Freehand Sketching Portrait Recognition with Least Square CycleGAN

Xiaosa Gou¹, Bingguo Liu*, Guodong Liu¹, Binghui Lu¹, Yu Gan¹ and Fengdong Chen¹

¹ School of Instrumental Science and Engineering, Harbin Institute of Technology, Harbin, Heilongjiang, 150001, China

*liu_bingguo@hit.edu.cn

Abstract. Freehand sketching portrait recognition refers to the recognition of sketched portraits and face photos drawn by artists. Existing research mainly involves inputting a given real portrait and converting it into a similar sketch portrait, and then matching the face with the sketch portrait in the real portrait database. This paper starts from the idea of inputting a sketch image to identify the real portrait in the database, focuses on the research based on the method of portrait synthesis, and introduces the existing methods of sketch portrait identification. We use the CycleGAN improved by least squares to achieve the translation from sketch to portrait, and finally PCA was used to complete face matching. The results show that LS_CycleGAN has certain advantages. Compared with other methods, its synthesis results are most close to the portrait. The average score of SSIM and PSNR is 0.844 and 17.331, and the average success rate of recognition is 88.2% for Rank10.

1. Introduction

Sketch portrait, as a major branch of heterogeneous portrait, has important research value in face images in heterogeneous mode, such as thermal infrared portrait, sketch portrait, 3D portrait and cartoon portrait. Especially in recent years because of its important role in law enforcement activities has attracted extensive attention to the current research on sketch portrait identification[1]. In the process of law enforcement, effective facial photos cannot be obtained due to the criminal suspect's anti-detection awareness or other reasons. When the police have limited clues, portrait experts will be invited to paint a portrait of the suspect. According to the portrait, facial photo database will be used to further narrow the scope of investigation. However, due to the large difference in the modal of sketch and real portrait, the traditional homologous face recognition method performs poorly in sketch portrait recognition. Owing to the images researched by the two have different ways of information expression:

(1) Their geometric dimensions are inconsistent. Face photos can basically accurately reflect the eatures of facial faces, such as the relative positions and sizes of five features, etc. Face sketch images are affected by some human factors, and their geometric dimensions are bound to exist in certain deviations.

(2) They contain different amounts of information, and facial photos can reflect detailed information about the face, such as the geometric size of facial features, facial contour, grayscale and facial structure. The face sketch image is only composed of lines and lacks some details of face features.
Their details differ greatly. In the description of facial features, there are great differences between face sketch image and face photo. Therefore, an automatic portrait recognition method is needed to accurately and quickly retrieve facial photo data set or surveillance video. At present, there are three strategies to improve the modal differences: the method based on portrait composition, the method based on portrait features, and the method based on portrait common subspace.

For feature-based recognition method. That is to find some features in the portrait, which are less affected by the difference between the sketched portrait and the real portrait. The matching of the portrait can be completed by extracting these features. In 2010, Bhat et al.[2] proposed a sketched face recognition based on segmented LBP. In 2011, Klare et al.[3] proposed a discriminant analysis based on local features. However, the characteristics of manual design could not adapt to different painting styles, that is, the results obtained by the same method in different databases are very different, and lack of universality.

For subspace-based recognition method. A subspace between the sketch and the true image is constructed to reduce the difference between them by projecting them into the space. At this point, we can directly compare two images of a common subspace. In 2011, Sharma et al.[4] used partial least squares (PLS) to linearly map images in different forms to a common linear subspace. In 2013, Huang et al.[5] proposed a unified model of coupling learning between dictionary and feature space. However, some important information is missing in the constructed public subspace of the target portrait, which is bound to have a great impact on the accuracy of sketch face recognition.

Due to the lack of generality and reliability of the above two methods. Therefore, there are more and more researches based on image synthesis. Unfortunately, most of the existing research is based on synthesizing portraits into sketches. In the case of a large database, converting all portraits into sketches is a time-consuming process with poor usability. This paper research the method of sketch synthesis portrait and proposes LS_CycleGAN model. In this research, face matching can be accomplished by only synthesizing a portrait.

2. Synthesis Methods

2.1. Related Work

The method based on portrait synthesis is the most widely used scheme in this field at present. It is to synthesize a portrait in one mode into a portrait in another mode, realize the mutual transformation of face images in different modes, and then use the traditional face recognition algorithm for face matching. In 2004, Liu et al.[6] proposed image composition method of segmental Linear approach to global nonlinear. The algorithm framework is as shown in the Figure 1. Firstly, all portraits are divided into \( N \) image blocks according to a certain size. The relation between the input sketch \( I_i \) and the output image block \( I_o^t \) of portrait \( I_o \) can be obtained as follows:

\[
I_o^t = \sum_{n=1}^{K} w_{in}^{\xi} I_{pn}^t
\]  

(1)

\( I_o^t \) represents the \( t \)-th image block in the output portrait, \( K \) represents the number of similar image blocks searched, \( w_{in}^{\xi} \) represents the weight relationship between \( I_i \) and \( I_{pn}^t \) of the image block \( p \) in the real portrait library in the training set, and \( I_{pn}^t \) represents the image block close to the \( N \)-th Euclidean distance from \( I_o^t \). Finally, \( \{I_1^t, ..., I_k^t\} \) is obtained by the best suture method to output portrait \( I_o \).

In 2009, Wang et al. [7] proposed Markov Random Fields (MRF) algorithm to model the relationship between sketched portrait and real portrait to represent the relationship between different image blocks in an image. In 2012, Zhou et al. [8] proposed Markov Weight Fields (MWF) in view of the MRF algorithm that did not carry out the research on the Weight factors between image blocks. The algorithm framework is shown in the Figure 1. Firstly, all portraits are divided into \( N \) image blocks according to a certain size. The image block \( I_i^m \) of sketch \( I_i \) corresponds to the image block \( I_{x1}^m, t = 1,2, ..., k \), of the
sketch portrait library in the training set with $K$ image blocks with the nearest Euclidean distance. The relation between $I^m_0$ and weight $w^m_0$ is shown in the following formula:

$$p(I^m_0, I^m_1, \ldots, I^m_n) \propto \prod_{m=1}^K \Phi(I^m_0, w^m) \prod_{(m,n) \in \Xi} \Psi(w^m, w^n)$$

$(m, n) \in \Xi$ means that $m$th and $n$th patches are adjacent blocks.

$$\Phi(I^m_0, w^m) = \exp\left\{-||I^m_0 - \sum_{l=1}^K w^m_l I^m_l||^2 / 2 \sigma_D^2\right\}$$

$$\Psi(w^m, w^n) = \exp\left\{-||\sum_{l=1}^K w^m_l o^m_{ln} - \sum_{l=1}^K w^n_l o^n_{lm}||^2 / 2 \sigma_S^2\right\}$$

$o^m_{ln}$ denotes the overlapping area of the $l$th candidate for the $m$th sketch patch with the $n$th patch, $\sigma_D$ and $\sigma_S$ is the preset two parameters, it solve for $\{w^m_0, \ldots, w^m_M\}$ and $\{I^m_0, \ldots, I^m_N\}$ into the (1). Finally, $\{I^1_0, \ldots, I^N_0\}$ is obtained by the best suture method to output portrait $I_p$.

In 2014, Song et al. [9] proposed Spatial Sketch Denoising (SSD) in view of the noisy image synthesis using LLE method. In 2017, Gao et al. [10] proposed a synthesis method based on random sampling to solve the problem of slow speed of searching the $K$ most similar image blocks, and replaced the searched $K$ most similar image blocks with the randomly sampled $M$ image blocks, which improved the synthesis speed.

With the development of deep learning, the appearance of GAN[11] in 2014 provided a new method for portrait synthesis. Typical methods include supervised image translation-Pix2Pix[12], that algorithm framework is shown in the Figure 2. Supervised training data are in pairs, Pix2Pix consists of
discriminant model $D$ and generation model $G$, random noise generation model is input, and then as far as possible to fit the real data as possible to trick the discriminant model, the distribution of discriminant model is used to determine the input data from real training or generation model output data, eventually making $G$ has ability to generate strong enough sample that there is no way for model $D$ to distinguish between real samples and generated samples. Although the method of synthesis reduces the difference of portraits in different modes, it makes the portraits from different sources comparable. However, there are still some disadvantages:

1. different synthesis algorithms need to be designed for different modes.
2. There is still a certain gap between synthetic portraits and real ones.
3. The fidelity and robustness of the synthesis method directly affect the accuracy of face matching.

2.2. LS_CycleGAN
The model in this paper is an improvement on CycleGAN[13], and the algorithm framework is shown in Figure 2. CycleGAN is an unsupervised image translation that does not need paired data, only one set of input images (such as sketch portraits) and a set of output images (such as real portraits). namely a Sketch→Photo one-way GAN plus a Photo→Sketch one-way GAN. Two GAN’s share two generators, and then take one discriminator each, so there are two discriminators and two generators. We use the Least Squares Loss instead of the original GAN Loss, that make network training more stable and improve image quality.

\[
\min_D J(D) = \min_D \frac{1}{2} E_{x-p_x}[D(x) - a]^2 + \frac{1}{2} E_{z-p_z}[D(G(z) - b)]^2
\]

\[
\min_G J(G) = \min_G \frac{1}{2} E_{x-p_x}[D(G(z) - c)]^2
\]

$D(x)$ represents the discriminator, $G(z)$ represents the generator, and the random variable $z$ obeys the standard normal distribution. Constants $a$ and $b$ represent the markers of the real images and the generated images. $c$ is the value set by the generator $G$ to let the discriminator $D$ think the generated image is real image. $p_x$ is the probability distribution that $x$ obeys, $p_z$ is the probability distribution that $z$ obeys, $E_{x-p_x}$ and $E_{z-p_z}$ is the expected value.

In order to keep the content in the sketch and only change the style. Cycle Consistency Loss can solve this problem.

\[
J_{cycle}(G, F) = E_{x-p_x}[||F(G(x) - x)||_1] + E_{y-p_y}[||G(F(y) - y)||_1]
\]

The input to $G$ is $x$, which is a generator that creates a false graph of $Y$, and the input to $F$ is $y$, which is a generator that creates a false graph of $X$.

In order to prevent the generator autonomously to modify the image color so as to change the overall color. We restrict it by adding Identity Loss.

\[
J_{identity}(G, F) = E_{x-p_x}[||F(x) - x||_1] + E_{y-p_y}[||G(y) - y||_1]
\]

3. Results and analysis

3.1. Implementation details
The experiment adopted the CUFs database of The Chinese University of Hong Kong as its Face Sketch that is open as a photo that is collected from three sub-databases: CUHK database, AR database and XM2VTS database, which contain a single positive neutral expression photo of 188, 123 and 295 people respectively. Each photo in CUFs database contains a hand-drawn portrait by an artist, that is, there are 606 pairs of sketched portraits and real portraits, and the size of the images is 200×250. The style of sketch portraits varies between the three databases, but the same database has the same style. This paper selects some representative methods from various methods for comparison: based on LLE, RSLCR, MRF, SSD, Pix2Pix, LS_CycleGAN. In order to conduct face portrait synthesis, the database needs to be divided into training data and testing data according to training/testing, so CHHK 88/100, AR 72/43, XM2VTS100/195.
For LLE, we set the number of candidate correspondences $K$ is set to 5, the search radius of displacement vector to 5 pixels in the $K$-NN search process, Image block size $N \times N =400$, In order to enhance the local compatibility and smoothness between adjacent synthetic sketch blocks, we use the average algorithm for overlapping areas in the final reconstruction results, so overlap =14. For RSLCR, we set the number of random samples $M =500$, the weight between reconstruction error and locality constraint $\alpha =0.5$, and the remaining parameters are the same as LLE. For SSD, we set the scope of denoising $\psi =8 \times 8$, and the remaining parameters are the same as LLE. For MRF, we set $\sigma_D = 0.2$, $\sigma_S = 1$, and the remaining parameters are the same as LLE. For Pix2Pix, we use original GAN loss function, Adversarial Loss weight=1, $L_1$ Loss weight=10, batch_size=1, learning rate=0.0002. For LS_CycleGAN, $a=-1$, $b=1$, $c=0$, Cycle Consistency Loss weight=10, Identity Loss weight=5, batch_size=1, learning rate=0.0002.

3.2. Evaluation methods

For the practicability of the experimental algorithm, the synthetic results were used for human visual representation, image quality evaluation and face recognition, and the results of the three databases were scored for image quality evaluation and compared respectively. The image quality evaluation algorithm uses SSIM[14] and PSNR[15] evaluation method. The face recognition algorithm uses Principal Component Analysis[16].

![Figure 3. Synthesized photos on the CUHK, AR, XM2VTS. (a) Input sketch. (b) Truth photo. (c) LLE. (d) RSCLR. (e) MRF. (f) SSD. (g) Pix2Pix. (h) LS_CycleGAN.](image)

3.3. Experiment results

The result of composition of AR and XM2VTS database is shown in Figure 3. It can be seen from the figure that there are many noise or block effects in LLE algorithm. RSLCR algorithm synthesis details (such as nose) is better, but there is also more noise. The synthesis results of SSD algorithm are very delicate, but the image is relatively fuzzy and many features are missing. The synthesis result of MRF algorithm has deformation, which cannot well synthesize the main structure of the face (such as eyes or hair). Pix2Pix algorithm can keep the texture features of the portrait well and the details are relatively complete. Human eyes can intuitively feel close to the portrait, but the deformation is serious. LS_CycleGAN algorithm overcomes some deformations of GAN, but it may cause color distortion. In general, each algorithm is closer to the real portrait than the sketch portrait, among which LS_CycleGAN has the best visual perception.
The image quality score is shown in Figure 4. SSIM is an index to measure the similarity of two images. Its value range is [-1,1]. When two images are identical, the value of SSIM is 1. PSNR is based on the error between corresponding pixels. The larger the value is, the better the image quality is. Both structure similarity and peak SNR are improved after adopting the synthetic method, especially the SSIM score in XM2VTS database is greatly improved. In general, the best performance of the two indexes is RSLCR algorithm, whose average SSIM score reaches 0.8685 and average PSNR score reaches 18.2315. The reason why Pix2Pix and LS_CycleGAN image quality is not as good as that of the machine learning algorithm is that the GAN generated image is difficult to build good details, and the image often lacks some information and features, so the image quality is slightly lower than that of the machine learning algorithm based on sample synthesis.

The success rate of identification is shown in Figure 5. Comparing with the direct recognition of sketch map, 6 methods enhance the recognition accuracy 15-30% in the database of real portrait database.
Among them, the recognition rate of LS_CycleGAN reached 78% for Rank10, which is the highest among algorithms. In the AR database, thanks to the painter's style being close to the real portrait, the synthesis effect of each algorithm is very good, and the identification success rate reaches 100%. In XM2VTS database, LS_CycleGAN has a good performance, and the recognition success rate is higher than other algorithms, up to 86.7%, which is 72.06% higher than the direct recognition of sketch portraits.

4. Conclusion
This paper research the transformation from sketch to portrait, and the proposed LS_CycleGAN algorithm is superior to other algorithms. Human eyes can intuitively see that the synthesized portrait is closer to the real photo, and it improved by 4.7% on average compared to other algorithms for Rank10.

The rationale behind this is to use cyclic consistency loss, which makes the resulting image not only realistic but also contains information about the original image. Moreover, the least square loss is used to replace the cross-entropy loss, which improves the quality of the generated images and makes the training more stable. However, due to the limitation of database size, the success rate of recognition is not significantly improved compared with other algorithms. In the future, we will further improve the recognition accuracy and realize the synthesis of different sketch styles.

Reference
[1] Wang, Nannan, et al. "A Comprehensive Survey to Face Hallucination." International Journal of Computer Vision 106.1(2014):9-30.
[2] Bhatt, Himanshu S., et al. "On matching sketches with digital face images." Fourth IEEE International Conference on Biometrics: Theory Applications & Systems IEEE, 2010.
[3] Klare, Brendan, Z. Li, and A. K. Jain. "On Matching Forensic Sketches to Mugshot Photos." IEEE Transactions on Pattern Analysis & Machine Intelligence 33.3(2011):639-46.
[4] Sharma, Abhishek, and D. W. Jacobs. "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch." IEEE Conference on Computer Vision & Pattern Recognition IEEE, 2011.
[5] Huang, De An, and Y. Frank Wang. "Coupled Dictionary and Feature Space Learning with Applications to Cross-Domain Image Synthesis and Recognition." (2013).
[6] Liu, Qingshan, et al. "A nonlinear approach for face sketch synthesis and recognition." IEEE Computer Society Conference on Computer Vision & Pattern Recognition IEEE, 2005.
[7] Wang, Xiaogang, and X. Tang. "Face Photo-Sketch Synthesis and Recognition." IEEE Transactions on Pattern Analysis & Machine Intelligence 31.11(2009):1955-1967.
[8] Zhou, H., Z. Kuang, and K. Wong. "Markov Weight Fields for face sketch synthesis." IEEE Conference on Computer Vision & Pattern Recognition IEEE, 2012.
[9] Song, Yibing, et al. "Real-Time Exemplar-Based Face Sketch Synthesis." (2014).
[10] Gao, X., et al. "Face Sketch – Photo Synthesis and Retrieval Using Sparse Representation." IEEE Transactions on Circuits & Systems for Video Technology 22.8(2012):1213-1226.
[11] Goodfellow, Ian J., et al. "Generative Adversarial Networks." Advances in Neural Information Processing Systems 3(2014):2672-2680.
[12] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," arXiv preprint arXiv: 1611.07004, Tech. Rep., 2016.
[13] Zhu, Jun Yan, et al. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks." (2017).
[14] Wang, Zhou. "Image Quality Assessment: From Error Visibility to Structural Similarity." IEEE Transactions on Image Processing (2004).
[15] Alain Horé, and D. Ziou. "Image quality metrics: PSNR vs. SSIM." 20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010 IEEE Computer Society, 2010.
[16] Shlens, Jonathon. "A Tutorial on Principal Component Analysis." International Journal of Remote Sensing 51.2(2014).