An ensemble approach for art face recognition problem

Cunzheng Wang*

School of Mechanical Engineering and Automation, Beihang University, Beijing, 100191, China

*Corresponding author’s e-mail: 18376028@buaa.edu.cn

Abstract. In recent years, with the development of the digitalization of cultural relics, many cultural relic restoration models based on different calculation methods have been proposed. Although the models for feature classification (such as VGG16, Xception) or ensemble models (such as SVM and ANN) are relatively mature, multiple classification problems and accuracy problems are still far from ideal. This paper proposes an ensemble method for artistic face recognition based on multiple models. Six convolutional neural network models are used as basic models to compare their performance in gender and status classification. Based on the KaoKore data set, after training them to obtain the corresponding characteristics, the ensemble model is used for weight combination to obtain a more accurate model. It has achieved 98.10% and 90.13% accuracy on gender and status classification respectively, which is the best in the current literature.

1. Introduction
Deep learning, especially Convolutional Neural Networks (CNN), as the research focus in pattern recognition in recent years, has attracted more and more attention, and related references are also emerging in an endless stream. The earliest CNN can be traced back to the proposal of the BP algorithm in 1986, and then LeCun used it in a multilayer neural network in 1989, until LeCun proposed the LeNet-5 model in 1998, and the prototype of the neural network was completed. In the next ten years, the related research of convolutional neural network tended to stagnate for two reasons. First, researchers realized that the calculation of multi-layer neural network during BP training was extremely large. Hardware computing power is completely impossible to achieve. Secondly, shallow machine learning algorithms including SVM have gradually begun to show their advantages. In 2006, Hinton finally made a breakthrough and published an article in Science. CNN once again awakened and made great progress. In 2012, CNN won the ImageNet competition. In 2014, Google developed a 20-layer VGG model. In the same year, the DeepFace and DeepID models were born, directly brushing the correct rate of face recognition and face authentication on the LFW database to 99.75%, almost surpassing humans.

The problem to be dealt with in this paper is the classification of gender and identity of characters related to artworks. Although there are many researches related to face recognition, there are relatively few researches on face classification in artworks. Because it can improve the efficiency of cultural relic restoration and reduce the secondary damage of cultural relics. At the same time, with the continuous excavation of cultural relics, the public are paying more and more attention to the protection of cultural relics, and the supply of professionals who can deal with them is insufficient. In recent years, computer-aided cultural relics restoration has received extensive attention from researchers. Identifying the identity of people in cultural relics and artworks can help archaeologists identify the year of the work, restore ancient works, and improve efficiency. On the other hand, it can also set prices for works.
In this paper, we used a Japanese artwork face data set named KaoKore [1], which contains 5,823 images for training. We tried the ensemble of different deep learning models to improve the classification accuracy of characters' gender and identity. Compared with a single model, we achieved better results. This paper further explored the application of artificial intelligence technology in archaeology. An ensemble model with higher accuracy is proposed to make the task of cultural relic image recognition more accurate and efficient. In the era of the explosion of cultural relics information, the burden on cultural relics restoration personnel has been reduced.

2. Related Work

2.1. Face Recognition

Face recognition has been heavily studied in the literature. Both machine learning and deep learning models have been applied for this problem. Because of the success of convolutional neural networks in the ImageNet competition, deep learning has become dominant in a series of problems, including face recognition [2-7], forecasting problems [8-10], classification problems [11-13], etc.

The combination of KPCA and SVM has been used for gender classification, and the accuracies for the AT@T, Faces94, Georgia Tech Face Database are 97.375%, 99.7%, and 96.67%, respectively [2]. VGG-Face is used and fine-tuned for gender classification in [3]. The modified model is trained and tested on open datasets including Adience, Wikipedia and LFW. The result for LFW is as high as 98.45%, which is higher than the traditional models, but the result for Adience is unsatisfactory, which is only 87.08%.

A novel multi-task deep architecture is proposed in [4], for facial gender and smile classification problems, based on two CNNs, namely, GNet and SNet. The test on the FotW shows that the new architecture can achieve an accuracy of 91.32% for the gender classification and 89.34% for the smile classification problem.

A fully connected CNN structure is proposed in [5], and adapted for both gender and age classification. The adapted network can achieve a competitive performance with notably fewer training epochs and fewer network variables. Finally, the gender classification error can be controlled within 2%, but to achieve a similar result for the age classification error, more training time and a more complex model are still needed.

Three new KNN classifiers, namely, KNN-Distance, Modified-KNN and KNN-Regression are proposed in [6]. They are tested on FG-NET Database with Mean Absolute Error (MAE) as the evaluation metric. The result shows that the gender estimation is improved and the age estimation is also improved if the gender is known. However, the Modified-KNN shows a decreasing accuracy.

There are also some models proposed for 3D face recognition. The SVM classifier in [7] is used for the 3D face geometric features. However, the performance is not satisfactory. The multiple training on the GavabDB dataset shows a final error ratio above 15%.

2.2. Applications of Artificial Intelligence in Historical Artwork

In fact, the idea of using artificial intelligence technology to restore and protect historical relics has been proposed very early. For example, in the 1990s, Jinshi Fan, the former president of the Dunhuang Academy, proposed the concept of "Digital Dunhuang", which is to use computer technology to permanently preserve the murals and paintings of the Mogao Grottoes in Dunhuang [14]. At the same time, Microsoft Research Asia uses the precious data provided by the Dunhuang Research Institute, combined with its own technical advantages, to provide an interactive information exchange platform for Dunhuang Mogao Grotto lovers, and help more people understand the history and culture of Dunhuang Mogao Grottoes in depth.

More research exists in this area. A three-dimensional convolutional neural network model based on residual learning is proposed in [15] and applied to the field of cultural relic point cloud type recognition. It optimizes and improves the three-dimensional convolutional neural network and introduces residual learning, which effectively avoids the degradation of deep-layer three-dimensional convolutional neural
networks, and improves the accuracy of object point cloud type recognition to a certain extent. This three-dimensional residual neural network is applied to the point cloud data of terracotta warriors and horses fragments to identify the types of terracotta warrior fragments. The experimental results show that the recognition accuracy rate reaches 83.59%, which is aimed at the complex cultural relics of terracotta warriors and horses. Fragment recognition and matching and splicing are the key technologies. In addition, in the recognition of fragment parts, deep learning has a better recognition effect, but the recognition accuracy of shallow networks is low, and deep networks will have a certain degree of network degradation. In fragment matching and splicing, for complex geometric contours, it is often difficult to achieve better splicing results based on a single feature.

Single-label classification and multi-label classification algorithms are also widely used in the field of cultural relics restoration [16]. Aiming at the current status of large-scale, multi-category collections of cultural relics image data sets that are not publicly available, they are constructed separately for domestic and foreign collection types through the Internet. Two representative data sets, DPM data set and MET data set, are used for single-label and multi-label classification research respectively. Aiming at the small sample problem of DPM data set, it first classifies the DPM data set against mainstream deep learning models by means of deep transfer learning, and the ResNet50 model reaches nearly 87% accuracy. Aiming at the problem of large differences within the categories of cultural relics and small differences between categories, a multi-feature fusion classification method combining point convolution and ensemble learning is proposed. The method based on local connection point convolution will finally classify the accuracy on the DPM data set, which achieves an increase of nearly 5 percentage points. Aiming at the characteristics of cultural relics images, which are mostly single-object images, a multi-label classification neural network based on RNN iterative prediction is proposed from the perspective of using label correlation. The experimental results on the MET dataset show that the introduction of RNN can effectively improve F1 score, accuracy, Hamming loss and Ranking loss, which are four multi-label classification evaluation metrics.

3. Dataset
KaoKore [1] we use in this study is a relatively novel data set, which consists of faces extracted from Pre-modern Japanese artworks. The KaoKore dataset is constructed based on the facial expression collection, which comes from the ROIS-DS Humanities Open Data Center (CODH) and has been publicly available since 2018. It provides a dataset of cropped facial images of Japanese art publicly extracted from the National Institute of Literature of Japan, the Digital Archives of Rare Materials of Kyoto University, and the Media Center of Keio University from the late 16th century to the early 17th century, in order to promote the study of art history, especially the study of artistic style. It also provides corresponding metadata annotated by researchers with domain knowledge, e.g. gender, social class, etc.

The KaoKore dataset used in this study contains 5823 image files, each of which is a colorful RGB image with a resolution of 256×256. At the same time, each file also has two sets of labels, namely, gender and social status.

For social status, we choose noble, warrior, incarnation, commoner, each of which is associated with at least 600 images, and abandon the rare ones, such as priest and animal, each with only a dozen images. This is to avoid unbalanced distribution on the label.

4. Models

4.1. Base Model
In this study, we use six convolutional neural network models as the base models and compare their performance for both the gender and status classifications.

4.1.1. DenseNet201 [17]
The model is a brand new connection structure proposed by Cornell University, Tsinghua University, and Facebook FAIR. In order to maximize the information flow between all layers in the network, the
authors connect all layers in the network in pairs so that each layer in the network receives the features of all the layers before it as input. Because there are a large number of dense connections in the network, the authors name this network structure as DenseNet. It mainly has the following two characteristics. To a certain extent, it alleviates the problem of gradient vanishing problem during training. Each layer will receive the gradient signals of all subsequent layers during the back propagation, so as the network depth increases, the gradient near the input layer will become smaller and smaller.

4.1.2. InceptionResNetV2 [18]
This model was announced by the Google team in February 2015, when batch normalization appeared. It is to solve the problem of Internal Covariate Shift problem (i.e., internal neuron data distribution changes). The Google team proposed for the first time that the $5 \times 5$ convolution kernel can be replaced by two $3 \times 3$ convolution kernels and applied to V2. In general, it has the following advantages: 1) A higher learning rate can be used; 2) Remove or use a lower dropout; 3) Reduce the L2 weight attenuation coefficient; 4) Cancel the Local Response Normalization layer; 5) Reduce the use of image distortion.

4.1.3. ResNet152V2 [19]
ResNet (Residual Neural Network) was proposed in a 152-layer neural network, which won the championship in the ILSVRC2015 competition. The key idea of ResNet is to add a direct connection channel to the network, which retains a certain proportion of the output of the previous network layers. That is, the neural network of this layer does not need to learn the entire output, but learns the residual of the previous network output, so ResNet is also called the residual network.

4.1.4. VGG16 [20]
VGG16 is a very classic model and is the runner-up of 2014 ImageNet. The core idea is small core and stacking, which are mainly divided into 5 stages and 13 convolutional layers. The number of 16 means adding 3 FC layers. Each stage is followed by a pooling operation to reduce the size. In terms of parameters, fully connection accounts for a lot, so most of the latter are replaced by convolution or global pooling. When doing segmentation, the stride of the fourth pooling is generally set to 1, and then the convolution of stage 5 uses dilation convolution to increase the size. The receptive field of 192 is still relatively small, and according to the analysis of some papers, the actual receptive field is generally much smaller than the theoretical one. Therefore, if it is used for tasks such as detection, the receptive field may be a problem.

4.1.5. InceptionV3 [21]
This model was announced by the Google team in December 2015. V2 has two 3x3 convolution kernels instead of a 5x5, while V3 will decompose more thoroughly. The core idea of V3 is to first use two 3x3 convolution kernels instead of 5x5 convolution kernels, and three 3x3 convolution kernels instead of 7x7 convolution kernels to reduce the amount of parameters and speed up calculations. V3 further decomposes the nxn convolution kernel into 1xn and nx1 convolution kernels, while reducing the size of the feature map and increasing the number of channels.

4.1.6. Xception [22]
Xception uses Separable Convolutiono replace the convolution operation in Inception-v3. In Inception, features can be extracted through 1ith separately. Then they use Separable Convolution (the ultimate Inception module) toe leaves the selection of feature types to the network's own training, that is, one input is input at the same time. Give several ways to extract features, and then concat. Compared with Inception V3, in terms of classification performance, Xception has a smaller lead on ImageNet, but has a much lead on JFT. In terms of parameters and speed, Xception has fewer parameters than Inception, but is faster.
4.2. Ensemble Model
We also propose an ensemble approach for the classification problems, based on the predicted probabilities from the base models and the ensemble module. The structure is shown in Figure 1.

![Figure 1. The structure of the ensemble model](image)

We use the following models for the ensemble module:

4.2.1. AdaBoost
The full name of Adaboost is adaptive boosting, which is derived from the improvement of the Boosting algorithm by Freund and Schapire in 1995. The algorithm process is to first train a model as a weak learner, and then evaluate the model. The problem of doing the right in this model will reduce the attention, and the problem of doing it wrong will increase the attention. The latter model of this algorithm is always completed on the basis of the previous model, which belongs to the sequential cascade structure. The advantage is that it can fully consider the weight of each sample point, is good at solving difficult problems, and the model performance ceiling is high. But it is easy to overfit, too sensitive to outliers of abnormal points, low starting point of model performance and slow speed.

4.2.2. k-nearst neighbors (kNN)
The core idea of k-nearst neighbors (kNN) is very simple and intuitive: if most of the K most similar samples of a sample in the feature space belong to a certain category, the sample also belongs to this category. Although this algorithm only needs the local distribution and does not need the overall distribution, the local estimation cannot conform to the global distribution, so the probability cannot be calculated.

4.2.3. Random Forest (RF)
Random forest was first proposed by Leo Breiman and Adele Cutler. This model process is to randomly sample from the data set to train each decision tree in the model (the model contains multiple decision trees), and then preset the hyperparameters of the model (regression: average; classification: mode) to make decisions.

This model has strong randomness and noise resistance, is not easy to overfit, and is sensitive to outliers of abnormal points. Processing high-dimensional data is also relatively fast. At the same time, due to the tree structure, the model has a high degree of interpretability and can tell the importance of each feature. But the disadvantage is that the model is too general, does not have the ability to deal with difficult samples, and the starting point of the model is high and the ceiling is low.

4.2.4. Support Vector Machine (SVM)
Support Vector Machine refers to a series of machine learning methods, originally developed by Vapnik of AT&T and colleagues in the 1990s. When this algorithm is used for classification, it is what we call
SVC (Support Vector Classification). The SVM classifier can be binary or discriminant. The advantage is that it is a strong classifier that can maximize the distinction between two data sets.

4.2.5. XGBoost [23]
The algorithm was developed by Dr. Tianqi Chen from the University of Washington and open sourced for everyone to use. Its core steps are as follows: (1) constantly performing feature splitting and adding trees; (2) predicting a sample score; (3) adding up the separate scores.

4.2.6. Deep Neural Network (DNN)
The "depth" in deep learning refers to the number of layers in a neural network. The good property of this system is a set of simple mathematical functions that can be trained. DNNs are suitable for many machine learning tasks.

5. Experiments

5.1. Settings
The models are implemented with Python, using scikit-learn and TensorFlow packages. All the experiments are conducted with a Windows OS personal computer.

We use the accuracy and confusion matrix for evaluation metrics. The accuracy here refers to the degree of conformity between the results produced by testing the entire data set and the real results after the training of each integrated model is finally completed. As an example, if a data set of character gender assumes a total of 10 characters, the actual situation is 8 females and 2 males; using the trained model to test and get 7 females, then its accuracy is 7/8.

5.2. Results for base models
For the base models, we use the following hyper-parameter settings: optimizer=Adam, epochs=40, batch_size=8, and earlyStopping is used.

The models we compare are summarized in Table 1.

| Model Name         | Gender Classification Accuracy | Status Classification Accuracy |
|--------------------|--------------------------------|--------------------------------|
| DenseNet201        | 0.9506                         | 0.8558                         |
| InceptionResNetV2  | 0.9526                         | 0.8311                         |
| ResNet152V2        | 0.9298                         | 0.7780                         |
| VGG16              | 0.9526                         | 0.7799                         |
| Xception           | 0.9791                         | 0.8843                         |
| InceptionV3        | 0.9715                         | 0.7685                         |

5.3. Results for ensemble models
The parameter settings for the ensemble modules we use are summarized in Table 2. For DNN, it contains only one input layer, one hidden layer and one output layer. We use 512 neurons and relu activation function in the hidden layer, and softmax activation function in the output layer. The other parameters are as follows: optimizer=Adam, epochs=40, batch_size=8. We also use earlyStopping here.

| Model Name         | Parameter                  | Search Space                  |
|--------------------|----------------------------|------------------------------|
| AdaBoost           | base_estimator__max_depth  | [5, 10, 15, 20, 25]           |
|                    | n_estimators               | [20, 50, 100, 200]            |
| k-nearest neighbors| n_neighbors                | [3, 5, 10]                    |
| Random Forest (RF) | max_depth                  | [5, 10, 15, 20, 25]           |
|                    | n_estimators               | [20, 50, 100, 200]            |
The final gender and status classification accuracies achieved by different ensemble modules are shown in Table 3. As we can see from Table 4, the DNN module achieves the best result, beating all the individual models as well as other ensemble modules. This is also the best result for the KaoKore dataset in the literature as far as the author knows.

| Model Name                  | Gender Classification Accuracy | Status Classification Accuracy |
|-----------------------------|--------------------------------|-------------------------------|
| AdaBoost                    | 0.9791                         | 0.8918                        |
| k-nearest neighbors (kNN)   | 0.9791                         | 0.8843                        |
| Random Forest (RF)          | 0.9772                         | 0.8994                        |
| Support Vector Machine (SVC)| 0.9772                         | 0.8956                        |
| XGBoost                     | 0.9734                         | 0.8994                        |
| Deep Neural Network (DNN)   | 0.9810                         | 0.9013                        |

We show the confusion matrix for the gender and status classification for our best results achieved by the DNN module in Figure 2 and Figure 3, respectively.

Figure 2. The confusion matrix for gender classification of the DNN ensemble model.
Figure 3. The confusion matrix for status classification of the DNN ensemble model.

6. Conclusion
In this paper, we used a Japanese artwork face data set named KaoKore [1] as a case study for characters' gender and identity in the artwork. We compared the ensemble of different deep learning models and individual models. Compared with an individual model, the DNN ensemble model achieved the best results in the literature. Our results demonstrate the promising application of artificial intelligence technologies in the culture relics protection.

References
[1] Tian Y, Suzuki C, Clanuwat T, et al. KaoKore: A Pre-modern Japanese Art Facial Expression Dataset[J]. arXiv preprint arXiv:2002.08595, 2020.
[2] Goel A, Vishwakarma V P. Gender classification using KPCA and SVM[C]//2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). IEEE, 2016: 291-295.
[3] Lee B, Gilani S Z, Hassan G M, et al. Facial Gender Classification—Analysis using Convolutional Neural Networks[C]//2019 Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2019: 1-8.
[4] Zhang K, Tan L, Li Z, et al. Gender and smile classification using deep convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016: 34-38.
[5] Downton W, Vadapalli H. Age and Gender Estimation using Optimised Deep Networks[M]//Proceedings of the South African Institute of Computer Scientists and Information Technologists 2019. 2019: 1-9.
[6] Tharwat A, Ghanem A M, Hassanien A E. Three different classifiers for facial age estimation
based on k-nearest neighbor[C]//2013 9th International Computer Engineering Conference (ICENCO). IEEE, 2013: 55-60.

[7] Han X, Ugail H, Palmer I. Gender classification based on 3D face geometry features using SVM[C]//2009 international conference on cyberworlds. IEEE, 2009: 114-118.

[8] Jiang W, Zhang L. Geospatial data to images: A deep-learning framework for traffic forecasting[J]. Tsinghua Science and Technology, 2018, 24(1): 52-64.

[9] Jiang W. Applications of deep learning in stock market prediction: recent progress[J]. arXiv preprint arXiv:2003.01859, 2020.

[10] Yan H, Ouyang H. Financial time series prediction based on deep learning[J]. Wireless Personal Communications, 2018, 102(2): 683-700.

[11] Jiang W, Zhang L. Edge-siamnet and edge-triplenet: New deep learning models for handwritten numeral recognition[J]. IEICE Transactions on Information and Systems, 2020, 103(3): 720-723.

[12] Jiang W. Time series classification: nearest neighbor versus deep learning models[J]. SN Applied Sciences, 2020, 2(4): 1-17.

[13] Jiang W. Evaluation of deep learning models for Urdu handwritten characters recognition[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1544(1): 012016.

[14] Sheng J. Enlightenment Digital Dunhuang of China's Cultural Heritage Protection[J]. China Cultural Relics News, 2020, 4.

[15] Zha X. Digital Image Processing: A Case Study of Tibetan Murals[D]. 2017.

[16] Sun L. Single-label and Multi-label Classification for Cultural Relic Images Based on Deep Learning: 2019.

[17] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[18] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[J]. arXiv preprint arXiv:1602.07261, 2016.

[19] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//European conference on computer vision. Springer, Cham, 2016: 630-645.

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[21] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.

[22] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.

[23] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.