Deep Graphical Feature Learning for Face Sketch Synthesis

Mingrui Zhu[†], **Nannan Wang**[‡*], **Xinbo Gao**[†], **Jie Li**[†]
[†] State Key Laboratory of Integrated Services Networks,
School of Electronic Engineering, Xidian University, Xi'an 710071, China
[‡] State Key Laboratory of Integrated Services Networks,
School of Telecommunications, Xidian University, Xi'an 710071, China
mrz.edu@gmail.com, nnwang@xidian.edu.cn, {leejie,xbgao}@mail.xidian.edu.cn

Abstract

The exemplar-based face sketch synthesis method generally contains two steps: neighbor selection and reconstruction weight representation. Pixel intensities are widely used as features by most of the existing exemplar-based methods, which lacks of representation ability and robustness to light variations and clutter backgrounds. We present a novel face sketch synthesis method combining generative exemplar-based method and discriminatively trained deep convolutional neural networks (dCNNs) via a deep graphical feature learning framework. Our method works in both two steps by using deep discriminative representations derived from dCNNs. Instead of using it directly, we boost its representation capability by a deep graphical feature learning framework. Finally, the optimal weights of deep representations and optimal reconstruction weights for face sketch synthesis can be obtained simultaneously. With the optimal reconstruction weights, we can synthesize high quality sketches which is robust against light variations and clutter backgrounds. Extensive experiments on public face sketch databases show that our method outperforms state-of-the-art methods, in terms of both synthesis quality and recognition ability.

1 Introduction

Face sketch synthesis aims at solving two problems: 1)given a face photo, synthesize a sketch drawing; 2)given a query face sketch drawing, retrieve face photos in the database. It can effectively assist law enforcement [Wang *et al.*, 2014]. In many criminal cases, suspects are well disguised so that the clear photo images of their faces are difficult and sometimes even impossible to acquire. Under these circumstances, normal face recognition methods based on photo images would not be effective. The best substitute is often a sketch drawing based on the recollection of an eyewitness [Tang and Wang, 2009]. Due to the great texture discrepancy between the query sketch drawing and gallery mug shots, the specially devised face sketch synthesis method is extremely important. In addition, transforming photos into sketches is also a useful application in the entertainment [Song *et al.*, 2014].

Over the past decade, substantial advances have been made in face sketch synthesis. Among the various face sketch synthesis methods, exemplar-based methods achieves the best performance which benefit from their success in detecting and exploiting patch correspondences within a training database or calculating optimized dictionaries allowing for highly sparse data representation. Another growing branch formulate the map between photos and sketches as a regression problem [Zhu and Wang, 2016], whose advantage lies in their computational speed. Recently, convolutional neural network (CNNS) has also been used as a nonlinear regression model [Zhang *et al.*, 2015]. However, due to the limited representation ability of their loss function (mean squared error, MSE), their results got blur effect and thus have less satisfied perceptual quality.

The most striking successes in deep learning have involved discriminative models for tasks such as image classification and object recognition [Goodfellow *et al.*, 2014]. It was shown that such models could learn to extract high-level image content in generic feature representations which simultaneous possess discriminative capacity. Gatys *et al.* [2016] have recently utilized the filter pyramid of the VGG network [Simonyan and Zisserman, 2014] as a higher-level feature representations for image style transfer. The method yields very impressive artistic style results. However, feature pyramid at different layers are directly used to capture only pixel correlations of the image "style" and "content" so that the results lack of structural similarity.

In this paper, we combine the the generative exemplar-based method with discriminatively trained deep convolutional neural networks for face sketch synthesis. Feature maps (deep representations) derived from the VGG network, which was trained to perform ImageNet [Deng *et al.*, 2009] classification, are used to represent the face photos. Such feature maps with multi channels will have a large amount of redundant information. So we linearly combine those channels weighted by a weight vector. At the beginning, all channels have the same proportion, but a better weight combination will be learnt later. All face photos and sketches are divided into patches with overlap between the neighboring patches so that the transformation problem of the whole image is reduced to the local patch level. Similarly, deep representations

*Corresponding author: Nannan Wang (nnwang@xidian.edu.cn)

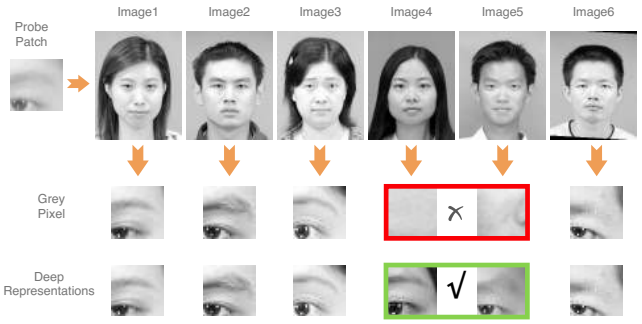

Figure 1: Comparison of neighbor selection between deep representations and pixel intensity.

of the whole photo is reduced to the patch level. Then, for each test photo patch, K nearest patches are selected from the training photo patches according to the Euclidean distance between their deep representations. Figure 1 shows that candidate patches selected based on their deep representations is superior to the candidates selected based on their pixel intensities. Sketch patches corresponding to the candidate photo patches are taken as the candidates for sketch patch synthesis and the prevalent way to synthesize the target sketch patch is the linear combination of selected candidate sketch patches. In order to jointly model the distribution over the weights for deep representations and the distribution over the weights for sketch reconstruction, we designed a graphical model-based deep feature learning framework.

The main contributions of this work are summarized as follows:

1) Deep discriminative representations derived from the deep Convolutional Neural Networks, which performs more robust respect to light variations and clutter backgrounds, are used to represent face photos. More accurate candidate sketch patches and weight combination for sketch patch reconstruction could be obtained benefiting from the deep representations.

2) A deep feature learning framework based on graphical model are proposed to jointly learn weights for deep representations and reconstruction weights.

3) Extensive experimental results illustrate the superior performances of our method compared with other state-of-the-art approaches in terms of both synthesis quality and recognition performance.

The rest of this paper is organized as follows. Section 2 outlines the existing literature on face sketch synthesis. Section 3 introduces the proposed deep graphical feature learning (DGFL) method. Experimental results and analysis are given in section 4 and section 5 concludes this paper.

2 Related Work

Recent works on face sketch synthesis include exemplar-based approaches and regression-based approaches.

Tang and Wang [2003] pioneered exemplar-based approach by the work of Eigen-transformation which assumed that the mapping between a photo and its corresponding sketch is a linear transformation when their shape and texture were processed separately. The reconstruction weights are learned by projecting the input photo onto the training photos through principal component analysis. However, whole face photos and face sketches cannot be simply represented by a linear transformation, especially when the hair region is considered. Liu *et al.* [2005] presented a locally linear embedding (LLE) based face sketch synthesis approach, with the idea of locally linear approximating global nonlinear. This method works on the image patch level. Each patch of the test photo was reconstructed from the linear combination of several selected training photo patches. Then the corresponding synthesized sketch patches could obtained from the linear combination of corresponding candidate sketch patches in the training set. This approach had the weaknesses that each patch was independently synthesized and thus neglected compatible relationships between the neighboring image patches. Therefore, neighborhood information is not well utilized and block effect was produced. Liu *et al.* [2007] developed a statistical Bayesian Tensor Inference between the photo patch space and the sketch patch space to capture and learn the inter-space dependencies. In order to introducing neighborhood information, Wang and Tang [2009] employed Markov random field (MRF) to model the distribution from two aspects: the distribution between test photo patches and nearest photo patches and the distribution between adjacent synthesized sketch patches. But this approach can not synthesize new patches existing not in the training set and its optimization is NP hard. Zhou *et al.* [2012] introduce the linear combination into the MRF model (namely Markov weight field, MWF) by formulated their model into a convex quadratic programming (QP) problem and proposed a cascade decomposition method (CDM) to solve it. Wang *et al.* [2017] proposed a Bayesian framework which provided an interpretation to existing face sketch synthesis methods from a probabilistic graphical view. Another branch of exemplar-based approaches is sparse representation-based methods. Sparse representation has been applied to various computer vision tasks, in which subsets weighted by a sparse vector are selected to represent the input signal. Chang *et al.* [2010] build a coupled dictionary with the photo and sketch patch pairs using sparse coding. The input test photo was decomposed on the photo elements in the coupled dictionary using sparse coefficients. The sketch patch could then be computed using the sketch elements in the coupled dictionary and the previously obtained sparse coefficients. Gao *et al.* [2012] proposed to adaptively determine the number of nearest neighbors by sparse representation and proposed a sparse-representation-based enhancement strategy to enhance the quality of the synthesized photos and sketches.

Regression-based approaches formulate the map between photos and sketches as a regression problem. They first learn a regression model between face photos (photo patches) and their corresponding face sketches (sketch patches) in the training stage. Then given a test face photo (photo patch), the regression model can predict the synthesized sketch (sketch patch). These approaches usually has a very fast speed in test stage because all time-consuming works are completed in training stage. Zhu and Wang [2016] proposed to use a

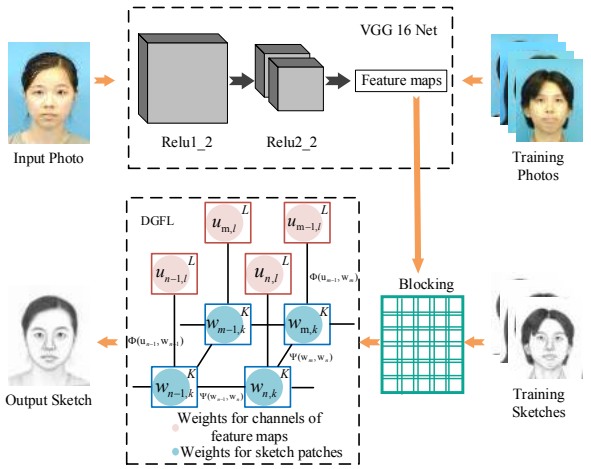

Figure 2: Framework of the proposed deep graphical feature learning for face sketch synthesis.

ridge regression model to learn the mapping between photo patches and their corresponding sketch patches in the same cluster. Convolutional neural networks (CNNs) has brought widespread attention in recent years and it also can be used as a nonlinear regression model. Zhang *et al.* [2015] proposed an end-to-end sketch generation model via fully convolutional networks (FCN) to directly model the complex nonlinear mapping between face photos and sketches. However, due to the lack of deep layers, this model failed to fully express the nonlinearity between face photos and sketches and thus its results have poor perceptual quality. In addition, the loss function of mean squared error (MSE) of regression model leads to blur effect.

3 Deep Graphical Feature Learning for Face Sketch Synthesis

In this section, we would introduce in detail how to represent photos with deep feature maps and then how to optimize the reconstruction weights and deep representations jointly in the proposed deep graphical feature learning framework. The overall framework of the proposed method is shown in Figure 2.

3.1 Deep Feature Extraction

Supposing there are M training photo-sketch pairs and a given test photo, which are geometrically aligned according to three points: two eye centers and the mouth center. Each image is cropped to the size of 250×200. We put all photos into a pre-trained 16-layer VGG network, which was trained to perform ImageNet classification and is described extensively in the original work [Gatys *et al.*, 2016], and perform forward propagation. Feature maps (deep representations) which have 128 channels derived from the relu2_2 layer of the network are used to represent the input photos. Each feature map generated by a convolution kernel could be regarded as a feature that concerned with a specific characteristic of the photos. Figure 3 provides some channels of deep representations for a test photo. As the size of convolution kernel in

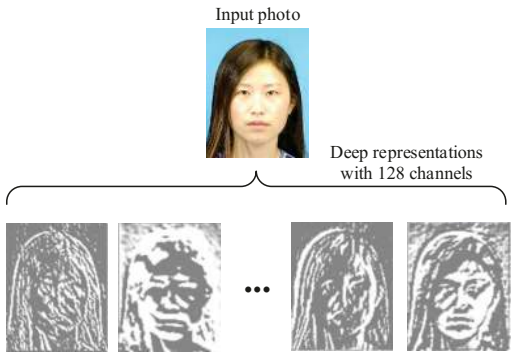

Figure 3: Deep representations of a test photo derived from VGG 16 Net.

VGG-16 network is 3×3, the feature maps we get would decrease in size. Thus, we resize each feature map back to 250×200 in order to consistent with the size of input photos. Then we divide all the photos, feature maps and sketches into N overlapping patches. Let $\boldsymbol{x}_i$ be the ith test photo patch, where $i = 1, 2, ..., N$. The deep representations of the test photo patch can be represented by the linear combination of 128 feature maps weighted by the 128 dimensional vector $\boldsymbol{u}_i$:

$$\mathrm{D}(\boldsymbol{x}_i) = \sum_{l=1}^{L} u_{i,l} \mathrm{d}_l(\boldsymbol{x}_i) \tag{1}$$

where $\mathrm{d}_l(\boldsymbol{x}_i)$ means lth feature map of patch $\boldsymbol{x}_i$, $u_{i,l}$ denotes the weight of lth feature map, $l = 1, 2, ..., L$, and $\sum_{l=1}^{L} u_{i,l} = 1$. Here L equals 128. Note that the weights of deep representations at various locations in photo patches are different, and each feature map has uniform weight in initial. Convergent weights for deep representations and weights for sketch reconstruction will be learnt later.

3.2 Deep Graphical Feature Learning

Given a test photo patch $\boldsymbol{x}_i$ and its deep representations $\mathrm{D}(\boldsymbol{x}_i)$, our target is to synthesize the corresponding sketch patch $\boldsymbol{y}_i$, where $i = 1, 2, ..., N$. Firstly, we find K candidate photo patches $\{\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, ..., \boldsymbol{x}_{i,K}\}$ that most like $\boldsymbol{x}_i$ according to the Euclidean distance of their deep representations and their corresponding sketch patches $\{\boldsymbol{y}_{i,1}, \boldsymbol{y}_{i,2}, ..., \boldsymbol{y}_{i,K}\}$ within the search region around the location of $\boldsymbol{x}_i$. We assume that deep representations of a test photo patch and its target sketch patch can be represented by the same linear combination of the K candidate photo patches' deep representations and sketch patches respectively. Then, the target sketch patch $\boldsymbol{y}_i$ can be obtained by the linear combination of K candidate sketch patches weighted by the K-dimentional vector $\boldsymbol{w}_i$:

$$\boldsymbol{y}_i = \sum_{k=1}^{K} w_{i,k} \boldsymbol{y}_{i,k} \tag{2}$$

where $w_{i,k}$ denotes the weight of the kth candidate sketch and $\sum_{k=1}^{K} w_{i,k} = 1$.

Now, we get two weight vectors to be optimized: weights of deep representations $\boldsymbol{u}_i$ and weights of sketch patch recon-

struction $\boldsymbol{w}_i$. $\boldsymbol{w}_i$ directly determines the quality of the synthesized sketch patch while $\boldsymbol{u}_i$ determines the representation ability and indirectly influence the reconstruction weights. Similar with the graphical network in the work [Zhou *et al.*, 2012], we design a deep graphical feature learning framework to jointly model the distribution of the weights for deep representations and the distribution of the weights for sketch reconstruction. The joint probability of $\boldsymbol{u}_i$ and $\boldsymbol{w}_i$, $\forall i \in \{1, ..., N\}$, is formulated as:

$$\begin{aligned} &p(\boldsymbol{u}_1, ..., \boldsymbol{u}_N, \boldsymbol{w}_1, ..., \boldsymbol{w}_N) \\ &\propto \prod_{i=1}^{N} \Phi(\boldsymbol{u}_i, \boldsymbol{w}_i) \prod_{(i,j)\in\Xi} \Psi(\boldsymbol{w}_i, \boldsymbol{w}_j) \prod_{i=1}^{N} \Upsilon(\boldsymbol{u}_i), \end{aligned} \tag{3}$$

where $\Phi(\boldsymbol{u}_i, \boldsymbol{w}_i)$ is the local evidence function:

$$\begin{aligned} &\Phi(\boldsymbol{u}_i, \boldsymbol{w}_i) \\ &= \exp\{-\sum_{l=1}^{L} u_{i,l} \|\mathrm{d}_l(\boldsymbol{x}_i) - \sum_{k=1}^{K} w_{i,k} \mathrm{d}_l(\boldsymbol{x}_{i,k})\|^2 / 2\delta_D^2\} \end{aligned} \tag{4}$$

and $\Psi(\boldsymbol{w}_i, \boldsymbol{w}_j)$ is the neighboring compatibility function:

$$\begin{aligned} &\Psi(\boldsymbol{w}_i, \boldsymbol{w}_j) \\ &= \exp\{-\|\sum_{k=1}^{K} w_{i,k} \boldsymbol{o}_{i,k}^{j} - \sum_{k=1}^{K} w_{j,k} \boldsymbol{o}_{j,k}^{i}\|^2 / 2\delta_S^2\} \end{aligned} \tag{5}$$

and $\Upsilon(\boldsymbol{u}_i)$ is the regularization function:

$$\Upsilon(\boldsymbol{u}_i) = \exp\{-\lambda \|\boldsymbol{u}_i\|^2\}. \tag{6}$$

Here $\mathrm{d}_l(\boldsymbol{x}_i)$ means the lth feature map of patch $\boldsymbol{x}_i$ and $\mathrm{d}_l(\boldsymbol{x}_{i,k})$ means the lth feature map of kth candidate patch. $(i, j) \in \Xi$ means the ith and jth patches are neighbors. $\boldsymbol{o}_{i,k}^{j}$ represents the overlapping area of the candidate sketch patch $\boldsymbol{y}_{i,k}$ with the jth patch. λ balances the regularization term with the other two terms.

In order to obtain the optimal weights $\boldsymbol{u}_i$ and $\boldsymbol{w}_i$, the joint probability in (3) should be maximized. We detail the optimization strategy in Section 3.3.

3.3 Optimization

The problem of maximizing the posteriori probability (3) can be converted to minimizing the following problem:

$$\begin{aligned} &\min_{\boldsymbol{u},\boldsymbol{w}} \frac{1}{2\delta_S^2} \sum_{(i,j)\in\Xi} \|\sum_{k=1}^{K} w_{i,k} \boldsymbol{o}_{i,k}^{j} - \sum_{k=1}^{K} w_{j,k} \boldsymbol{o}_{j,k}^{i}\|^2 \\ &+ \frac{1}{2\delta_D^2} \sum_{i=1}^{N} \sum_{l=1}^{L} u_{i,l} \|\mathrm{d}_l(\boldsymbol{x}_i) - \sum_{k=1}^{K} w_{i,k} \mathrm{d}_l(\boldsymbol{x}_{i,k})\|^2 \\ &+ \sum_{i=1}^{N} \lambda \|\boldsymbol{u}_i\|^2 \\ s.t. &\sum_{k=1}^{k} w_{i,k} = 1, 0 \le w_{i,k} \le 1 \\ &\sum_{l=1}^{L} u_{i,l} = 1, 0 \le u_{i,l} \le 1 \end{aligned} \tag{7}$$

where $i = 1, 2, ..., N$, $k = 1, 2, ..., K$, $l = 1, 2, ..., L$.

An alternating optimization strategy can be used to solve this problem.

1)Fix $\boldsymbol{u}$, the problem (6) can be simplified to:

$$\begin{aligned} &\min_{\boldsymbol{w}} \sum_{i=1}^{N} \|\mathrm{D}(\boldsymbol{x}_i) - \sum_{k=1}^{K} w_{i,k} \mathrm{D}(\boldsymbol{x}_{i,k})\|^2 \\ &+ \alpha \sum_{(i,j)\in\Xi} \|\sum_{k=1}^{K} w_{i,k} \boldsymbol{o}_{i,k}^{j} - \sum_{k=1}^{K} w_{j,k} \boldsymbol{o}_{j,k}^{i}\|^2 \end{aligned} \tag{8}$$

where $\alpha = \delta_D^2 / \delta_S^2$. This problem can be solved using the CDM proposed in [Zhou *et al.*, 2012].

2) Fix $\boldsymbol{w}$, the problem (6) can be simplified to:

$$\begin{aligned} &\min_{\boldsymbol{u}_i} \sum_{l=1}^{L} u_{i,l} \boldsymbol{p}_{i,l} + \lambda \|\boldsymbol{u}_i\|^2 \\ &\Rightarrow \min_{\boldsymbol{u}_i} \lambda \boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{u}_i + \boldsymbol{p}_i^{\mathrm{T}} \boldsymbol{u}_i \end{aligned} \tag{9}$$

where

$$\boldsymbol{p}_{i,l} = \frac{1}{2\delta_D^2} \|\mathrm{d}_l(\boldsymbol{x}_i) - \sum_{k=1}^{K} w_{i,k} \mathrm{d}_l(\boldsymbol{x}_{i,k})\|^2 \tag{10}$$

The problem (9) is a standard convex QP problem which can be effectively solved.

The initial $u_{i,l}$ are set to $1/L$ and $w_{i,k}$ are set to $1/K$. Then we alternate execute 1) and 2) until convergence. Once we get the optimal weights w_i, the target sketch patch can be synthesized by (2). After obtaining all target sketch patches $\{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N\}$, they are stitched into a whole sketch with overlapping area averaged. We summarize the proposed DGFL algorithm in Algorithm 1.

Algorithm 1 Deep Graphical Feature Learning for Face Sketch Synthesis

Input: Deep representations of training photo patches, training sketch patches, test photo patches, deep representations of test photo patches, the number of neighboring patches K, the search region, α, λ;
Initialize: $u_{i,l} = 1/L$, $w_{i,k} = 1/K$;
Step 1: For each test photo patch $\boldsymbol{x}_i$, find its K candidate neighboring patches from the training photo patches within the search region around the location of $\boldsymbol{x}_i$ according to the Euclidean distance of their deep representations;
repeat
  Step 2: Fix $\boldsymbol{u}$, optimize the problem (8) to compute $\boldsymbol{w}$;
  Step 3: Fix $\boldsymbol{w}$, optimize the problem (8) to compute $\boldsymbol{u}$;
until convergence
Step 4: Synthesized all target sketch patches with w. Stitch them into a whole sketch with overlapping area averaged.
Output: The target sketch.

4 Experimental Results and Analysis

In this section, we report the experimental results of the proposed DGFL method quantitatively and qualitatively. We

conducted our experiments on the Chinese University of Hong Kong (CUHK) face sketch database (CUFS) [Tang and Wang, 2009]. The CUFS database consists of face photos from three databases: the CUHK student database [Tang and Wang, 2002] (188 persons), the AR database [Martinez and Benavente, 1998] (123 persons) and the XM2VTS database [K. Messer and J. Luettin, 1999] (295 persons). Persons in the XM2VTS database are different in ages, skins (races) and hair styles. Each person in the database has a face photo and corresponding face sketch. All these face photos and sketches are geometrically aligned relying on three points: two eye centers and the mouth center and they are cropped to the size of 250 $\times$ 200. In the following context, we firstly explain the experimental settings. Then we perform face sketch synthesis on the CUFS database to qualitatively illustrate the superiority of the proposed DGFL method compared with state-of-the-arts. Subsequently, objective image quality assessment and face recognition experiments are conducted to quantitatively illustrate the superiority of the proposed method.

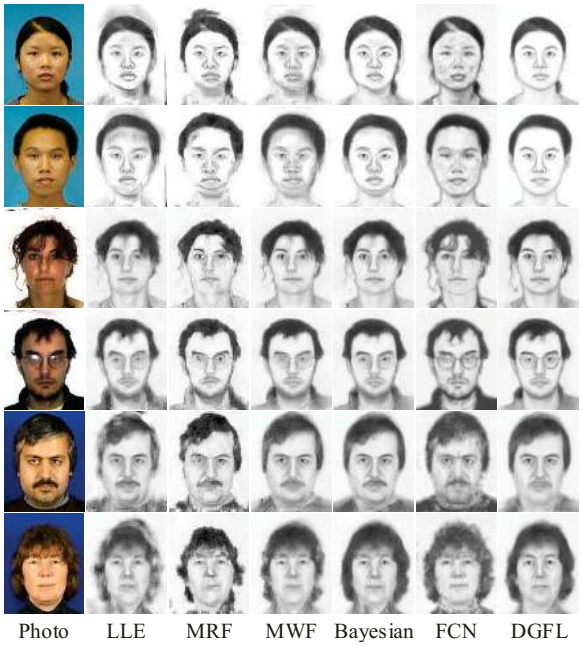

Figure 4: Synthesized sketches on the CUFS database by LLE, MRF, MWF, Bayesian, FCN, the proposed DGFL method.

Table 1: AVERAGE SSIM SCORE (%)

Methods	LLE	MRF	MWF	Bayesian	FCN	DGFL
SSIM (%)	52.58	51.32	53.93	55.43	52.13	**56.45**

4.1 Experimental Settings

The parameters used were set as follows: the image patch size was 10, the overlap size was 5, the size of search region was 5, the number of candidate patches K was set to 10, the α was set to 0.25, the λ was set to 2. For the CUHK student database, 88 pairs of face photo-sketch are taken for training and the rest for testing. For the AR database, we randomly choose 80 pairs for training and the rest 43 pairs for testing. For the XM2VTS database, we randomly choose 100 pairs for training and the rest 195 pairs for testing. Five state-of-the-art methods are compared: the LLE method [Liu *et al.*, 2005], the MRF method [Tang and Wang, 2009], the MWF method [Zhou *et al.*, 2012], the Bayesian method [Wang *et al.*, 2017] and the FCN method [Zhang *et al.*, 2015]. All synthesized sketches by the MWF method and Bayesian Method are generated from the source codes provided by the authors. For the MRF method, we use the codes from the implementation provided by authors of [Song *et al.*, 2014]. Results of the LLE method and FCN method is based on our implementations. All experiments are conducted using Python on Ubuntu 14.04 system with i7-4790 3.6G CPU and 12G NVIDIA Titan X GPU. GPU was only used in the deep feature extraction stage. The deep feature extraction of training photos can be conducted offline. With the help of GPU, deep feature extraction of a input test photo can be finished less than 1s. So, the most time-consuming part of our proposed method lies in two phases: 1) the neighbor selection phase; and 2) the optimization phase. The neighbor selection phase depends on the number of training photos and the size of search regions. More training photos and larger search regions would increase in time-consuming. We use kd-tree [Beis and Lowe, 1997] to conduct K-neighbor searches in order to improve efficiency. The optimization phase mainly depends on the number of iterations. In our experiments, it always converges after five to eight iterations.

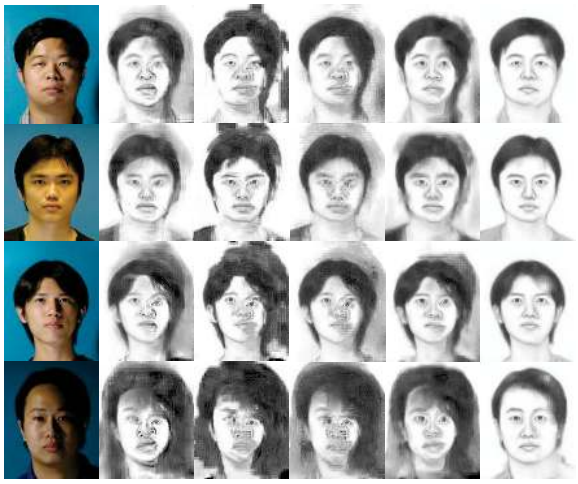

Figure 5: Synthesized sketches of the photos with extreme lighting variance.

4.2 Face Sketch Synthesis

Figure 4 shows some synthesized face sketches from different methods on the CUFS database. As we can see, results from other methods contain significant amount of noise on the nose, mouth and the hair region while the proposed DGFL method achieves much better performance. We have further

validated the performance of the proposed DGFL method on face photos with extreme lighting variance, as shown in Figure 5. When the test photos is affected by extreme environmental noises, such as strong lighting variance, other methods failed to discern the noises and thus produced distortion results. Deep representations is more robust to these noises than pixel intensity, therefore the proposed DGFL method could synthesize sketches with virtually no distortion.

4.3 Objective Image Quality Assessment

We utilize structural similarity index metric (SSIM) [Wang *et al.*, 2004] to evaluate the quality of synthesized sketches by different methods on CUFS. There are 338 (100 + 43 + 195) synthesized sketches for each method generated from the CUFS database. Figure 6 gives the statistics of SSIM scores on the databases. The horizontal axis labels represent the SSIM score from 0 to 1. The vertical axis means the percentage of synthesized sketch whose SSIM scores are not smaller than the score marked on the horizontal axis. Table I gives the average SSIM score on the CUFS database. It can be seen from Figure 6 and Table I that the proposed DGFL method outperform five other state-of-the-art methods.

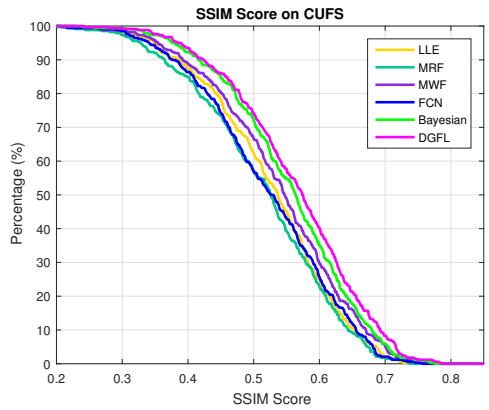

Figure 6: Statistics of SSIM scores on the CUFS database.

4.4 Face Recognition

Sketch based face recognition is always used to assist law enforcement. Given a sketch drawn by the artist, we can retrieve the corresponding photo in the mug shots based on its synthesized sketch. Null-space linear discriminant analysis (NLDA) [Chen *et al.*, 2000] is employed to conduct the face recognition experiments. As aforementioned, we have 338 synthesized sketches for each face synthesis method. We randomly choose 150 synthesized sketches and their corresponding sketches drawn by the artist for classifier training and the rest 188 synthesized sketches are as the gallery set. We repeat each face recognition experiment 20 times by randomly partition the data. Figure 7 gives the face recognition accuracy against variations of the number of reduced dimensions by NLDA on the CUFS database. It can be seen that on the CUFS database, the proposed DGFL method has the highest recognition rate. This accuracy can be further improved by using some advanced face recognition methods [Lei *et al.*, 2014] and heterogeneous face recognition methods [Klare *et al.*, 2011].

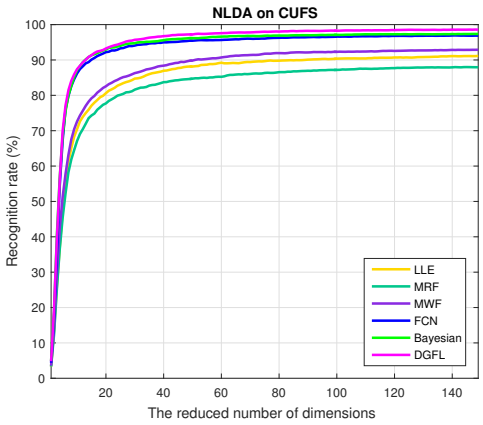

Figure 7: Face recognition accuracy against variations of the number of reduced dimensions by NLDA on the CUFS database.

5 Conclusion

In this paper, we presented a deep feature learning framework for face sketch synthesis. Deep discriminative representations derived from the deep Convolutional Neural Networks are used to represent face photos, which performs more robust respect to the noises than pixel intensity. In order to jointly model the distribution of the weights for deep representations and the distribution of the weights for sketch reconstruction, we present a deep feature learning framework based on the graph model and adopt an alternating optimization strategy to solve this model. Optimal weight combination for sketch reconstruction can be obtained after convergence. Quantitative and qualitative results showed that the proposed method outperforms existing state-of-the-art methods, especially under the noise environment. In the future, we would further explore the discrimination ability of the optimal deep representations and apply it to the identification problem.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (under Grant 61671339, 61501339, 61432014, U1605252, and 61601158), in part by the Fundamental Research Funds for the Central Universities under Grant JB160104, in part by the Program for Changjiang Scholars, in part by the Leading Talent of Technological Innovation of Ten-Thousands Talents Program under Grant CS31117200001, in part by the China Post-Doctoral Science Foundation under Grant 2015M580818 and Grant 2016T90893, and in part by the Shaanxi Province Post-Doctoral Science Foundation.

References

[Beis and Lowe, 1997] Jeffrey S. Beis and David G. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1000–1006, 1997.

[Chen *et al.*, 2000] Li-Fen Chen, Hong-Yuan Mark Liao, Ming-Tat Ko, Ja-Chen Lin, and Gwo-Jong Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33(10):1713–1726, 2000.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[Gao *et al.*, 2012] Xinbo Gao, Nannan Wang, Dacheng Tao, and Xuelong Li. Face sketch-photo synthesis and retrieval using sparse representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(8):1213–1226, 2012.

[Gatys *et al.*, 2016] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances In Neural Information Processing systems*, pages 2672–2680, 2014.

[K. Messer and J. Luettin, 1999] J. Kittler K. Messer, J. Matas and G. Maitre J. Luettin. Xm2vtsdb: the extended m2vts database. In *Proceedings of the International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 72–77, 1999.

[Klare *et al.*, 2011] Brendan Klare, Zhifeng Li, and Anil K. Jain. On matching forensic sketches to mugshot photos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):639–646, 2011.

[Lei *et al.*, 2014] Zhen Lei, Matti Pietikainen, and Stan Z. Li. Learning discriminant face descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):289–302, 2014.

[Liang Chang *et al.*, 2010] Mingquan Zhou Liang Chang, Yanjun Han, and Xiaoming Deng. Face sketch synthesis via sparse representation. In *Proc. 20th Int. Conf. Pattern Recognit*, pages 2146–2149, Istanbul, Turkey, August 2010.

[Liu *et al.*, 2005] Qingshan Liu, Xiaoou Tang, Hongliang Jin, Hanqing Lu, and Songde Ma. A nonlinear approach for face sketch synthesis and recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1005–1010, 2005.

[Liu *et al.*, 2007] Wei Liu, Xiaoou Tang, and Jianzhuang Liu. Bayesian tensor inference for sketch-based facial photo hallucination. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 2141–2146, 2007.

[Martinez and Benavente, 1998] Aleix Martinez and Robert Benavente. The ar face database. In *CVC*, Barcelona, Spain, 1998.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[Song *et al.*, 2014] Yibing Song, Linchao Bao, Qingxiong Yang, and Ming-Hsuan Yang. Real-time exemplar-based face sketch synthesis. In *Proceedings of Eureopean Conference on Computer Vision*, pages 800–813, 2014.

[Tang and Wang, 2002] Xiaogang Tang and Xiaoou Wang. Face photo recognition using sketch. In *Proceedings of IEEE International Conference on Image Processing*, pages 257–260, 2002.

[Tang and Wang, 2003] Xiaoou Tang and Xiaogang Wang. Face sketch synthesis and recognition. In *Proceedings of IEEE International Conference on Computer Vision*, pages 687–694, 2003.

[Tang and Wang, 2009] Xiaoou Tang and Xiaogang Wang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, November 2009.

[Wang *et al.*, 2004] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2014] Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. A comprehensive survey to face hallucination. *International Journal of Computer Vision*, 106(1):9–30, January 2014.

[Wang *et al.*, 2017] Nannan Wang, Xinbo Gao, Leiyu Sun, and Jie Li. Bayesian face sketch synthesis. *IEEE Transactions on Image Processing*, PP:1–11, January 2017.

[Zhang *et al.*, 2015] Liliang Zhang, Liang Lin, Xian Wu, Shengyong Ding, and Lei Zhang. End-to-end photo-sketch generation via fully convolutional representation learning. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 627–634, 2015.

[Zhou *et al.*, 2012] Hao Zhou, Zhanghui Kuang, and Kwan-Yee K. Wong. Markov weight fields for face sketch synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1091–1097, 2012.

[Zhu and Wang, 2016] Mingrui Zhu and Nannan Wang. A simple and fast method for face sketch synthesis. In *International Conference on Internet Multimedia Computing and Service*, pages 168–171, 2016.