Intricate Face Recognition Based On Virtual Sample Generation

Qionglin Tong, Jinshan Su*, Lan Ma and Zhenhua Wang
School of electronic and Information Engineering, Micro-nano Electric Sensing Technology and Bionic Devices Key Lab, Ili Normal University, Yining, Xinjiang, China

*Corresponding author e-mail: sqsjs1968@aliyun.com

Abstract. This paper mainly focuses on the problem of low recognition rate and making deep hidden features unable to learn caused by the lack of training samples in intricate facial recognition, a method of virtual image sample generation is proposed. The starGAN network based on style migration is used to generate the image that does not exist in the original database, and multiple images with different facial attributes are added to a single sample, combine the obtained virtual samples with the original database to build a larger face database for face recognition. Experiments on the CelebA face database show that the proposed method can effectively improve the face recognition rate in complex situations.

Keywords: Intricate Facial Recognition, Virtual Samples, Style Migration, Deep Learning

1. Introduction
In actual recognition occasions, people who normally face the lens often have lighting, expressions, poses, and occlusions on their faces, as shown in Fig. 1. There are some difficulties in the recognition of these complex faces. Deep Learning (DL) [1] has been proposed as a new machine Learning algorithm, which has attracted wide attention from researchers in the industry. Deep learning can automatically learn features from training samples, which mainly benefit from multi-level feature representation and training with massive sample data, thus learning robust facial representation of changes in training data, to make the extracted features more efficient. If complex face recognition is the key problem that face recognition system must solve, then how to obtain a large number of high-quality occlusion images is the main way that must be solved in occlusion recognition. Although the existing open source data set [2-3] contains hundreds of thousands of face images, it is difficult to find the recognition rules due to lack of pose changes and occluded images. At the same time, the images manually collected or obtained from the network are confusing and need to be manually labeled. For small laboratories conducting academic research, obtaining a large number of complex samples and labeling them is time-consuming and laborious. Based on the above content and the idea of enhancing the training samples by the Masi [4] synthesis data, this paper proposes to generate a style sample in the adversarial network to generate virtual sample pictures with different poses and occlusions that are
indistinguishable from the real samples, increasing the diversity of samples, making up for the lack of sample size in the data set.

The innovation and contribution of this paper can be summarized as follows:

1. In order to solve the problem of sample shortage in deep learning for complex facial recognition, a virtual sample generation method based on Generative Adversarial Networks (GANs) is proposed. Depending on the recognition object, virtual samples are selectively generated to ensure the diversity and realism of virtual samples.

2. Complex face recognition systems such as large pose and large area occlusion are constructed to achieve high recognition accuracy and strong robustness.

Fig 1. An example of how incomplete facial may be presented as input probe images for face recognition

2. Relevant Work of the Proposed Methodology

2.1 Nonlinear Mapping in Deep Learning

One of the main advantages of deep learning methods is that deep layer features in image data can be captured through the learning of multi-level neural networks, and detailed feature decomposition is performed on each layer of the image to obtain more accurate features. Secondly, a large amount of data can be used to train the data in order to learn the nonlinear laws of complex faces.

Many scholars realize that the process of extracting image features is actually the process of sparse approximation of the image, considering the normed space \( X_0 \) and its subspace \( X_n \subset X, n=0,1,2,\ldots \), and setting the approximated image \( f \subset X \), the element of subspace \( X_n \) is used as the approximation element to approximate the image \( f \), and \( g \) is the approximation image composed of the elements of \( X_n \), the approximation error is:

\[
E_n(f)_X := \text{dist}(f, X_n)_X := \inf_{g \in X_n} \| f - g \|_X
\]  

(1)

The element of subspace \( X_n \) in Eq. (1) is the extracted feature, and feature selection is the key preprocessing link of classification and recognition. Because deep learning is the decomposition of multi-level structure, each layer can get features on different levels, the correct and effective features can be extracted and the generalization ability can be improved. It can be seen from Fig.2 that these features, as invisible texture features like hen, Yorkshire terrier, Shetland sheepdog, fountain, theater curtain and geyser, have the characteristics of local irregularity and macro regularity. These images are useful for understanding and diagnosing network behavior. The example uses Deep Learning Toolbox™ and Deep Learning Toolbox Model for AlexNet Network to generate the images [12].
Fig 2. Multiple images at once generated by selecting multiple classes by using AlexNet Network

The classification and recognition of complex faces is a high-dimensional mapping problem, and the number of samples increases exponentially with the increase of sample dimension. If the sample distribution over the whole high-dimensional space, then the "dimensional disaster problem is inevitable" at this time, massive samples must be collected, and the samples should cover the entire high-dimensional space as much as possible to learn the hidden regularity in the samples. However, in this study, the samples are only distributed in a subspace or a local sub-region in a high-dimensional space, which are correlated with each other, thus reducing the dimension and avoiding the need for massive samples. Based on the deep learning methods changes in expression [5], illumination changes [6], age [7] of face recognition has made a historic breakthrough, but the application of complex scenarios and a way to go, because of the deep learning method is limited by the lack of high quality, large-scale training data, to posture change, keep out of the facial recognition rate needs to be improved.

2.2 Style Migration of Virtual Samples

Although CGAN can generate "specified" face images, at the same time, the model will also learn other noises, which will cause other face changes in the generated image. Moreover, for data sample sets that contain multiple classes, it is necessary to learn one category by one and generate virtual samples of corresponding categories. However, the feature of multi-posture and occlude face image is that the local face changes, and the generation of complex face samples can be realized through local image migration, which is called style migration. The main task of image style transfer is to combine the given content image (C) and style image (S) to generate a style transfer image with content features and style features, the final result image (R) has both content features of content image (C) and style image (S).

Let $X_s$ be the subspace of the higher dimensional space $X$, $X_s \subset X$. The source image $f$ be satisfy $f \subset X_s$ with migration part $f_c$, have $f_c \subset X_s$, $f_i \subset f$, the part $f_i$ moved to target image as migration conditions C, then the formed image (virtual samples) $f'$ is

$$f' = (f - f_d) \cup f_c$$

Where $f_d$ is the replaced portion in target image $f'$, $f' \subset X_s$, as shown in the Fig.3.
Fig 3. Style migration and formed virtual samples

The error of approximation is:

$$E_{n}(f)_{X_i} := \inf_{g_{x_{n}}} \| f_{i} - g_{i} \|_{X_{i}}$$  \hspace{1cm} (2)

Where the elements of $X'_{n}$ are the migrated features, $g_{i}$ is the approximate image of the combination of features in $X'_{n}$, in this case, let $X_{i}$ be a subspace of the high dimensional space $X$, according to the above, we can learn the hidden regularity of the sample without the need for massive samples.

3. Identification System

The face recognition model flow of this paper is shown in fig.4.

Fig 4. Face recognition system process

Fig.4 shows the process of face recognition system in this paper. The model structure used in the recognition stage is ResNet50. This model was proposed by Kaiming He and other four Chinese from Microsoft Research Institute in 2015. We know that the deeper the network, the more information can be obtained from it, and the richer the feature information. According to experiments, as the network deepens, the optimization effect of the deep convolutional network becomes worse, and the test data and training data are accurate. The rate is also relatively lower, because the deepening of the network layer will cause the problem of gradient explosion and disappearance. In the past, the solution for more than ten layers of network is to normalize the input data and the data of the middle layer, so as to make the network converge. For the deeper network, this method has no effect and significance. Therefore, compared with the deep neural network which has made good progress in face recognition, the ResNet50 are easier to optimize, facilitate training, and improve accuracy.
3.1 Feature Classification
The core of classification task is to assign a label to an image from a given classification set. In recent years, classification algorithms that have performed well in face recognition tasks include decision trees [8], SVM [9], and KNN [10]. In the work of this paper, all the features extracted during training and testing are used for classification tasks. In our experiment, for the classification method, we used cosine similarity (CS) [11] and K-Nearest Neighbor (KNN) [10]. We have tested other classifiers, using CS and KNN to achieve the best results, which is the reason why we choose these two algorithms. Moreover, through experiments and analysis, we found that the combination of these two classifiers can better overcome the problem of facial occlusion in complex situations. The KNN method mainly depends on the limited adjacent samples, rather than on the method of identifying the class field. Therefore, KNN method is more suitable than other methods for the sample set to be divided with more overlapping or overlapping class fields.

4. Experiments and Results

4.1 The CelebA Dataset
The experiment was performed on a large face attribute dataset of CelebFaces Attributes Dataset (CelebA) (Fig. 5), which contains 202,599 face pictures of 10,177 people, and the resolution of the images is $178 \times 218$. This dataset covers all types Face images, each of which has been labeled with features, including a face b-box labeling box, five facial feature point coordinates and 40 binary attribute annotations. CelebA is provided by the Chinese University of Hong Kong and is widely used in computer vision training tasks related to face technology, such as face attribute identification training and face detection training. The following image shows some sample images from the CelebA face dataset.

![Fig 5. Typical face samples of CelebA dataset](image)

4.2 Conditional Virtual Sample Generation
**Fig 6.** Loss values of discriminator and generator against generated virtual samples
Figure 6 shows the number of iterations and the loss value between the generator and the discriminator. It can be seen from the figure that when the training samples are enough and the sample attributes are single and clear, no matter how many iterations, the loss value of the generator and discriminator will keep balance, and the loss difference is about 0.5.

4.3 Experiments on intricate face using the CelebA dataset
In our experiment, the following training and experiments are carried out for face recognition in virtual images with occlusion. First, the Resnet50 network is trained using the original CelebA data set, and then the virtual samples with sunglasses occlusion and mask occlusion generated by starGAN based on style transfer (Fig.7). The image is added to the original training image to construct a new occlusion face experimental data set. This step increases the amount of data required for deep learning, and also avoids the overfitting phenomenon that occurs when the recognition is too close to the training data.

It is necessary to adjust parameters in training network model. The size of learning rate setting and the number of iterations will affect the final training model. Through many experiments, we choose a real number between [0,1] for the selection of learning rate. The learning rate controls the update step size of each epoch. If the learning rate is set too large, the network will vibrate. If the setting is too small, the convergence speed of the network will be very slow, so the loss response is too slow. In this paper, in the initial training model, the learning rate is set to 0.01. When the model achieves a good recognition rate, the learning rate is reduced to 0.001 and the momentum is adjusted to 0.9.

**Fig 7.** Virtual face image generation with sunglasses and masks
After the features are extracted from the model, resnet50 is used to study the recognition rate of human face in the shade of sunglasses and masks.

**Fig 8.** Face recognition based on sunglasses occlusion
Finally, in the recognition experiment, we divide the combined test data set into two independent data sets, male and female. In Figure 8, the Cd in the recognition image represents the recognition confidence, and the yellow box represents the face location area detected from the image by resnet50 network. M1, M2 ,M1000,…… ,M2000,…… , denoted as identity information in the male dataset, while W1, W2 ,W1000,…… W2000,…… , is the identity information in the female dataset. As can be
seen from Fig. 8, the recognition has achieved high robustness based on a large number of occluded face image data.

5. Conclusions
In order to increase the diversity of samples and make up for the lack of sample numbers in the data set, a style transfer is presented by using StarGAN to generate adversarial networks to generate virtual sample pictures with different poses and occlusions that are not different from real samples. Although the ResNet50 model achieves the expected classification and recognition effect, the data set used in this article is manually collected on the Internet and part of the self-photographed, the background interference intensity is not very large, and the number of data sets is not very large. It needs to be increased in the future. More images are used for training and verification, and higher recognition accuracy can be made for occluded faces in intricate situations.

Acknowledgement
This research was funded by the Natural Science Foundation of Xinjiang (Grant No. 2016D01C382).

References
[1] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, Nature. 2015, 521(7553):436-444.
[2] GB Huang, et al., Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, E Learned-Miller - cics.umass.edu ,2008.
[3] Z. Liu, et al., Deep Learning Face Attributes in the Wild, Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015, pp. 3730-3738
[4] I. Masi, et al., Do We Really Need to Collect Millions of Faces for Effective Face Recognition? ECCV 2016: Computer Vision – ECCV. 2016, pp 579-596
[5] A. Fathallah, L. Abdi and A. Douik, Facial Expression Recognition via Deep Learning, 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA). 2161-5330
[6] N. P. Ramaiah, E. P. Lijina and C. K. Mohan, Illumination invariant face recognition using convolutional neural networks, 2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES).
[7] H. Zhai, et al., Face Verification Across Aging Based on Deep Convolutional Networks and Local Binary Patterns, Intelligence Science and Big Data Engineering. 2015:341-350.
[8] S.K. Murthy, Automatic construction of decision trees from data: A multidisciplinary survey, Data Min. Knowl. Discov. 2 (4) (1998) 345–389.
[9] X. Han, H. Ugail and I. J. Palmer, Gender classification based on 3D face geometry features using SVM, 2009 International Conference on CyberWorlds. Bradford, UK, 2009, pp. 114–118.
[10] Z. Zhang, Introduction to machine learning: k-nearest neighbors, Ann.Transl. Med. 4 (11) (2016) 218
[11] G. Sidorov, et al., Soft similarity and soft cosine measure: Similarity of features in vector space model, Comput. Sist. 18 (2014) 491–504, http://dx.doi.org/10.13053/CyS-18-3-2043.
[12] Deep Learning Toolbox Model for AlexNet Network, https://ww2.mathworks.cn/matlabcentral/fileexchange?q=AlexNet