Face Emotion Recognition Using Dataset Augmentation Based on Neural Network

Mengyu Rao  
Fuzhou University, Fuzhou, China

Ruyi Bao  
University of Nottingham, Nottingham, United Kingdom

Liangshun Dong∗  
Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Face expression plays a critical role during the daily life, and people cannot live without face emotion. With the development of technology, many methods of facial expression recognition have been proposed. However, from traditional methods to deep learning methods, few of them pay attention to the hybrid data augmentation, which can help improve the robustness of models. Therefore, a method of hybrid data augmentation is highlighted in this paper. The hybrid data augmentation is a method of combining several effective data augmentation. In the experiments, the technique is applied on four basic networks and the results are compared to the baseline models. After applying this technique, the results show that four benchmark models have higher performance than those previously. This approach is simple and robust in terms of data augmentation, which makes it applied in the real world in the future. Besides the results show versatility of the technique as all of our experiments get better results.

CCS CONCEPTS

• Computing methodologies → Artificial intelligence.

KEYWORDS

Deep learning, Computer vision, Facial expression recognition, Facial emotion

ACM Reference Format:

Mengyu Rao, Ruyi Bao, and Liangshun Dong. 2022. Face Emotion Recognition Using Dataset Augmentation Based on Neural Network. In 2022 The 6th International Conference on Graphics and Signal Processing (ICGSP) (ICGSP 2022). July 01–03, 2022, Chiba, Japan. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3561518.3561519

1 INTRODUCTION

Facial expression is one of the most external indications of a person’s feelings and emotions. In daily conversation, according to the psychologist, only 7% and 38% of information is communicated through words and sounds respective, while up to 55% is through facial expression [1]. It plays an important role in coordinating interpersonal relationships.

1 Related Work

As can be seen from Figure 1, ResNet usually contains 5 blocks. There are 5 different network depths in ResNet: 18, 34, 50, 101, and 152. ResNet50 is the most used network depth and many methods or algorithms are based on it. In this paper, ResNet18 and ResNet50 are selected as benchmark models.

2 RELATED WORKS

2.1 VggNet

The VGG model [8] was posted by the Visual Geometry Group team at Oxford University. The primary goal of this architecture is to demonstrate how the its final performance can be impacted by increasing network depth. In VGG, 7×7 convolution kernels are replaced by three 3×3 convolution kernels, and 5×5 convolution kernels are replaced by two 3×3 convolution kernels. The main goal of the change is to make sure that the depth of the network and the impact of the neural network can be ameliorated with the condition of the same perceptual field.

2.2 ResNet

The ResNet [9] model won first place in the ImageNet competition [10] held in 2015. The problem that deepening the model can decrease the accuracy was solved by this work. Due to the proposed residual block, it is easy to learn the identity mapping, even though stacked. If there are numerous blocks, redundant blocks can also learn the identity mapping with the help of the residual block. Furthermore, it improves the effectiveness of SGD optimization, which can optimize the network in deeper. What is more, no additional parameters and computational complexity are introduced. Only a very simple addition operation is performed and the complexity is negligible compared to the convolution operation. The ResNet architecture is shown in Figure 1.

As can be seen from Figure 1, ResNet usually contains 5 blocks.

2.3 Xception

The Xception [12] model is an upgraded version of the InceptionV3 [11] model. Chollet F offers a new structure of deep convolutional...
neural network named Xception that replaces the Inception module with a depthwise separable convolution. The residual network and the depthwise separable convolution are the fundamental components of this network. Xception is typically composed of 36 convolutional layers grouped into 14 blocks, with 12 blocks in the middle containing all linear residual connections. Simultaneously, the model holds the properties of depthwise separable convolution [13] since the model executes spatial layer-by-layer convolution on every channel of the inputs individually, and then conducts point-by-point convolution on the output.

The structure between the ordinary convolution and the above separable convolution of Xception is demonstrated in Figure 2 [12]. The first step is to separate the channels through $1 \times 1$ convolution, and the second step independently draws the spatial correlation of each output channel. Processed individually with $3 \times 3$ and combined are the last step.

3 APPROACH

3.1 HDA: hybrid data augmentation

3.1.1 Horizontal Flip. Geometric transformation is one of the most basic methods to augment data. Because of the particularity of the images, which means that facial expression images emoticons do not undergo a large degree of distortion and rotation in most cases, the horizontal flip (HF) is used to ensure that these images are consistent. Every image in the original dataset is horizontally flipped to create a mirror image. The formula of this method is shown in Eq. 1).

$$\begin{vmatrix} x' \\ y' \end{vmatrix} = \begin{vmatrix} \text{width} - 1 & 0 \\ 0 & 1 \end{vmatrix} \begin{vmatrix} x \\ y \end{vmatrix}$$ (1)

where width equals 48 pixels, x and y represent the position of a pixel in the image.

3.1.2 Gaussian Noise. Gaussian noise (GN) represents a kind of statistical noise that has a probability density function equivalent to that of the normal distribution. In the proposed approach, the training images are added with Gaussian noise to simulate noise that may happen in the real world so that the model can become robust against the original images. The formula of Gaussian noise can be written as Eq. 2).

$$GN(x, y) = \left(2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}\right)^{-1} e^{- \frac{1}{2(1-\rho^2)} \left( \frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} \right)}$$ (2)

3.2 Algorithm

Algorithm 1 Hybrid Data Augmentation
Input: Face emotion dataset D
Output: Hybrid face emotion dataset $D'$
begin:
if size(D) \neq 0 then
for i in all images in D:
\quad x = resize(D_i) // resize to 48*48 pixels
\quad add x to $D'$
\quad add HF(x) to $D'$
\quad add GN(x) to $D'$
end for
endif
end

4 EXPERIMENTS AND RESULTS

4.1 Datasets
Following related work on facial emotion recognition, these experiments are conducted on the two benchmark public face emotion datasets: Ck+ dataset [14], and Fer2013 dataset [15].

Ck+: The most widely utilized laboratory-controlled dataset for facial expression recognition is the Extended Cohn–Kanade (Ck+) Dataset [14] (some samples are shown in Figure 3). Sequences that change from neutral to peak expression are included in the Ck+ dataset. Extraction of the final 1 to 3 frames which have peak formation and the first frame of every sequence is the most common data selection approach for evaluation. Then, people are divided into n groups for person-independent n-fold cross-validation experiments, where n is typically between 5, 8, and 10.

Fer2013: Fer2013 [15] is a large-scale dataset acquired automatically by the Google image search API (some samples are shown in Figure 4). 35887 images are contained in the Fer2013 dataset, and each image is labeled as one of the seven basic emotions. Furthermore, this dataset contains 547 disgusted images, 5121 fear images, 4953 angry images, 8989 happy images, 4002 surprised images, 6077 sad images, and 6198 neutral images.
Some information about the datasets are shown in Table 1.

### 4.2 Experimental settings

The experiments are implemented via PyTorch [16], and the NVIDIA GTX 2080Ti with 4 CPU cores and 13 Gigabytes of RAM is used for experiments. The Adam optimizer [17] is used to train the networks with a learning rate of 2e-4 and betas of 0.9 and 0.999. The best model is selected through the principle of selection of the best accuracy of several experiments.

### 4.3 Experimental evaluation metric

Face emotion recognition can be viewed as a multi-classification problem. Accuracy (Acc) is used as the evaluation metric in this paper, and the calculation formula is as follows:

\[
Acc = \frac{TP + TN}{TP + TN + FP + FN}
\]

In this formula, TP stands for a positive sample predicted by the model as a positive sample, TN stands for a negative sample predicted by the model as a negative sample, FP stands for a negative sample predicted by the model as a positive sample, and FN stands for a positive sample predicted by the model as a negative sample.

### 4.4 Performance on Ck+ dataset

This dataset consists of 8 emotions with a total of 981 trainable images, all of which are 640 pixels × 490 pixels in size. 593 trainable images from 7 expressions are selected for the experiment (shown in Figure 5). However, the background of the volunteers in the pictures is larger than the face images. Therefore, the image size of the dataset is processed to 48×48 pixels, which is convenient for the model input size to be uniform.
The results of the Ck+ dataset and the HDA CK+ dataset are shown in Table 2, where both the Ck+ dataset and the HDA CK+ dataset are divided in the ratio of 8:1:1 for training, validating, and testing respectively. The batch size used is 32, and 20 epochs are used for training.

Tables 2 and Figure 6 indicate the accuracies of the Ck+ dataset and those of the HDA CK+ dataset. The dataset with data augmentation always has a higher performance in most models, and the ResNet18 model achieved 100% testing accuracy in the HDA CK+ dataset. After analyzing the results, significant improvement in accuracy can be seen on all models, especially for Vgg19. The accuracy comparison of Vgg19 shows that the accuracy improves by 23.61%. The ResNet50 network is the least improved, but the accuracy is also 4.03% higher than the original data.

Furthermore, the figures show that the data augmentation improves the feature learning for every model. For the left one of each column, the Ck+ dataset is used. The training accuracies have an upward tendency, but the validation accuracies always have some shocks, even increasing slowly. Especially for ResNet18, the curve always fluctuates around 70%. However, when comes to the right one of each column, the HDA CK+ dataset is used. Both the curve of training accuracy and the curve of validation have an increasing toward, and the validation accuracies of all figures in the right position come to a peak of around 98%.

### 4.5 Performance on Fer2013 dataset

After rejecting wrongly labeled frames, each image was registered and scaled to 48×48 pixels. There are 28,709, 3,589 and 3,589 images respectively for training, validation, and testing with seven expression labels in the Fer2013 dataset (shown in Figure 7).

The results of the Fer2013 dataset and the HDA Fer2013 dataset are presented in Table 3, where both the Fer2013 dataset and the HDA Fer2013 dataset are divided in the ratio of 8:1:1 for training, validating, and testing respectively. The batch size used is 32, and 20 epochs are used for training.

Tables 3 and Figure 8 indicate the accuracies of the Fer2013 dataset and those of the HDA Fer2013 dataset. The dataset used data augmentation always has a higher performance in most models. The results show that improvement in accuracy can be seen in all models. The accuracy comparison of ResNet50 shows that the accuracy improves by 25.62%. The Xception network is the least improved, but the accuracy is also 22.33% higher than that on the original data.

Furthermore, the figures show that the data augmentation improves the feature learning for every model. For the left one of each column, the Fer2013 dataset is used. The training accuracies have an upward tendency, but the validation accuracies always fluctuate around 63%. However, when comes to the right one of each column, the HDA Fer2013 dataset is used. Both the curve of training accuracy and the curve of validation have an increasing toward, and the validation accuracies of all figures in the right position come to a peak above 82%.

### 5 CONCLUSION

In this study, a hybrid data augmentation method of the dataset, which has a high performance in some models, is presented. Because convolutional neural networks require more samples for training to get accurate and robust result, the hybrid data augmentation method is used to enlarge the number of samples. After applying the technique, the numbers of images in both the Ck+ dataset and the Fer2013 dataset have increased, and four benchmark models have higher performance than those previously. This approach is simple and robust in terms of data augmentation, which makes it applicable in the real world in the future.

### REFERENCES

[1] A. Mehrabian and S. Ferris, "Inference of attitudes from nonverbal communication in two channels." Journal of consulting psychology, vol. 31 3, pp. 248–52, 1967.
Figure 6: Accuracy comparison of different models on Ck+ dataset and HDACK+ dataset. (Best viewed in color)

Figure 7: Images of different expressions in the Fer2013 dataset
Figure 8: Accuracy comparison of different models on Fer2013 dataset and HDAFer2013 dataset. (Best viewed in color)

[2] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," Journal of personality and social psychology, vol. 17 2, pp. 124–9, 1971.