most existing FER databases are collected in a controlled laboratory environment, limiting the scale and diversity of expression data. Although recent databases, such as FERPlus and RAF-DB, have been labeled more reliably, they are still limited in scale. Additionally, the quality of expression labels in some databases, such as AffectNet, is poor, making it challenging for models to learn real expression features.

The variability of facial expressions in uncontrolled environments is also a significant challenge. Factors such as lighting, occlusion, posture, and individual differences can affect the recognition of facial expressions. For example, wearing masks or sunglasses can make it challenging to apply FER in real-life situations.

To address these issues, this research project proposes a method to improve the FER recognition rate based on the static image library and existing FER methods. The proposed solution includes using transfer learning to transfer object or face recognition models to the expression recognition task to address the lack of reliable expression data. Utilizing the pre-trained weights of existing models via transfer learning can enhance the FER model's accuracy. Furthermore, to tackle the issue of variability in facial expressions caused by occlusion and posture, local facial information can be employed for expression recognition. The proposed approach can enhance the model's ability to recognize facial expressions even when a portion of the face is occluded.

The present study employs a methodology that integrates two state-of-the-art deep learning techniques: the region attention network (RAN) and transfer learning. Firstly, RAN is utilized to process most facial images, producing facial feature representations with enhanced expressive power via the incorporation of both the self-attention mechanism and the region-attention mechanism, which weight different regions in the face image. Subsequently, the outputs of RAN are fed into the facial expression recognition (FER) model based on a large-scale image dataset obtained through transfer learning. Aiming at a higher accuracy rate of emotion recognition.

The proposed research has the potential to enable innovative applications in various fields. For example, accurate FER can aid in diagnosing and monitoring neurological disorders like Parkinson's disease that affect facial expressions. It can also be useful in mental health assessments by providing an objective measure of emotional state.

Research Problem

A. Sub-problems

To address the limitations of current FER models and improve their performance, the main research question is "How can we enhance face emotion recognition with advanced machine learning methods?" Sub-problems are proposed to guide the research:

RQ1: How can we increase the size of reliable expression data for FER?

The question arises from the requirement of a considerable amount of dependable expression data to effectively train a robust and accurate FER system. However, due to factors such as diversity, subjectivity, and temporality of human expressions, producing high-quality expression datasets is an arduous task, resulting in datasets that are both scarce and noisy. Hence, it is crucial to investigate approaches for augmenting dependable expression datasets to elevate the performance of FER systems.

RQ2: How can we improve FER accuracy in uncontrolled environments?

The question arises because FER systems are typically trained in controlled lab settings with standardized conditions, but in real-world scenarios like social robots or uncontrolled environments, subjects display diverse expressions under varying conditions. Thus, improving the robustness and generalizability of FER systems is crucial for their practical applicability.

RQ3: How to obtain the facial expression recognition model through transfer learning?

The reason for asking this question is that transfer learning is a valuable technique that can enhance the accuracy of a model trained on a limited dataset by utilizing insights gained from a larger dataset. In facial expression recognition, transfer learning involves using a pre-trained model from a large image dataset, like FaceNet, and fine-tuning it on a smaller expression dataset. It is important to explore how to effectively obtain and use pre-trained models through transfer learning.

B. Previous work

Ng and colleagues (2015)[4] applied transfer learning using a pre-trained CNN on ImageNet to improve emotion recognition on small datasets, achieving better accuracy than the baseline. Ding and colleagues (2017)[2] proposed an algorithm, FaceNet2ExpNet, that leverages facial domain knowledge and standardized training of expression recognition networks to improve performance on four public expression databases. Uçar (2017)[6] improved facial expression recognition with a deep CNN algorithm, utilizing the CNN model architecture and NVIDIA Tegra TX1 platform. Jain and colleagues (2018)[3] proposed a deep learning-based facial emotion(FER) recognition model that uses a hybrid model of convolutional neural network and recurrent neural network, as well as explores new feature-level fusion methods to

enhance emotion prediction accuracy. Wang and colleagues (2020)[7] introduced a novel facial expression recognition method, Regional Attention Network (RAN), which adapts to facial regions under occlusion and pose changes, achieving superior results in multiple popular datasets with the use of Region Bias Loss (RB-Loss) to emphasize high attention weighting on the most important regions. Akhand et al. (2021)[1] proposed a FER system that utilizes transfer learning with deep CNNs to avoid training on large amounts of data. The system achieves remarkable accuracy on benchmark face images from front and side views, with promising performance on the side-view KDEF dataset for industrial applications.

C. Advancements of my Work over Previous Studies

Previous studies have demonstrated significant advancements in facial emotion recognition through the use of sophisticated techniques, such as neural networks and deep learning. Nevertheless, challenges still exist, including inadequate size and diversity of training datasets, as well as low recognition rates of facial expressions in uncontrolled environments. This article aims to address the limitations of current methods by focusing on enhancing the scale of existing datasets and improving techniques for processing occluded images, ultimately leading to improved accuracy of facial emotion recognition. We aim to improve the accuracy and reliability of facial emotion recognition, making it more useful in real-world applications such as healthcare, social robotics, and security systems. The proposed approach has the potential to significantly advance the field of facial emotion recognition and help overcome some of the current limitations in this area.

Methodology

A. overview

In order to improve the accuracy of facial emotion recognition (FER), and solve the problems of insufficient scale of existing data sets and possible occlusion of pictures, etc.,

The combination of the emotion recognition model（RAN） obtained by FaceNet through transfer learning and Region Attention Network can further enhance the performance of facial emotion recognition. Region Attention Network, which is an attention mechanism, can extract region-level features from images.

Specifically, the facial image is initially input into a Region Attention Network, which generates a set of feature maps, with each feature map corresponding to a unique region of the face. After extracting the feature map of each facial region, the emotion recognition model obtained by FaceNet through transfer learning utilizes these features as input to obtain the corresponding facial expression features. These facial expression features are then merged to form a comprehensive representation of the overall facial expression. The schematic representation illustrating the operational principle is presented in the abstract as follows.

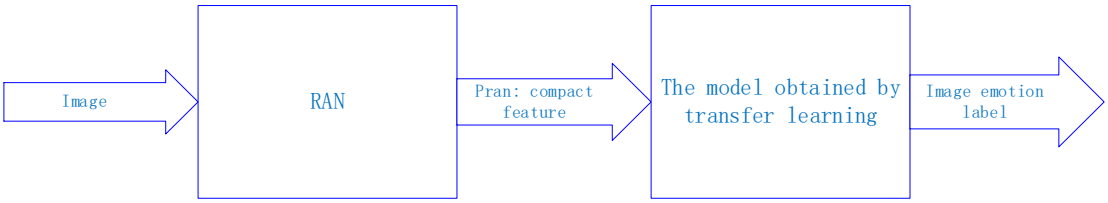


Figure 1 Method Abstract Diagram

B. Dataset

AffectNet is a facial expression dataset that contains around 400,000 images, each of which has been manually labeled for eight facial expressions, including neutrality, happiness, anger, sadness, fear, surprise, disgust, and contempt, as well as the intensity of valence and arousal. Since the transfer model used in this study is based on Facenet, there is no need for an additional training set. Therefore, the AffectNet dataset was used as a testing dataset. This dataset was chosen for its large size, comprehensive labeling, and the inclusion of some occluded portraits, making it a suitable choice for both training and evaluating facial expression recognition models.

Figure 2 https://www.catalyzex.com/paper/arxiv:2103.16854

C. RAN

Wang et al. (2020)[7] proposed a approach for facial expression recognition called the regional attention network (RAN), which can effectively handle occlusion and pose changes by adaptively capturing the importance of different facial regions. RAN is comprised of two stages: the first stage involves a self-attention module that roughly calculates the significance of each region, while the second stage uses a relational attention module to model the relationship between regional features and global features. Figure 3 illustrates the two stages of RAN and how they work. This approach has been shown to outperform existing methods and is a promising direction for future research in the field of facial expression recognition(FER).

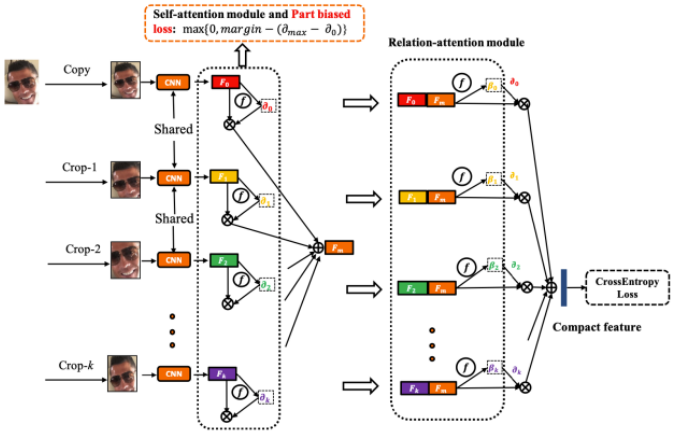


Figure 3 The framework of RAN. (Quoted from Wang et al., 2020)

a) Region Generation

The fundamental operation of RAN involves partitioning a portrait image into multiple segments. Since the images in the selected image library predominantly comprise faces as the primary subject, cropping at fixed positions is a feasible approach that preserves a significant portion of the portrait's feature information. The fixed position cropping process divides the selected image into six distinct regions, namely Top left, Top middle, Top right, Bottom left, Bottom middle, and Bottom right, which are labeled as I1 to I6. The CNN takes in the six segments shown in Figure 4, which are first cropped to 3/4 of the image size and then adjusted in size to correspond with the input dimensions of the network.

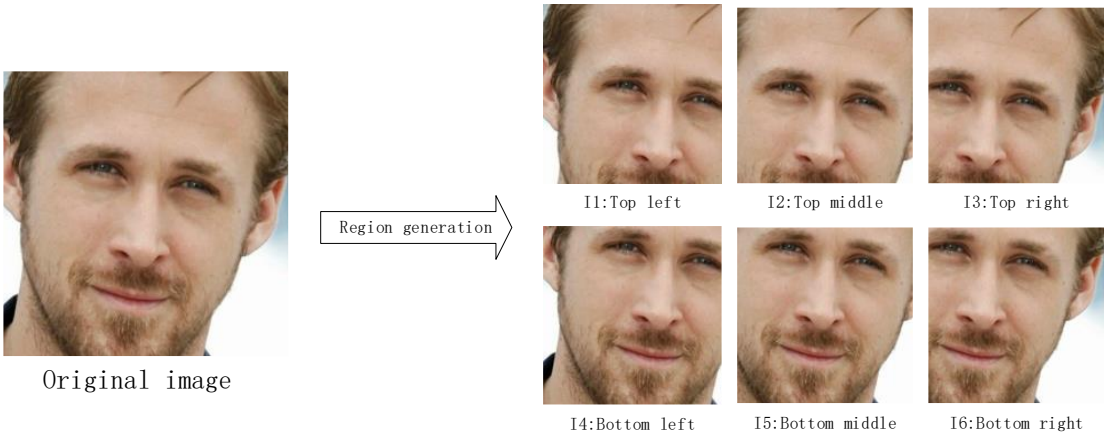


Figure 4: Schematic diagram of Region Generation

b) Region Feature Extraction

After cropping, put the pictures of the six regions into the backbone CNN for feature extraction, the formula is as follows

$I1, I2, \cdots, Ik$, *and the self attention as* $r(\cdot;\ \theta)$. *The feature set* X *of* I *is defined by*:

$$X = [F0, F1, \cdots, Fk] = [r(I0;\ \theta), r(I1;\ \theta), \cdots, r(Ik;\ \theta)]$$

c) Self-attention module

The self-attention module utilizes region features to compute rough attention weights by employing a fully connected (FC) layer and a sigmoid function. Specifically, the attention weight for the i-th region is defined mathematically as follows:

$$\mu i = f\ (Fi.T * q0)$$

The parameter q0 of the fully connected (FC) layer and the sigmoid function f are involved in the computation of attention weights.

The passive summarization of region features, along with their corresponding attention weights, results in a global representation Fm, which is obtained as follows:

$$Fm = \frac{1}{\sum_{i=0}^{n} \mu i} \sum_{i=0}^{n} \mu i\ Fi$$

The global representation Fm is a condensed representation and can be utilized as the final input for the classifier.

The purpose of utilizing the RB-Loss is to enhance the accuracy of facial expression recognition by assigning higher attention weights to the most significant regions and adaptively addressing errors caused by occlusions and pose variations.

$$Lrb = max\{0, \alpha - (\mu max - \mu 0)\}$$

d) Relation-Attention Module

The self-attention module employs a coarse-grained approach to learn weights for individual features through non-linear mapping. However, as the aggregated representation Fm already encapsulates the contents of all facial regions, attention weights can be enhanced by incorporating the relationship between region features and the global representation Fm.

The sample concatenation and another FC layer are utilized to estimate new attention weights for region features. The formulation of the new attention weight for the i-th region in the relation-attention module is as follows:

$$vi = f([Fi : Fm].T * q1)$$

which q1 is the parameter of FC, and f is the sigmoid function.

The region information and the coarse global representation from self-attention are combined to obtain a new compact feature as follows:

$$Pran = \frac{1}{\sum_{i=0}^{n} \mu i \; vi} \sum_{i=0}^{n} \mu i \; vi[Fi: Fm]$$

e) Output of RAN

Pran refers to a set of vectors that are derived from the weights assigned to six distinct regions after they have passed through the RAN. These weights correspond to the information content of each region, with higher weights indicating lower occlusion and, consequently, greater availability of information for facial emotion recognition. The use of such a weighting scheme facilitates the representation of information content in a manner that is well-suited for the accurate recognition of facial emotions.

D. Classifier

a) Transfer learning

Transfer learning refers to the method of applying a model that has been trained on one task to another related task. In transfer learning, already trained models are often referred to as pre-trained models, which can be fine-tuned on new tasks by sharing the underlying feature extractors.

b) Model choice for preprocessing

Our selected preprocessing model is the FaceNet model, which trains a direct mapping from facial images to a condensed Euclidean space. In this space, the distances between points indicate the degree of similarity between faces, as demonstrated by Schroff et al. in 2015.[5] The rationale for selecting the Facenet model lies in the fact that it is trained on the Labeled Faces in the Wild (LFW) and Youtube Faces DB datasets, which are widely used as benchmarks for face recognition and verification. In particular, FaceNet was trained on approximately 1-2 million face thumbnails, representing approximately 8 million distinct identities. The LFW dataset includes more than 16,000 facial images from over 13,000 individuals, while the Youtube Faces DB comprises approximately 3,500 hours of video

featuring 3,425 identities, with 1,595 of those identities appearing in multiple videos.

Consequently, leveraging transfer learning based on the Facenet model can address the issue of limited size of the emotion recognition database.

c) FaceNet2ExpNet

According to Ding et al. (2017)[2], the FaceNet2ExpNet algorithm is used to convert the FaceNet model into an emotion recognition model by training the expression recognition network on static images to improve its performance, while utilizing facial domain knowledge to normalize the training. This is achieved by modeling the higher-level neurons of the expression network using a new distribution function based on information derived from fine-tuning the face network. The algorithm trains the convolutional layer and the fully connected layer through a two-stage training process, which improves the performance of the network. We will use this algorithm to process the results obtained from RAN and assign emotional labels to portraits

d) FER Pseudocode using Face2ExpNet algorithm

'example.jpg' is the region with the highest weight value after RAN processing.

```
# Load FaceNet2ExpNet model
model = load_model('FaceNet2ExpNet.h5')
# Load image to classify
image = load_image('example.jpg')
# Preprocess image
image = preprocess_image(image)
# Classify emotion from the image
emotion = model.predict(image)
# Convert the result to emotion label
#neutrality, happiness, anger, sadness, fear, surprise, disgust, and contempt
emotion_labels = ['Neutrality', 'Happiness', 'Anger', 'Sadness', 'Fear', 'Surprise', 'Disgust', 'Contempt']
emotion_label = emotion_labels[np.argmax(emotion)]
# Output emotion label
print(emotion_label)
```

E. Baseline select

To establish a baseline for comparison, we adopt four existing models, processing model using RAN（Wang et al., 2020)[7], CNN FER model (Uçar , 2017)[6] and FER model with DCNN and TL(Akhand et al. , 2021)[1]. These models have been previously proposed and evaluated in the literature for the same task as ours. Our strategy is to evaluate the efficacy of various proposed methods by comparing my model performance with that of established baseline models in the field. The baseline models will be implemented and evaluated under the same experimental conditions as our proposed method to ensure a fair comparison. The purpose of using a baseline is to provide a reference for evaluating the effectiveness of our proposed approach and to demonstrate the superiority of our method over the existing ones.

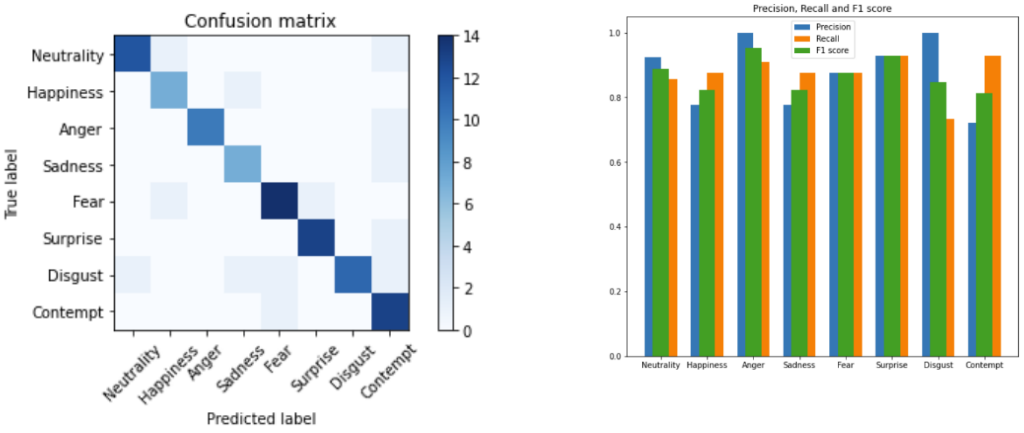
Evaluation Criteria

A. Calculate method

After processing the AffectNet library with RAN and Face2ExpNet models, the FER results were obtained. The predicted labels were saved in the 'predict.csv' file, while the ground truth labels were saved in the 'ground_truth.csv' file.

We will write some code to visualize and generate confusion matrix, accuracy, recall and F1 scores for my model and three baseline models.

The result example is as follows



The performance of a classification model is assessed using a confusion matrix, which compares the model's predicted results with the actual results. Key metrics such as recall, precision, and F1 score can be computed from the confusion matrix. Recall measures the classifier's ability to identify a category correctly. Precision measures the accuracy of the classifier in predicting a category. F1 score is the weighted average of recall and precision and is more suitable for unbalanced datasets. These metrics help us better understand the classifier's performance and improve its accuracy.

B. Evaluation Criteria

Higher values on the main diagonal of the confusion matrix, which correspond to correctly predicted classes, indicate better model performance. Conversely, smaller values on the off-diagonal indicate superior classification accuracy.

The model is deemed to perform well when both the recall and precision values are close to 1, and the F1 score value is also near 1. Furthermore, high consistency across categories is maintained for recall, precision, and F1 scores.

Upon completing all computational procedures, we generated a comparative table displaying the classification results of our proposed method and several state-of-the-art models. Should our model exhibit superior performance in comparison to the aforementioned state-of-the-art models, it may be concluded that the proposed method is both efficacious and

feasible.

Model	Author, year	Recall	Precision	F1-score
My_model				
processing model using RAN	Wang et al., 2020[7]			
CNN FER model	Uçar , 2017[6]			
FER model with DCNN and TL	Akhand et al. , 2021[1]			

The successful improvement of facial emotion recognition through our proposed model would contribute to the enhancement of human-computer interaction experience. Furthermore, it would have broader implications for related research in fields such as medicine, psychology, and sociology. For instance, the model could aid medical practitioners in diagnosing mental illnesses, and researchers could leverage it to enhance our comprehension of human emotions and social behavior.

Conclusion

During the writing process of my case study, I used ChatGPT to help correct any grammatical errors and improve my language. My proposal focuses on advanced machine learning techniques such as predecessors RAN and transfer learning, and presents a novel approach for enhancing facial emotion recognition accuracy. Following this, I compare the advanced facial emotion recognition algorithms with my proposed method. If the results of the comparison confirm that my method performs better, it will demonstrate a valuable contribution to the field of facial emotion recognition.

Reference list

[1] Akhand, M. A. H., Roy, S., Siddique, N., Kamal, M. A. S., & Shimamura, T. (2021). Facial emotion recognition using transfer learning in the deep CNN. *Electronics*, *10*(9), 1036.

[2] Ding, H., Zhou, S. K., & Chellappa, R. (2017, May). Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 118-126). IEEE.

[3] Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., & Zareapoor, M. (2018). Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters*, *115*, 101-106.

[4] Ng, H. W., Nguyen, V. D., Vonikakis, V., & Winkler, S. (2015, November). Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 443-449).

[5] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).

[6] Uçar, A. (2017, July). Deep Convolutional Neural Networks for facial expression recognition. In *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 371-375). IEEE.

[7] Wang, K., Peng, X., Yang, J., Meng, D., & Qiao, Y. (2020). Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, *29*, 4057-4069.