ISSN: 2424-8460(O)
2251-2608(P)

Original Research Article

Facial expression recognition of masked faces using transfer learning

Denezhkina Lidia, Luo Ye
Tongji University, Shanghai, 200092, China

Abstract: Human emotions, reflected through facial expressions, provide valuable insight into an individual's state of mind. Automatic facial emotion recognition (FER) has made significant progress. However, the recognizing emotions on occluded faces remains an area that requires further exploration. We propose a method to address the challenges of partial occlusion, that is, to recognize emotions on faces where the lower part is obscured by a mask. We tackle the classical FER task by employing a transfer learning approach: first, training a teacher model on a dataset of unmasked faces, and then fine-tuning a student model with the learned weights on a dataset of masked faces. Our model integrates a combination of convolutional layers, Inception blocks, Residual blocks, and Squeeze-and-Excitation (SE) networks. We evaluate the proposed approach on three datasets: JAFFE, KDEF, and Oulu-CASIA, achieving accuracy rates of 80.00%, 78.57%, and 89.38%, respectively.
Keywords: facial emotion recognition; occluded facial emotion recognition; face masks, transfer learning; convolutional neural network

1. Introduction

FER is a classification task that aims to identify human emotions from facial images. Similarly, to interpersonal communication in human social activities, machine learning-based facial expression recognition helps AI understand emotions. FER has achieved significant advancements, with models demonstrating increasingly high levels of accuracy and reliability.

The ongoing global COVID-19 pandemic has required the widespread use of face masks as a preventive measure during public interactions. In addition, people may wear masks for various reasons that are not related to the pandemic. Although face masks have become prevalent in many aspects of daily life, they inevitably obscure facial expressions, posing significant challenges to emotion detection. Consequently, FER has become more complex, revealing the limitations of existing methods for recognizing emotions on covered faces. This growing challenge has sparked an increase of interest in developing approaches for recognizing emotions on masked faces.

Most existing FER models are trained with datasets of unmasked faces, leading to a notable decline in performance when used for masked faces. As a result, the focus of this study is on recognizing emotions in mask-covered faces. The six fundamental emotions: anger, disgust, fear, happiness, sadness, surprise and the neutral state, serve as the classification categories for this research, were originally identified in 1976 by Paul Ekman [1].

To address these challenges, this study introduces a transfer learning approach based on a teacher-student model paradigm.

The teacher model is trained with an unmasked dataset, while the student model is fine-tuned for recognizing emotions on masked faces. By leveraging the weights of the teacher model, was trained specifically for the emotion classification, the student model inherits the ability to effectively extract features relevant to this task. It

ISSN: 2424-8460(O)
2251-2608(P)

optimizes the training process for the final model, requiring fewer epochs and yielding better results compared to training from scratch.

2. Methodology

The model includes four sequential convolutional layers, followed by an SE block, three Inception blocks, another SE block, and three Residual blocks, ultimately concluding with three fully connected layers. The teacher and student models have identical architectures. The architecture of the model is illustrated in **Figure 1**.

2.1. Transfer Learning

In our research, we utilize transfer learning. First, we train the teacher model by combining three datasets containing faces without masks. The student model, on the other hand, addresses the final task — FER on masked faces. The student model is also trained with a combination of three datasets, but these datasets exclusively consist of images with masked faces. We ensure that the datasets used for the teacher and student models do not contain overlapping images, thereby avoiding repetition and preventing overfitting.

After training the teacher model, the learned weights are transferred to the student model. In the student model, the weights of the first three convolutional layers are frozen, and the remaining layers are fine-tuned using a masked dataset.

2.2. Convolutional layers

In the first block of our model, four convolutional layers are used as feature extractors, with kernel sizes of 7, 5, 3, and 3. Each convolutional layer is followed by batch normalization and the ReLU activation function. Max pooling with a kernel size of 2 is used after the second, third, and fourth convolutional layers. Dropout with a rate of 0.2 is incorporated after the second and fourth layers to mitigate overfitting and improve generalization.

2.3. Inception blocks

Following the convolutional layers, three Inception blocks are employed for efficient feature extraction and representation. Inception architecture first was introduced in [2].

The first Inception block receives an input of 128 channels, while the second and third blocks process 256 channels each.

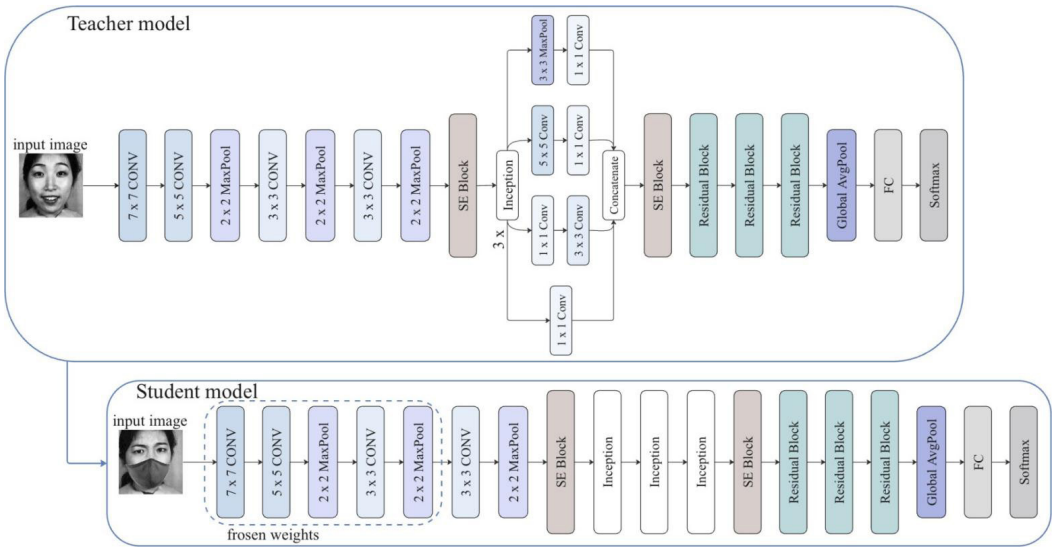

Figure 1. Architecture of the proposed model.

ISSN: 2424-8460(O)
2251-2608(P)

Notably, the second and third Inception blocks do not increase the number of channels. Following all Inception blocks, a dropout with a rate of 0.2 is used. Our block has four parallel convolutional paths. The structure of the block is as follows:

a) The first path: implements a 1 × 1 convolutional layer for dimensionality reduction. This convolution is followed by a batch normalization layer.

b) The second path: begins with a 1 × 1 convolution that reduces the number of input channels. This is followed by a 3 × 3 convolution, enabling the block to capture intermediate spatial features with reduced computational cost. Each convolution is followed by a batch normalization layer.

c) The third path: mirrors the second path but employs a 5 × 5 convolution, allowing the block to capture larger spatial features. Similar to the second path, channel reduction is performed via a 1 × 1 convolution, followed by a 5 × 5 convolution and batch normalization.

d) The fourth path: involves a pooling operation to aggregate global context. A 3 × 3 max pooling layer with a stride of 1 and padding of 1 ensures that the spatial dimensions remain unchanged. After the pooling operation, a 1 × 1 convolution is applied to reduce the dimensionality of the pooled feature maps, followed by batch normalization.

e) Channel dimension: The outputs of the four paths are concatenated along the channel dimension, forming the output of the Inception block.

f) Residal connection: is introduced to facilitate gradient flow and improve convergence. The input tensor is passed through a 1 × 1 convolution to align the number of channels with the concatenated output of the four paths, followed by batch normalization. This residual connection is added to the output.

2.4. Squeeze-and-excitation block

In our model, we utilize a Squeeze-and-Excitation Block twice: first after the convolutional layers and second after the Inception layers, to improve channel-wise feature representation in convolutional neural networks. The Squeeze-and-Excitation block consists of two main operations: squeeze and excitation.

a) Squeeze phase: each spatial feature map is reduced to a single representative value using adaptive average pooling.

b) Excitation phase: inter-channel dependencies are modeled using two fully connected (FC) layers. The first FC layer reduces the channel dimensionality by a factor of r (defined by the reduction parameter) and applies the ReLU activation function. The second FC layer restores the channel dimensionality and applies a sigmoid activation function to normalize the output in the range [0, 1]. The recalibration weights are reshaped to match the dimensions of the input tensor X. The final recalibrated feature maps are obtained through channel-wise multiplication.

2.5. Residual blocks

Before the fully connected layers, we employ three Residual Blocks, originally introduced in [3]. These blocks combine a direct path with a shortcut pathway to facilitate information flow.

Each block performs two sequential convolutional operations, each followed by batch normalization and a ReLU activation function. For an input tensor $X \in R^{N\times C_{in}\times H\times W}$, where N is the batch size, C_{in} the number of input channels, and H, W the spatial dimensions, the first convolutional operation transforms the input into feature representations F_1 using:

$$F_1 = \mathrm{ReLU}(\mathrm{BN}_1(\mathrm{Conv}_{3\times3}(X))) \quad (1)$$

ISSN: 2424-8460(O)
2251-2608(P)

The second convolutional operation further processes the intermediate features to produce F_2:

$$F_2 = BN_2(Conv_{3\times3}(F_1)) \quad (2)$$

To enable the residual connection, a shortcut pathway is introduced, directly connecting the input tensor X to the output tensor, bypassing the main convolutional operations. When the spatial dimensions or the number of channels in X do not match those of F_2 (e.g., due to a stride greater than 1 or differing input and output channels), the input tensor is transformed using a 1×1 convolution.

The transformed input from the shortcut pathway is then added channel-wise to the output of the second convolution, forming the residual connection. Finally, a ReLU activation is applied to the combined output to introduce non-linearity:

$$F_{output} = ReLU(F_2 + X_{shortcut}) \quad (3)$$

The three Res Blocks used in our model process inputs of 256, 512, and 1024 channels, respectively. Following the Res Blocks, a dropout layer with a rate of 0.2 is applied.

2.6. Fully connected layers

At the final stage, three fully connected layers with input channels of 2048, 1024, and 512, were utilized. Each of these layers employs the ReLU activation function, with a dropout rate of 0.5 applied after all layers. The final, fourth fully connected layer has 256 input channels and 7 output channels, corresponding to the number of emotion classes. The softmax activation function is used to this layer.

3. Datasets and experimental details

In this study, three datasets specifically designed for the task of facial emotion recognition were utilized. Examples of the utilized datasets are presented in **Figure 2**.

Figure 2. Sample images illustrating seven emotion classes across three datasets employed.

3.1. Datasets description

a) Japanese Female Facial Expression Database: The first dataset, Japanese Female Facial Expression Database (JAFFE) [4], is widely regarded as a benchmark in FER research. JAFFE has 213 grayscale facial expressions from ten Japanese women, presented in TIFF format with a resolution of 256×256 pixels. The dataset is categorized into seven classes, including six basic emotions and neutral. The distribution of images across the classes is well-balanced, making it suitable for classification tasks.

Although JAFFE contains a relatively small number of images, it was selected for this study due to its extensive use in prior research. Its inclusion facilitates a relevant comparison of results and the evaluation of model performance using a benchmark dataset.

b) Karolinska Directed Emotional Faces: The second dataset used is the Karolinska Directed Emotional Faces (KDEF), introduced in [5]. KDEF has 4,900 images featuring 70 individuals (35 males and 35 females) displaying six basic emotions and neutral. The dataset is perfectly balanced across all classes. The images are

ISSN: 2424-8460(O)
2251-2608(P)

presented in RGB color format, provided as JPEG files, and have a resolution of 562 × 762 pixels.

For the purposes of this study, we selected only 980 images for further analysis. This decision was made because the dataset includes photographs of faces captured from five different angles: frontal, three-quarter left, three-quarter right, profile left, and profile right. When the face is not oriented directly forward, applying a mask during subsequent processing becomes problematic, making such images less suitable for our research objectives.

c) Oulu-CASIA NIR&VIS Facial Expression Database: The third dataset used is the Oulu-CASIA NIR&VIS Facial Expression Database [6]. This dataset contains 2,880 video sequences recorded from 80 subjects, of whom 73.8% are males and the remainder are females. The videos were captured under three different lighting conditions: strong, weak, and dark, with 960 videos for each condition. Each video depicts the transition of a facial expression from a neutral state to one of six basic emotions.

For this study, we used a version of the dataset where each video was split into a sequence of still frames, yielding approximately 20 frames per video. The extracted images are grayscale, saved in JPEG format, and have a resolution of 320 × 240 pixels. From each sequence of images, we selected the second and fourth frames to represent the neutral class and the penultimate and fourth-to-last frames to represent the expressed emotion. It allows us to select two images per emotion from each video: one for the teacher model and another for the student model.

This methodology ensures that the dataset is expanded while preventing the reuse of identical images in the teacher and student models. By extracting neutral images from each video, we increase the overall dataset size. However, this process results in an unbalanced dataset due to the differing proportions of neutral and emotional images.

Table 1. The quantitative distribution of images across masked, unmasked datasets and train, test sets.

Dataset	Total amount of samples	Unmasked dataset		Masked dataset	
		Train set	Test set	Train set	Test set
JAFFE	214	71	35	72	35
KEDF	980	350	140	350	140
Oulu-CASIA	5760	2635	245	2635	245
Merged dataset	-	3056	-	3057	-

3.2. Mask application

All the datasets used originally do not include facial masks. First, unmasked faces are essential for training the teacher model, as it addresses the classical FER task on faces without masks. Second, while there are very few publicly available FER datasets labeled with seven emotion classes that also include facial masks.

Each dataset was split between the teacher model and the student model while maintaining the class distribution. The quantitative distribution of images across the models and the train and test sets is presented in **Table 1** For validation during training process, 15% of the train set was used in both models. During training, the three datasets were merged, testing was performed separately for the test set of each dataset.

The FaceTheMask method, introduced in 2020 year [7], was used for mask application. Using this method, three types of masks: KN95, blue surgical, and black cloth were applied in random order. Masks were applied exclusively to the images intended for the student model. Before mask application, all images were cropped to a resolution of 256 × 256 pixels so that the face occupied the majority of the frame. Initially, faces were detected using the dlib library, provided the coordinates of the detected face.

Figure 3 presents an example of the visual differences between original images and images after cropping

ISSN: 2424-8460(O)
2251-2608(P)

and mask application.

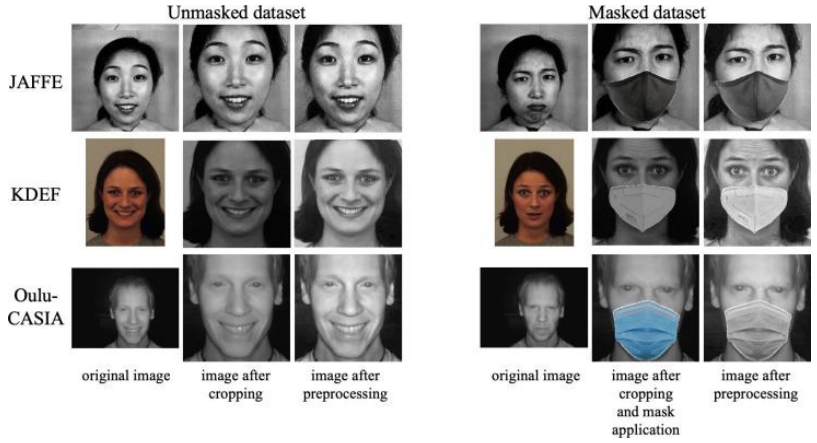

Figure 3. Example of differences between original images and images after cropping, mask application and preprocessing.

3.3. Preprocessing

Since all three datasets are combined during model training, it is essential to perform thorough data preprocessing. The datasets themselves are highly diverse, differing in lighting conditions, brightness, contrast, and image quality. To ensure consistency, these differences must be addressed. Without such preprocessing, the model may become biased toward one dataset, leading to an imbalance in its learning. This would undermine the goal of generalizing features that indicate emotions, as the model would fail to perform equally well across all datasets. Data preprocessing included sharpening and histogram equalization and brightness normalization. Examples of images from each dataset, both prior to and following preprocessing, are depicted in **Figure 3**.

4. Results

This section presents the results of our model. **Table 2** summarizes the metrics for accuracy, precision, recall, and F1 score for both the teacher and student models across all three datasets. Testing was conducted separately on each dataset. The highest performance is achieved when testing on the Oulu-CASIA dataset. This is likely because images from this dataset constitute the largest portion of the final merged dataset used for training the model.

Figure 4 presents the confusion matrices for all three datasets, illustrating the performance of the student model on the task of facial emotion recognition using images of faces with masks. The provided confusion matrices indicate that the model struggles with recognizing disgust on the JAFFE dataset and fear on the KDEF dataset. However, for all other classes, the model demonstrates relatively stable performance.

Table 3 provides a comparison of the achieved accuracy for the facial emotion recognition task on masked face images with other previously proposed methods. The comparison primarily focuses on the JAFFE dataset, as it is the most widely used and accessible dataset for this task. The masked facial FER has not yet been addressed on the KDEF and Oulu-CASIA datasets. Consequently, we are unable to compare our results on these datasets with prior work. This task is relatively specialized, with limited research available. It is also a challenging problem, explains the limited number of methods achieving particularly high performance.

Our results demonstrate that our model on the masked JAFFE dataset achieves higher accuracy that mentioned above articles. This can be attributed to the deeper architecture of our model and the use of transfer learning, where the teacher model's weights are more logically aligned with the task. In contrast, the weights of AlexNet, SqueezeNet, ResNet50, and VGGFace2 were for unrelated tasks and pre-trained with datasets that

ISSN: 2424-8460(O)
2251-2608(P)

do not directly correspond to FER.

Table 2. Masked FER results comparison of proposed model with previously proposed methods on JAFFE dataset.

Method	Accuracy in %
AlexNet [8]	69.92
SquuezeNet [8]	67.19
ResNet [8]	70.70
VGGFace 2 [8]	74.2.2
DCNN [9]	71.35
Proposed model	80.00

Table 3. Results of proposed model across three datasets.

Dataset name	Teacher model (not occluded FER)				Student model (occluded FER)			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
JAFFE	0.91	0.89	0.91	0.90	0.80	0.73	0.80	0.75
KDEF	0.87	0.90	0.87	0.88	0.79	0.82	0.79	0.78
Oulu-CASIA	0.93	0.93	0.93	0.93	0.89	0.89	0.89	0.89

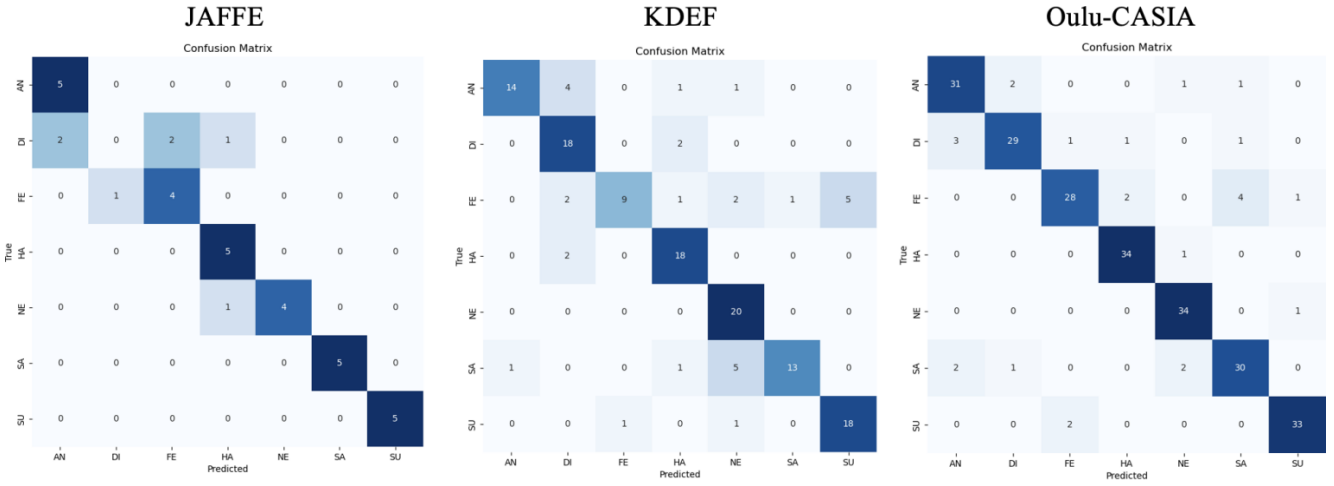

Figure 4. Confusion matrices across three datasets.

5. Conclusion

We propose a novel model architecture for emotion recogni- tion on masked faces. The proposed teacher-student framework effectively adapts the model to handle partially occluded faces. The model was evaluated on three datasets: JAFFE, KDEF, and Oulu-CASIA, achieving accuracy rates of 80.00%, 78.57%, and 89.38%. The analysis reveals that the proposed architecture outperforms previously reported methods on the JAFFE dataset.

References

[1] P. Ekman, Pictures of facial affect, Consulting Psychologists Press, 1976.

[2] C. Szegedy, W. Liu, Y. Jia, et al, Going deeper with convolutions, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 1-9, 2015.

[3] K. He, X. Zhang, S. Ren, et al, Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[4] M. J. Lyons, "Excavating AI"re-excavated: debunking a fallacious ccount of the JAFFE dataset, 2021.

[5] E. Goeleven, R. De, L. Rudi, et al, The Karolinska Directed Emotional Faces: A validation study, Cognition and emotion, vol. 22, pp. 1094-1118, 2008.

ISSN: 2424-8460(O)
2251-2608(P)

[6] G. Zhao, X. Huang, M. Taini, et al, Facialexpression recognition from near-infrared videos, Image and Vision Computing, vol. 29, issue 9, pp. 607-619, 2011.

[7] A. Anwar, A. Raychowdhury, Masked face recognition for secure authentication, 2020.

[8] C. Jiang, M. R. Hasan, T. Gedeon, et al, MaskTheFER: mask-aware facial expression recognition using convolutional neural network, 2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Port Macquarie, Australia, pp. 456-463, 2023.

[9] S. Tegani, T. Abdelmoutia, Using COVID-19 masks dataset to implement deep convolutional neural networks for facial emotion recognition, 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Alkhobar, Saudi Arabia, pp. 1-5, 2021.