Influence of Skin Color and Data Pre-processing on Gender Classification

Zihang Shao*

School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai, Shandong, China

* Corresponding Author Email: 202000800151@mail.sdu.edu.cn

Abstract. Face recognition and gender classification is one of the hot spots at present, how to get better results in the recognition process is the key. This research was based on neural networks. While some researchers have explored this issue through neural networks, this research focuses on the impact of data processing other than the model on the final results. First of all, through the manual review and screening of data sets, the training set of face images only retained deep color skin face images, and the test set did not do selection processing. The final results of the model which based on selected data set are not much different from that of the model based on the unselected training set, except for a slight decrease in the recall rate of female face recognition. Secondly, by the method of control variables, the effect of the three kinds of data pre-processing is compared. The results show that both random cropping and normalization can improve some attributes of the model to varying degrees, but only random horizontal flipping has a certain negative effect on the model. In this research, it was found that the smaller proportion of white skin face image training sets hardly affected the model performance. Data pre-processing can effectively improve the model, but the use of random horizontal flips in a data set with a great number of small tilt angle images may lead to the deterioration of the results on the training set.

Keywords: Gender classification; deep learning; data pre-processing; face recognition.

1. Introduction

In today's society, the use of technology to identify or authenticate identity and obtain information has become an indispensable part of life. Various ways to obtain personal identity characteristics are widely used, such as fingerprint, iris, face recognition, and so on. Due to the outbreak of COVID-19, the risk of infection increases when people contact with public things. Due to its advantages of non-contact, the application scope of facial recognition has been further expanded, and it has become a more mainstream identification technology. In the information age, the recommendation of corresponding services according to different information of different users has become the key to improve user experience. Facial gender information plays an important role in providing personalized service and personalized human-computer interaction experiences for users [1].

As a part of machine learning, deep learning is an algorithm based on neural network block to learn from data [2, 3]. Deep learning has better anti-noise ability in image classification [4, 5]. Compared with traditional methods, deep learning does not require the manual design of features and requires less prior knowledge of designers. Meanwhile, based on sufficient data, deep learning models can have better performance [6, 7].

Generally speaking, increasing the quantity of layers of the model can help the model learn features better [8]. However, when the number of layers has risen up to a certain extent, the "model degradation" problem that results become worse will occur. The ResNet [4] can well avoid this problem.

Based on the ResNet18 model, this article explores the impact of data pre-processing on the final results of the model, including the impact on the accuracy rate, precision rate, recall rate, and other model measurement indicators. Meanwhile, the impact of data pre-processing on the training speed of the model is also concerned [9]. On this basis, this article also explores the influence of the proportional distribution of different races in the training set on the classification results. After manually selecting the training set containing the ideal proportion of races, this training set is applied

to train the model. Using the same model, train it with an original training set. Under the condition that those two training sets have same data size, the indicators of the model are compared to explore the results.

2. Methodology

2.1. Data collection

In this research, the dataset is gender classification dataset [10]. The dataset is images of male and female. It is divided into two parts, training sets and test sets. The training set contains about 23,000 images for each class, and the test set contains about 5,800 images for each class, as shown in Table 1. In the dataset, the sample is distributed across all skin colors and age groups. In the dataset, most of the images were taken from the front side and no angel tilt, but a few of the images had a random degree of angle tilt.

Table 1. Initial data.

Samples	Male samples	Female samples	Total samples size
Training samples	23766	23339	47105
Testing samples	5808	5841	11649
Total samples size	29574	29180	28754

2.2. Data screening

The images of this data set have been pre-picked and are all usable data. In subsequent experiments, in order to investigate whether the proportion of different skin color sample will influence on the model. Table 2 demonstrates the data distribution after selection. By manually sifting through the data, only people of color were retained in the training set, in order to reduce the workload, researchers reduced the amount of data, 1272 images of people of color were selected, with a 1:1 ratio of male to female images. For the test set, 912 samples were randomly picked from the original test set, among which 456 were male and female. Considering the influence of data size on model training, aiming to control variables, the control group was set up in this study. 1272 images were randomly picked from the original training set as the training set of the control group, including 636 images of men and same quantity images of women, which were consistent with the experimental group, and the influence of other factors on the results was minimized. The test set in the control group was identical to the test set in the experimental group. The above data is only used to explore the influence of skin color. When exploring the influence of data pre-processing on the model, the original large-scale data is still used in the research.

Table 2. Data after selection.

Samples	Male samples	Female samples	Total samples
Training samples	636	636	1272
Testing samples	456	456	912
Total samples size	1092	1092	2184

2.3. Data processing

In order to increase the robustness of the model, random cropping and random horizontal flipping of the pictures were used before the training began. Random cropping of images can effectively reduce the negative effect of data noise on the model and increase the model stability. In the training set, although there are some non-positive face images, but in training neural network will affect greatly weakened, in order to enhance model to other the non-positive face images recognition ability, enhance the adaptability of the model, a random horizontal flip was applied to the training set. At the same time, in order to accelerate the training speed of the model, the normalization is used to the data before the training starts. At the same time in order to test the above three kinds of data pre-processing

effects on the actual model, this research also set up a control group, in the training data and test data are not changed and under the condition of using the same model, set up four kinds of data pre-processing methods combinations, respectively is: no data pre-processing, random cropping only, random horizontal flipping only and normalized only. The first three groups of comparison mainly focused on the test results, and the last group also paid attention to the change of model training time in addition to the test results.

2.4. Model architecture

The network model used in this study is resnet18, and the residual network is composed of a series of residual blocks. A residuals block can be represented by the following figure, with the input convolved multiple times and then added to the input.

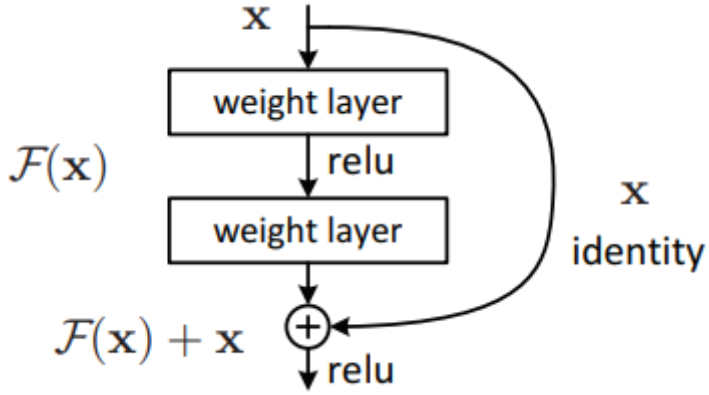

Fig. 1 Residual learning: a building block. Figure from [4].

Resnet18 is a relatively mature network model that can effectively solve the degradation problems of deep networks. Its basic module is demonstrated in Figure 1. The 18 layers of ResNet18 means the 18 layers with weights. All of these layers are convolution layers and the full connection layers. Pooling layer and BN layer are not included. ResNet18 is a relatively deep network, which is conducive to better learning effect of the model.

2.5. Evaluation indicators

The model evaluation indicators are as follows. (1) **Precision and Recall**. The accuracy rate represents the proportion that the model predicts correctly among all the results predicted by the model as positive. The recall rate represents the proportion of all outcomes that are actually positive that the model predicts is correct. (2) **Specificity**. Specificity is expressed in all negative classes of the model. Predict the correct ratio column. (3) **F1 value**. This is a value that ranges from 0 to 1 and can be viewed as a weighted average of model accuracy and recall rates. (4) **Geometric mean (G-mean)**, which can measure the average performance of the model in two categories. (5) **Confusion matrix**. The sample number of predicted and actual categories can be visualized, and the four data items are respectively.

TP: The probability that a positive class is predicted to be positive.
TN: The probability that a negative class is predicted to be negative.
FP: The probability that a negative class is predicted to be positive.
FN: The probability that a positive class is predicted to be negative.

3. Result

3.1. Result Hypothesis

First, as for the proportion of people with different skin colors in the training set, the input values generated by pictures with different skin colors are different. Therefore, if the proportion of people with different skin colors changes in the training set, it may have a negative impact on the test results.

Secondly, for the three data pre-processing methods, if no data pre-processing is applied on the data set, the model may not be able to correctly identify the test samples from non-positive faces or other angles, so the test results would be affected in a negative way. Random horizontal flipping and random cropping might both increase the adaptability of the model to some extent, making the model perform well in the face of more complex image input. Normalization may play an active role in shortening the time of network training and quickly searching for the optimal solution.

3.2. Result of different skin color

Table 3. The test results of different skin color samples in the training set

	Before data set selection	After data set selection
Accuracy	0.9539	0.9506
Precision	0.8427	0.8593
Recall	0.8393	0.8367
Specificity	0.8434	0.8609
F1-score	0.8411	0.8479
Gmean	0.8411	0.8480
Running time	178s	168s

In Table 3, the data on the left is the test result without manual selection of the data set, and the data on the right is the test result after selection of the data set. There was no significant difference in the results of the two tests. The confusion matrixes are shown in Figure 2 and Figure 3.

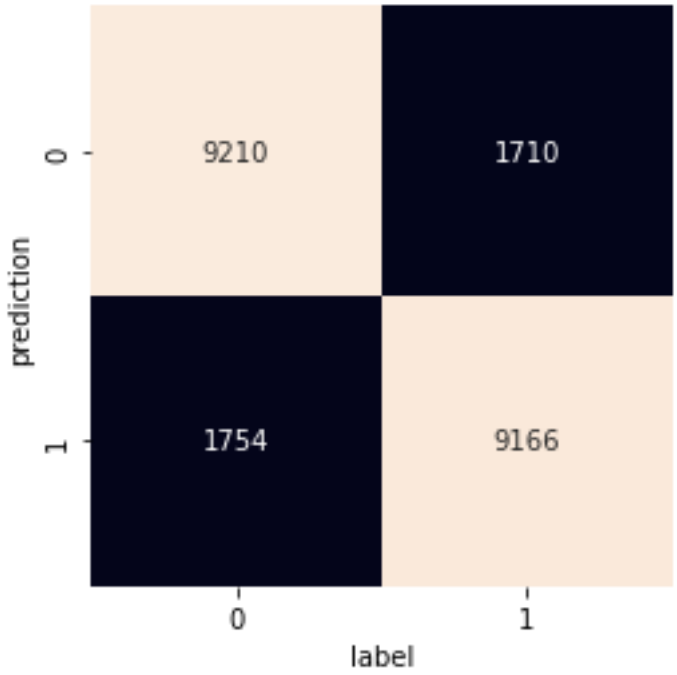

Fig. 2 Confusion matrix before selection.

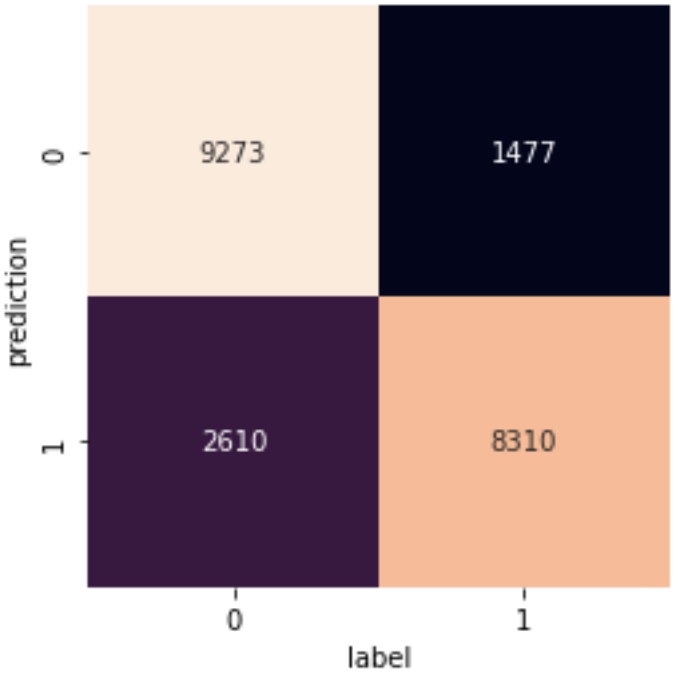

Fig. 3 Confusion matrix after selection.

Where 1 represents male and 0 represents female. The results showed that the model trained on the selected data set had a slightly negative effect on female recognition compared to the pre-selected test set.

3.3. Data pre-procession results

3.3.1 Result comparison before and after pre-procession

In Table 4, the left column shows the information of the model trained by the training set without any data pre-processing, and the right column shows the information of the model trained by the training set after random cropping, random horizontal flip and normalization.

Table 4. Results of models with and without data pre-processing

	Without any pre-processing	After complete pre-processing
Accuracy	0.6505	0.9751
Precision	0.9191	0.9404
Recall	0.8759	0.9379
Specificity	0.9218	0.9397
F1-score	0.8969	0.9392
Gmean	0.8972	0.9392
Running time	3612s	5692s

In any index except training time, the results after data pre-processing are better than the results without data processing, and the test accuracy increased by 32%.

3.3.2 Result of random cropping

In Table 5, the data on the left is the result of using random cropping on data set, the data in the middle column is the result of the model trained by the training set after random shearing, random horizontal flipping and normalization, and the data on the right column is the result without any processing.

Table 5. Results of models only with random cropping

	Cropping Only	After complete pre-processing	Without any pre-processing
Accuracy	0.8086	0.9751	0.6505
Precision	0.9253	0.9404	0.9191
Recall	0.8466	0.9379	0.8759
Specificity	0.9307	0.9397	0.9218
F1-score	0.8842	0.9392	0.8969
Gmean	0.8851	0.9392	0.8972
Running time	5697s	5692s	3612s

There is still a big gap in the accuracy, recall rate, F1-score and Gmean indexes, but the training time is almost the same. Compared with the results without any data processing, it is found that the accuracy is greatly improved, but the running time is longer.

3.3.3 Result of random horizontal flipping

In Table 6, the data on the left is the result after random flipping of the data set, the data in the middle column is the result of the model trained by the training set after random shearing, random horizontal flipping and normalization, and the data on the right column is the result without any processing.

Table 6. Results of models only with random horizontal flipping

	Flipping Only	After complete pre-processing	Without any pre-processing
Accuracy	0.6357	0.9751	0.6505
Precision	0.8586	0.9404	0.9191
Recall	0.9352	0.9379	0.8759
Specificity	0.8439	0.9397	0.9218
F1-score	0.8953	0.9392	0.8969
Gmean	0.8961	0.9392	0.8972
Running time	4008s	5692s	3612s

Except for the recall rate of the data on the left, which is not much different from the data after processing, the other indicators are inferior to the data results after processing. In terms of running time, the time required by the model on the left is significantly less than the time required by the model trained based on the data after processing. Compared with the unprocessed data, the results of the random horizontal flip only have better recall rate than the latter.

3.3.4 Result of normalization

In Table 7, the data on the left is the result of normalization was applied on the data set, the data in the middle column is the result of the model trained by the training set after random shearing, random horizontal flipping and normalization, and the data on the right column is the result without any processing.

Table 7. Results of models only with normalization

	Normalization Only	After complete pre-processing	Without any pre-processing
Accuracy	0.6617	0.9751	0.6505
Precision	0.8882	0.9404	0.9191
Recall	0.9058	0.9379	0.8759
Specificity	0.8845	0.9397	0.9218
F1-score	0.8970	0.9392	0.8969
Gmean	0.8970	0.9392	0.8972
Running time	3413s	5692s	3612s

As can be seen from the chart, compared with the data without any pre-processing, the data after normalization only takes less time in the running time, and the other indicators have little difference, and these indicators are far smaller than the results after complete data pre-processing.

4. Discussion

First of all, after the white skin samples in the training set are manually removed, the model which was trained through the training set after the selection is slightly negatively affected in the female recognition compared with the training set before the selection. This is similar to the results obtained by Joy Buolamwini, whose research found that dark-skinned women were the most misjudged category. The results obtained after random cropping are indeed improved, because the adaptability of the model is increased through random cropping, which makes the model perform better in the face which is not complete front side face images. Normalization also resulted in a slight improvement in training speed.

However, different from the hypothesis, the overall evaluation index of the model did not change significantly after manually removing the white skin samples in the training set. This may be because after the training, the judging criteria of the model are no longer based on a single color, but on the softness or sharpness of the facial outline and the facial organs distribution features. These features, however, are independent of skin color and therefore have less impact on the overall model. In the severely affected models mentioned in Joy's article, whites and males accounted for more than 70% of the training set, while in this research, both males and females accounted for 50% of the samples, and the training set is all dark-skinned race samples, which may be the reason for the different research results. Therefore, based on the above results, the researchers can assume that too many samples of white skin in the training set will affect the classification results, while too many people with deep color do not show such defects for the time being. Therefore, the proportion of pictures of white people may be appropriately reduced in the research to improve the model.

In addition, in the part of data pre-processing, the research found that only random horizontal flipping of data did not improve the model much, but increased the running time. This is because the inclination angle of most samples in the test set is very small, and the addition of random horizontal flipping has a certain negative impact on the model. However, in practical application, researchers should consider whether samples from other angles will appear, so as to decide whether to carry out random horizontal flip processing.

Due to limited resources, the samples obtained in this research are not big enough, especially in the research with different skin color samples taking up different proportions in the training set, which may lead to unrepresentative results to some extent. In addition, because the samples are manually screened, there is a possibility of error.

5. Conclusion

In this research, it is found that adjusting the proportion of different skin colors in the training set will affect the gender classification model to some extent, but the effect is light. In this research, only the colored skin samples were retained as the training set, and the results achieved on the selected training set were almost the same as the result which based on the data set which without the selection. The researchers speculate that the model may be classifying features that are not affected by skin color but only by sex, such as the softness of facial contours or the distribution of facial organs. In addition, the negative impact of too few white samples in the training set on the model is much smaller than the negative impact of too many white samples.

In terms of data pre-processing, data pre-processing can greatly enhance the results of the model, but it may increase the training time. At the same time, the characteristics of the test set and the training set should be fully considered. If there are no samples from other angles in the data, the data flipping may lead to negative effects.

Gender classification model plays an important role in information acquisition, electronic commerce and other fields. Improving the model by pre-processing the data set or adjusting the distribution of different kinds of samples in the data can increase the efficiency.

Due to the limit of resources, this research did not further research the reasons for the different proportions of samples with different skin colors cause different results. There is no research about the impact on the results when yellow skin, black skin, white skin dominates the training set in this study. At the same time, the lack of data volume also affects the representativeness of the results, which can be improved in the subsequent work.

References

[1] Kumar S, Singh S, Kumar J. Gender classification using machine learning with multi-feature method. IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). 2019: 0648-0653.

[2] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798-1828.

[3] Zhang D, Yin J, Zhu X, et al. Network representation learning: A survey. IEEE transactions on Big Data, 2018, 6(1): 3-28.

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

[5] Li S, Song W, Fang L, et al. Deep learning for hyperspectral image classification: An overview. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(9): 6690-6709.

[6] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444.

[7] Guo Y, Liu Y, Oerlemans A, et al. Deep learning for visual understanding: A review. Neurocomputing, 2016, 187: 27-48.

[8] Srivastava R K, Greff K, Schmidhuber J. Training very deep networks. Advances in neural information processing systems, 2015, 28.

[9] Too E C, Yujian L, Njuki S, et al. A comparative study of fine-tuning deep learning models for plant disease identification. Computers and Electronics in Agriculture, 2019, 161: 272-279.

[10] Ashutosh C, Gender Classification Dataset, URL: https://www.kaggle.com/datasets/cashutosh/gender-classification-dataset. 2019.