The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Face Sketch Synthesis from Coarse to Fine

Mingjin Zhang,[1] Nannan Wang,[1]* Yunsong Li,[1] Ruxin Wang,[2] Xinbo Gao[3]
[1] State Key Laboratory of Integrated Services Networks,
School of Telecommunications, Xidian University, Xi'an 710071, China
[2] Yunnan Union Vision Innovations Technology Company Limited, Kunming 650000, China
[3] School of Electronic Engineering, Xidian University, Xi'an 710071, China

Abstract

Synthesizing fine face sketches from photos is a valuable yet challenging problem in digital entertainment. Face sketches synthesized by conventional methods usually exhibit coarse structures of faces, whereas fine details are lost especially on some critical facial components. In this paper, by imitating the coarse-to-fine drawing process of artists, we propose a novel face sketch synthesis framework consisting of a coarse stage and a fine stage. In the coarse stage, a mapping relationship between face photos and sketches is learned via the convolutional neural network. It ensures that the synthesized sketches keep coarse structures of faces. Given the test photo and the coarse synthesized sketch, a probabilistic graphic model is designed to synthesize the delicate face sketch which has fine and critical details. Experimental results on public face sketch databases illustrate that our proposed framework outperforms the state-of-the-art methods in both quantitive and visual comparisons.

Introduction

Face sketch synthesis from photos is a popular and user-desired function in many applications such as on blogs and social media websites (*e.g.*, Twitter, Instagram, Facebook, and LinkedIn), where users would like to use vivid face sketches as their profile pictures and share their delicate face sketches with friends and family. Face sketch synthesis techniques enable users to effortlessly get sketch faces immediately after taking a photo or selecting a photo from gallery. Delicate and distinctive face sketches synthesized by the these techniques help users stand out and build their "personal brands". Thus, the requirement of fine details in the synthesized sketches is put forward.

However, existing methods fail to synthesize the face sketches with fine and delicate details especially on important facial components, such as eyebrows, eyes, nose and mouth. Regarding the architectures of the designed models, shallow models focus on how to transfer the common structure of training faces to a test face, while deep models pay more attention on how to keep specific structures of the test photo face. A critical issue is that the fine details of critical facial components in synthesized sketches are overlooked by almost all conventional methods.

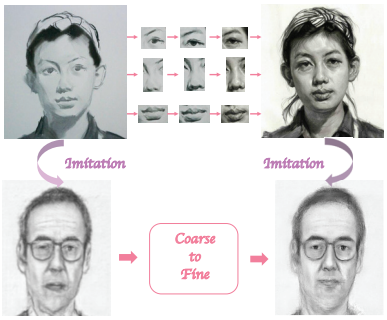


Figure 1: Imitating the coarse-to-fine drawing process of artists. First, we synthesize a coarse sketch which imitates an initial draft containing the coarse structure of face. Then, in the light of the second drawing step where the artist divides the draft into squares and paints the sketch detail by detail, we divide the coarse sketch into patches and synthesize the final fine face sketch.

Targeting on the above problem, we imitate the drawing process in which artists initially draws the coarse structure of face as a draft, and then divides it into squares and paint the face sketch detail by detail (Fig.1). Inspired by this, we propose a coarse-to-fine face sketch synthesis process. The method can be decomposed into two stages: a *coarse* stage and a *fine* stage. In the first stage, we employ the convolutional neural network which is used to synthesize a coarse face sketch. The structure of the synthesized face is approximated, *i.e.*, the spatial positions and sizes of facial components and the part-to-whole relationship are recovered. However, this process may loss or distort details on some facial components, which could play a significant role in specifying the characteristics of face sketches and which enable users to build personal brands. Taking eyes for example, all eyes are balls set in sockets, surrounded by lower and upper lid, but different eyes have different shapes. They are teardrop shaped or almond. In the second stage, inspired by the polishing process, we propose a probabilistic graphic model to recover the missing details and correct or adjust the distortions generated in the first stage. The candidate patches of the test photo and the coarse sketch synthesized in the first

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

stage are regarded as the inputs of the probabilistic graphic model. Bayesian inference are employed to erase the distortions or noises produced in the coarse stage and synthesize fine details especially on distinct edges of facial components.

The main contribution of this work is a face sketch synthesis process which is performed from coarse to fine. The experimental results on a database including 606 face photo-sketch pairs demonstrate the superior performance of our method in terms of image quality assessment compared with state-of-the-art methods. Our synthesized sketches contain more delicate details on facial components and fulfil user requirements in digital entertainment.

The remainder of the paper is organized as follows. Section II presents a review on state-of-the-arts of face sketch synthesis. Section III details the proposed the coarse-to-fine face sketch synthesis framework. Experimental results and comprehensive analyses are presented in Section IV. Section V draws the conclusion.

Related Work

Existing face sketch synthesis methods can be classified into two major categories: shallow learning-based methods and deep leaning-based methods, which are detailed as follows.

Shallow learning-based face sketch synthesis

Shallow learning-based face sketch synthesis generally assumes that face photos and sketches share a common facial structure. The relationship between training photos and the test photo is learned and then applied to the sketches for synthesis. This class of models can be further grouped into three types: subspace learning, Bayesian inference, and sparse representation (Wang et al. 2013).

In the subspace learning-based methods, principal component analysis (Jolliffe 2002)(PCA)-based and local linear embedding (LLE)-based methods are the typical ones. The PCA-based method proposed by Tang and Wang (2002) assumes a linear relationship between the test and training photos. It is also assumed that the photos and sketches share a common topological structure. Thus, in line with PCA, the eigenfaces can be combined linearly to synthesize the face sketches. Compared with the PCA-based method, the LLE-based methods presented by Liu *et al.* (2005) and Liu *et al.* (2007) keep the assumption that the face photos and sketches share common information, and borrow the idea of local linear embedding. These methods are performed on image patches instead of whole images, where a linear relationship among patches is constructed. Song *et al.* (2014) developed a spatial sketch denoising (SSD)-based method which could suppress noise in the synthesized face sketches by Liu et al. Note that the assumption of sharing a common topological structure is too restricted in these subspace learning-based methods, causing that the details which exist in the test sample but not in the training dataset are lost during synthesis.

Bayesian inference-based methods are solved via repeated calculation of the product and sum rule of probability. We shall begin by discussing the embedded hidden Markov model (E-HMM)-based methods. Gao *et al.* (2008a)(2008b)(2010) exploited E-HMM to model the relationship between the training sketches and photos. In the test stage, they utilized the relationship to synthesize a groups of face sketches before fusing them by a selective ensemble manipulation. The other major class of Bayesian inference-based methods are Markov random field (MRF)-based methods. Wang *et al.* (2009) presented the MRF model which enforced soft constraints between the sketch patches and their corresponding photo patches, as well as constrains between the sketch patches and their neighbouring sketch patches. The MRF-based family has many members, including the weighted MRF method improved by Zhou *et al.* (2012), the MRF model based on multiple features (Peng et al. 2016) or super-pixels (Peng et al. 2015), the alternating MRF-based model (Wang et al. 2013)(2017). The alternating MRF-based model results in improved performance by optimizing weights and searching similar candidates alternatively, but the global optimum of the problem is still not guaranteed due to non-convexity and its computational cost is high.

Sparse representation plays a critical role in face sketch synthesis. We begin by considering one typical manipulation which utilizes sparse coding to investigate the relationship of face sketch patches (Chang et al. 2010)(Wang et al. 2011)(Gao et al. 2012). As an illustration of the typical sparse coding-based method, we consider the two-step model proposed by Gao *et al.* (2012). In the first step, the sketch dictionary and the photo dictionary are learned after which the sparse representation of a test sketch is multiplied with the sketch dictionary for initial sketch synthesis. In the second step, the high frequency of the synthesized sketch is produced by support vector regression. Another type of sparse coding-based method is proposed by Zhang *et al.* (2015), which utilizes the sparse codes instead of the pixel intensities to find similar counterparts of face sketch patches. All of the sparse-based methods, no matter what the sparse codes are used for, amount to the restrict similarity assumption. Thus, in general, they are hard to synthesize details appearing only in the test photos but not in training photos.

Deep learning-based face sketch synthesis

Deep learning models are a powerful tool for analyzing complex data. In face synthesis, these models formulate the mapping relationship between face photos and sketches in an end-to-end manner. With the help of parallel computing, the target sketch can be synthesized through a single forward process, leading to fast inference speed.

Zhang *et al.* (2017) investigated a fully convolutional network (FCN) stacked by a series of convolutional layers. In addition, they designed a generative loss for the optimization of FCN, which can describe the person identity of a test photo from training photos. But the synthesized sketches have blurry contours due to the mean square error metric in the training loss. Goodfellow *et al.* (2014) presented the generative adversarial network (GAN) including two convolutional neural networks. The network of generator focus on how to synthesize the face sketches like those drawn by artists, while the other network, namely discriminator net-

work, pays more attentions on how to classify the synthesized sketches and the ones drawn by artists. In this way, the identity information only existing in the test database can be synthesized, but the common structural information of face is easily missing at some time.

Face Sketch Synthesis from Coarse to Fine

Compared with the face sketches drawn by artists, the sketches synthesized by the shallow learning methods may lose characteristics which exist only in the test samples, and the sketches generated by the deep learning models may lose some fine details on the facial components, such as eyebrows, eyes, nose and mouth. To overcome these shortcomings, we interview a number of artists and try to understand their drawing processes. These works inspire us to develop a coarse-to-fine face sketch synthesis process. Fig.2 gives a graphical demonstration to our proposed method. The synthesis process of a face sketch is decomposed into two stage:

- Coarse Stage: it builds the facial structure of a test photo and captures the characteristic features which are exclusive in the test faces and not in the training faces. Given a test photo $\mathbf{x}$, the task of this stage is to estimate the coarse sketch $\mathbf{y}^c$ from $\mathbf{x}$. The inference of $\mathbf{y}^c$ can be formulated as maximizing the posterior probability $P(\mathbf{y}^c|\mathbf{x})$.
- Fine Stage: it erases the noise produced in the coarse stage and recovers the distinctive edges and delicate facial details which are lost or distorted in the coarse stage. The task of this stage is to infer the fine sketch $\mathbf{y}^f$ from the coarse sketch $\mathbf{y}^c$ and test photo $\mathbf{x}$. We formulate this problem as maximizing the posterior probability $P(\mathbf{y}^f|\mathbf{y}^c, \mathbf{x})$.

Coarse stage

After interviewed the artists, we find it is necessary to copy the face sketches drawn by artists for the students when learn drawing. The copy in painting field can help the students be familiar with the basic drawing techniques and the drawing of face sketches based on photos.

Considering this, we train a convolutional neural network which directly learns the mapping relationship between face photos and sketches. This model exploits the potential and identity information of the test faces which is exclusive in the training data.

In particular, GAN (Goodfellow et al. 2014) shows promising performance for synthesizing images with characteristics of the test samples. This model is extended to conditional GAN (Mirza and Osindero 2017) which produces compelling results on super-resolution (Ledig et al. 2017) and image inpainting (Pathak et al. 2016). The conditional GAN is composed of two models: the generator G and the discriminator D. The posterior probability $P(\mathbf{y}^c|\mathbf{x})$ is expressed as:

$$P(\mathbf{y}^c|\mathbf{x}) = G(\mathbf{x}, \mathbf{z}), \tag{1}$$

where $\mathbf{z}$ is a noise term and plays a role as one of the inputs of G.

The generator G aims to synthesize face sketches as real as possible, while the discriminator D tries to distinguish the

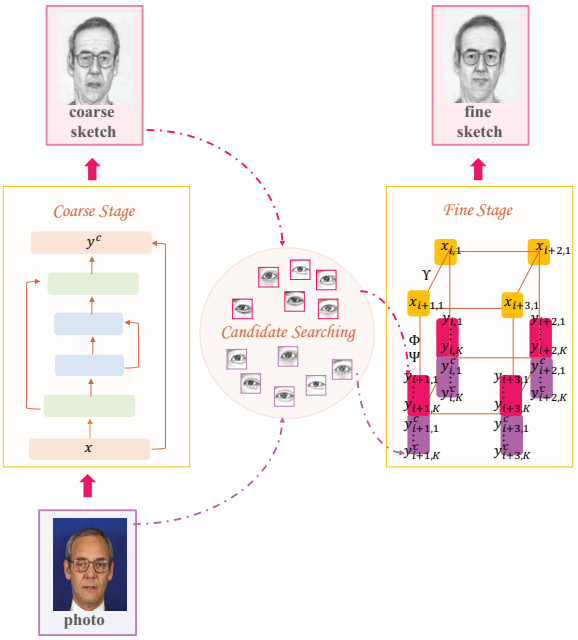


Figure 2: Framework of the proposed face sketch synthesis from coarse to fine. In the coarse stage, we utilize the convolutional neural network in the U-Net to synthesize a coarse sketch. In the fine stage, we investigate a probabilistic graphic model to generate a fine sketch.

sketches synthesized by G from those drawn by artists. The generator G and discriminator D compete with each other and play a two-player min-max game together. Experiences tell us that the L1 loss between $G(\mathbf{x}, \mathbf{z})$ and $\mathbf{y}$ can suppress the pixel-value difference between the sketches synthesized by generator G and those drawn by artists. Thus, the objective function of the whole model in the coarse stage can be written as

$$\begin{aligned}
&\min_G \max_D V(G, D) \\
&= E_{\mathbf{x}' \sim p_{data}(\mathbf{x}'), \mathbf{z} \sim p_z(\mathbf{z})}[log(1 - D(\mathbf{x}', G(\mathbf{x}', \mathbf{z})))] \\
&+ E_{\mathbf{x}', \mathbf{y}' \sim p_{data}(\mathbf{x}', \mathbf{y}'), \mathbf{z} \sim p_z(\mathbf{z})}(\|\mathbf{y}' - G(\mathbf{x}', \mathbf{z})\|_1) \\
&+ E_{\mathbf{x}', \mathbf{y}' \sim p_{data}(\mathbf{x}', \mathbf{y}')}[logD(\mathbf{x}', \mathbf{y}')],
\end{aligned} \tag{2}$$

where the distribution p_{data} is over the training face photo-sketch pairs $(\mathbf{x}', \mathbf{y}')$ and $p_{\mathbf{z}}$ is over the noise $\mathbf{z}$.

Note that the discriminator D regards the pair of the face photo and synthesized sketch as the positive example, while the pair of the face photo and the sketch drawn by artist is regarded as the negative example.

Consider that the input face photo $\mathbf{x}'$ and the output sketch $\mathbf{y}'$ of the generator G describe the same person and share a common facial structure, we design the architecture with mirrored symmetrical layers which can carry common information. The skip connections are built between the mirrored layers, as in the "U-Net"(Ronneberger, Fischer, and Brox 2015).

The generator G and discriminator D are optimized fol-

lowing the objective function in Eq. (2) in an alternating way. We update the discriminator D and the generator G by stochastic gradient ascending and descending, respectively.

Fine stage

In the above stage, the identity information only existing in the test database can be synthesized, but the distinctive and fine details are lost or distorted and the noise exist in the results. Taking the left eye in Fig. 1 as an example, the coarse stage can generate the coarse structures of eyes which, however, are dissimilar to the eyes drawn by artists. It cannot produce the distinct contours and cannot illustrate how much of the exposed eye the iris covers. Thus, it is necessary to synthesize the sketches with fine details via the proposed fine stage.

The drawing process of artists inspires us to regard the coarse sketches generated in the first stage as the drafts drawn by artists. Then it is needed to erase noise from the drafts and refine the details on the facial components before making a closer observation on the faces like artists.

Hence, the task of this fine stage can be defined as maximizing the probability of fine face sketches $\mathbf{y}^f$ given the coarse sketch $\mathbf{y}^c$ and the face photo $\mathbf{x}$. We can express it according to Bayes' theorem in terms of the posterior and prior probability:

$$P(\mathbf{y}^f|\mathbf{y}^c,\mathbf{x}) = \frac{P(\mathbf{y}^f,\mathbf{y}^c,\mathbf{x})}{P(\mathbf{y}^c|\mathbf{x})P(\mathbf{x})}, \tag{3}$$

where the prior probability $P(\mathbf{x})$ is a normalization term and $P(\mathbf{y}^c|\mathbf{x})$ is obtained in the coarse stage, we only need to maximize the joint probability $P(\mathbf{y}^f,\mathbf{y}^c,\mathbf{x})$ in this stage.

Enlightened by artists who divide the draft into squares and draw the face sketch detail by detail, we divide each photo $\mathbf{x}$, fine sketch $\mathbf{y}^f$ and coarse sketch $\mathbf{y}^c$ into N overlapping patches $\{\mathbf{x}_1,...,\mathbf{x}_N\}$, $\{\mathbf{y}^f_1,...,\mathbf{y}^f_N\}$ and $\{\mathbf{y}^c_1,...,\mathbf{y}^c_N\}$, respectively. These overlapping patches replace the images in the joint probability, so that the task is turned to maximize the likelihood function $P(\mathbf{y}^f_1,...,\mathbf{y}^f_N,\mathbf{y}^c_1,...,\mathbf{y}^c_N,\mathbf{x}_1,...,\mathbf{x}_N)$.

The target sketch is a weighted sum of several similar sketches from training data. For each sketch patch $\mathbf{y}^f_i$, we find $2K$ candidate sketch patches $\{\mathbf{y}^f_{i,1},...,\mathbf{y}^f_{i,2K}\}$, where $i = 1,2,...,N$. $\mathbf{y}^f_i$ is expressed in a linear combination form:

$$\mathbf{y}^f_i = \sum_{k=1}^{2K} \omega_{i,k}\mathbf{y}^f_{i,k}, \tag{4}$$

where the K candidate sketch patches come from training sketch patches that are similar to the coarse sketch patches $\mathbf{y}^c_i$ according to the Euclidean distance metric. These patches are denoted as $\{\mathbf{y}^c_{i,1},...,\mathbf{y}^c_{i,K}\}$. The other K candidate sketch patches are corresponding to the K candidate photo patches $\{\mathbf{x}_{i,1},...,\mathbf{x}_{i,K}\}$, which should be similar to the test photo patches $\mathbf{x}_i$. Regarding these settings, the problem is transformed to find the optimal weights ω_i for maximizing the probability $P(\omega_1,...,\omega_N,\mathbf{y}^c_1,...,\mathbf{y}^c_N,\mathbf{x}_1,...,\mathbf{x}_N)$.

Each patch has a close relation with their neighbors, and these relations form the facial structure of different components, such as eyebrow, eye, nose and mouth. Here we employ the probabilistic graphical model to formulate such facial structures. The sketch candidates, photo candidates and test photo patches act the nodes of the proposed probabilistic graphical model. The probabilistic relationship between these nodes are the links of our graph. In the probabilistic graphical model, due to the assumed independence between different $\mathbf{y}^c_i$'s and $\mathbf{x}_i$'s, we decompose the joint distribution $P(\omega_1,...,\omega_N,\mathbf{y}^c_1,...,\mathbf{y}^c_N,\mathbf{x}_1,...,\mathbf{x}_N)$ over all of variables ω_i, $\mathbf{y}^c_i$, and $\mathbf{x}_i$, $i = 1,...,N$, into a product of factors which only depends on a small subset of variables, which is formulated as

$$\begin{aligned}&\max_{\omega_i} P(\omega_1,...,\omega_N,\mathbf{y}^c_1,...,\mathbf{y}^c_N,\mathbf{x}_1,...,\mathbf{x}_N)\\ &\propto \max_{\omega_i} \prod_{i=1}^{N}\Phi(\mathbf{y}^c_i,\omega_i)\prod_{i=1}^{N}\Psi(\mathbf{x}_i,\omega_i)\\ &\qquad \prod_{(i,j)\in\Xi}\Upsilon(\omega_i,\omega_j)\end{aligned} \tag{5}$$

where $(i,j) \in \Xi$ denotes the ith photo patch and the jth neighbor. The linear combination of candidate sketch patches should be similar to the coarse sketch patch. The combination of candidate photo patches should also be close to the test photo patch. The overlapping area of neighbouring candidate sketch or photo patches should have the similar pixel intensities. Thus, the factors are expressed as

$$\Phi(\mathbf{y}^c_i,\omega_i) = \exp\{-\|\mathbf{y}^c_i - \sum_{k=1}^{K}\omega_i\mathbf{y}^c_{i,k}\|^2/2\sigma_C^2\}, \tag{6}$$

$$\Psi(\mathbf{x}_i,\omega_i) = \exp\{-\|\mathbf{x}_i - \sum_{k=1}^{K}\omega_i\mathbf{x}_{i,k}\|^2/2\sigma_D^2\}, \tag{7}$$

and

$$\Upsilon(\omega_i,\omega_j) = \exp\{-\|\sum_{k=1}^{K}\omega_{i,k}\mathbf{r}^j_{i,k} - \sum_{k=1}^{K}\omega_{j,k}\mathbf{r}^i_{j,k}\|^2/2\sigma_S^2\}, \tag{8}$$

where $\omega_{i,k} \geq 0$ and $\sum_{k=1}^{K}\omega_{i,k} = 1$. $\mathbf{r}^j_{i,k}$ and $\mathbf{r}^i_{j,k}$ mean the overlapping area between the ith and jth patches corresponding to the kth candidate photos. $\mathbf{r}^j_{i,k}$ belongs to ith patches and $\mathbf{r}^i_{j,k}$ locates on jth patches.

Maximizing the likelihood function in Eq. (5) is equivalent to minimizing the error function

$$\begin{aligned}\min_{\mathbf{W}} \quad & \alpha\sum_{i=1}^{N}\|\mathbf{y}^c_i - \mathbf{Y}^c_i\mathbf{W}\|^2 + \beta\sum_{i=1}^{N}\|\mathbf{x}_i - \mathbf{X}_i\mathbf{W}\|^2\\ &+ \sum_{(i,j)\in\Xi}\|\mathbf{R}^j_i\mathbf{W} - \mathbf{R}^i_j\mathbf{W}\|^2,\end{aligned} \tag{9}$$

where the balance parameters $\alpha = \sigma_C^2/\sigma_S^2$, and $\beta = \sigma_D^2/\sigma_S^2$. $\mathbf{W}$ is a vector. In particular, the $(k+(i-1)K)$th element of $\mathbf{W}$ is ω_i. $\mathbf{Y}_i^c$,$\mathbf{X}_i$,$\mathbf{R}_i^j$ and $\mathbf{R}_j^i$ are matrices. $\mathbf{y}_{i,k}^c$, $\mathbf{y}_{i,k}$, $\mathbf{r}_{i,k}^j$ and $\mathbf{r}_{j,k}^i$ are in the $(k+(i-1)K)$th column of ω_i. $\mathbf{Y}_i^c$,$\mathbf{X}_i$,$\mathbf{R}_i^j$, and $\mathbf{R}_j^i$, respectively. The remaining elements are zero. Eq. (9) is reformulate as a standard QP problem:

$$\begin{aligned} \min_{\mathbf{W}} \quad & \mathbf{W}^{\mathrm{T}}\mathbf{Q}\mathbf{W} - 2\mathbf{W}^{\mathrm{T}}\mathbf{G} + \mathbf{H} \\ s.t. \quad & \mathbf{A}\mathbf{W} = 1 \\ & \omega_{i,k} \geq 0 \\ & \forall i \in \{1,2,...,N\}, k \in \{1,2,...,K\} \end{aligned} \tag{10}$$

where $\mathbf{A}$ is a matrix, with the elements from $(1+(i-1)K)$ to iK for each ith row being 1 and all others being 0. The quantities are detailed as follows:

$$\begin{aligned} \mathbf{Q} &= \alpha \sum_{i=1}^{N} \mathbf{Y}_i^{c\mathrm{T}} \mathbf{Y}_i^c + \beta \sum_{i=1}^{N} \mathbf{X}_i{}^{\mathrm{T}} \mathbf{X}_i \\ &\quad + \sum_{(i,j)\in\Xi} (\mathbf{R}_i^j - \mathbf{R}_j^i)^{\mathrm{T}} (\mathbf{R}_i^j - \mathbf{R}_j^i), \end{aligned} \tag{11}$$

$$\mathbf{G} = \sum_{i=1}^{N} \mathbf{Y}_i^{c\mathrm{T}} \mathbf{y}_i^c + \sum_{i=1}^{N} \mathbf{X}_i{}^{\mathrm{T}} \mathbf{x}_i, \tag{12}$$

$$\mathbf{H} = \sum_{i=1}^{N} \mathbf{y}_i^{c\mathrm{T}} \mathbf{y}_i^c + \sum_{i=1}^{N} \mathbf{x}_i{}^{\mathrm{T}} \mathbf{x}_i. \tag{13}$$

Since the term $\mathbf{H}$ has no influence on minimizing (10), we can reformulate (10) as:

$$\begin{aligned} \min_{\mathbf{W}} \quad & \mathbf{W}^{\mathrm{T}}\mathbf{Q}\mathbf{W} - 2\mathbf{W}^{\mathrm{T}}\mathbf{G} \\ s.t. \quad & \mathbf{A}\mathbf{W} = 1 \\ & \omega_{i,k} \geq 0 \\ & \forall i \in \{1,2,...,N\}, k \in \{1,2,...,K\}. \end{aligned} \tag{14}$$

Eq. (14) is a standard convex QP problem and can be solved by the cascade decomposition method (Zhou, Kuang, and Wong 2012).

Experimental Results and Analysis

In this section, we conduct multiple experiments to demonstrate the effectiveness of the proposed face sketch synthesis from coarse to fine. The proposed method is compared with the previous synthesis methods both qualitatively and quantitatively. We conduct experiments on the Chinese University of Hong Kong (CUHK) face sketch database (CUFS) (Wang and Tang 2009), which consists of 606 face sketch-photo pairs. Specifically, this databased includes the CUHK student dataset (188 persons, 134 males and 54 females), the AR dataset (123 persons, 70 males and 53 females) (Martinez and Benavente 1998), and the XM2VTS dataset (295 persons, 158 males and 137 females) (Messer et al. 1999). All face photos are taken under well-controlled conditions. The sketches are drawn by artists according to the face photos. Both face sketches and photos are in the size of $250 \times 200 \times 3$.

Algorithm 1 Realization procedure

Input: training photos $\mathbf{x}'$, training sketches $\mathbf{y}'$, test photo $\mathbf{x}$, a well-trained generator G, number of similar patches K, parameters α and β;
Steps:
1. Reconstruct the coarse sketch $\mathbf{y}^c$ according to (1);
2. Divide the training photos $\mathbf{x}'$, training sketches $\mathbf{y}'$, test photo $\mathbf{x}$, and coarse sketch $\mathbf{y}^c$ into patches;
3. For each patch $\mathbf{x}_i$ in $\mathbf{x}$, do:
3.1. Search K similar patches $\{\mathbf{x}_{i,1},...,\mathbf{x}_{i,K}\}$ with the test photo patch $\mathbf{x}_i$ from training photo patches;
3.2. Collect K similar patches $\{\mathbf{y}_{i,1}^c,...,\mathbf{y}_{i,K}^c\}$ with the coarse sketch patches $\mathbf{y}_i^c$ from training sketch patches;
3.3. Reconstruct the fine sketch patch $\mathbf{y}_i^f$ according to (14);
4. Stitch fine sketch patches $\{\mathbf{y}_1^f,...,\mathbf{y}_N^f\}$ into the fine sketch $\mathbf{y}^f$ by averaging the overlapping areas;
Output: fine sketch $\mathbf{y}^f$.

Table 1: SSIM AND VIF VALUES ON CUFS

Comparison Methods	SSIM	VIF
MRF-based method	0.4282	0.0693
MWF-based method	0.4605	0.0786
Bayesian-based method	0.4622	0.0790
FCN-based method	0.4254	0.0707
Coarse stage	0.4118	0.0736
Our method	**0.4718**	**0.0837**

Experimental Settings

In the proposed approach, we set six parameters. The number of candidates K is 10, the patch size is 11, the overlap size is 7, the search region is 5, and the balance parameters α and β are 100 and 1, respectively. The coarse stage is conducted using Torch on Ubuntu 14.04 system with 12G NVIDIA Titan X GPU, whereas the fine stage is tested using Matlab on Window 7 System with i7-4790 3.6G CPU.

Face Sketch Synthesis

For the CUHK student database, we randomly choose 88 sketch-photo pairs to form the training set, and the remaining 100 sketch-photo pairs compose the test set. For the AR database, the training and test set are split sequentially. Specifically, 100 sketch-photo pairs are selected for training, and the remaining 23 sketch-photo pairs are used for testing. This split manipulation is repeated until all the sketch-photo pairs of the AR database are collected as the test data once. 100 sketch-photo pairs are chosen from the XM2VTS database for training, and the remaining 195 sketch-photo pairs are the test data.

Visual comparisons of the proposed coarse-to-fine face sketch synthesis method with the MRF-based method (Wang and Tang 2009), the MWF-based method (Zhou, Kuang, and Wong 2012), the Bayesian-based method (Wang et al. 2017),

Figure 3: Comparison between the proposed method and the conventional methods for synthesizing sketches on CUFS. (a) Input photos. (b)-(e) Results of MRF, MWF, Bayesian, and FCN based method. (f) Coarse sketches of proposed method. (g) Fine sketches of proposed method.

and the FCN-based method (Zhang et al. 2017) are illustrated in Fig.3. As seen, our method produces more delicate details on the facial components. Although the sketches synthesized in the coarse stage have the identical information of the test photos in comparison with the conventional face sketch synthesis methods, there exist some distortions and missing parts in several facial components. The MRF-based method cannot generate the hairstyle well. The reason behind this is that it can choose only one candidate which may be not suited to match the test patch in the hair region. The MWF-based and Bayesian-based methods can compute a patch by local linear combination, but the details of the final results are not distinctive, such as glasses and mouths. The main rationale behind this is that they overlook the mapping relationship between face photos and sketches and lose the identity information which only exists in the test data. The results of the FCN-based method are noisy and unclear due to the use of mean square error as training loss.

We further validate the performance of the proposed method in Fig.4. In the coarse stage, the facial structure with the identity information which is not presented in the training data can be synthesized. In the next stage, we erase the noise in the face and enhance the details on the eyebrows, eyes, nose and mouth. For instance, the coarse structure of eyes can be generated in the first stage, and then the eyes synthesized in the fine stage show how much of the exposed eyes do the iris cover clearly.

Image Quality Assessment

To evaluate the proposed method quantitatively, we utilize the full reference image quality assessment (FR-IQA) (Gao, Tao, and Li 2015). The original sketches drawn by artists are regarded as the reference images, while the synthesized sketches play as the distorted images. The average FR-IQA value is the mean value of all quality values between the reference and distorted images on CUFS. Specifically, we apply the structural similarity index metric (SSIM) (Wang et al. 2004) and the visual information fidelity index (VIF) (Sheikh and Bovik 2006) to evaluate the performance of the sketches synthesized by different face sketch synthesis methods.

We compare our face sketch synthesis method with the MRF-based method (Wang and Tang 2009), the MWF-based method (Zhou, Kuang, and Wong 2012), the Bayesian-based method (Wang et al. 2017), and the FCN-based method (Zhang et al. 2017) on CUFS. The SSIM values and VIF values of the synthesized sketches are shown in Fig. 5. It can be seen that both the SSIM and VIF values of our sketches

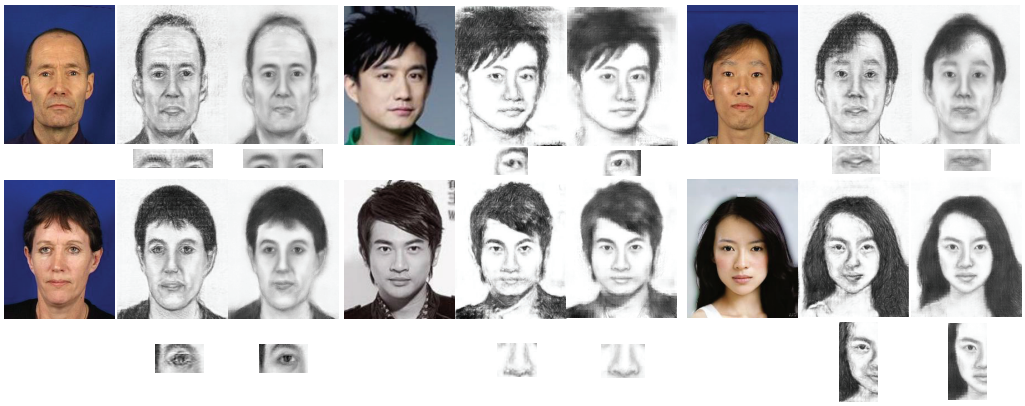
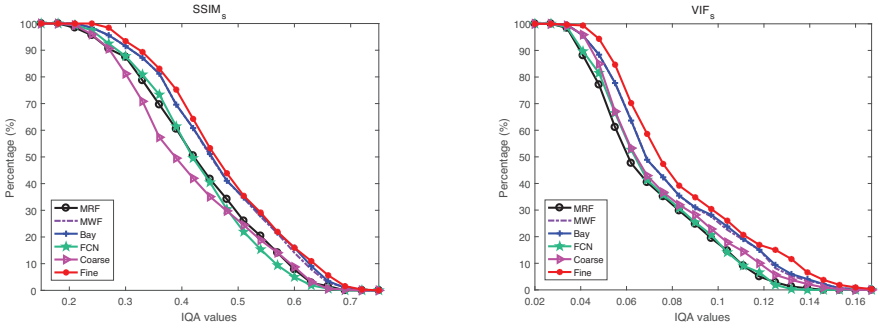


Figure 4: Results of our face sketch synthesis from coarse to fine. (a)(d)(g) Input photos. (b)(e)(h) Coarse sketches of proposed method. (c)(f)(i) Fine sketches of proposed method. The eyebrows, eyes, nose and mouth are refined while the noise in the coarse sketch is erased.



(a) (b)

Figure 5: Comparison of the SSIM (a) and VIF (b) values for synthesizing sketches on CUFS.

are higher than all other competitors. The average SSIM and VIF values of the synthesized sketches are listed in Table 1, which indicates that the coarse-to-fine face sketch synthesis outperforms the state-of-the-arts.

Conclusion

In this paper, we propose a coarse-to-fine face sketch synthesis method imitating the drawing process of artists. The proposed framework is composed of two stages. In coarse stage, the coarse common structure of the face sketch is captured. In fine stage, we begin by erasing the noise of the coarse synthesized sketches, and then, the delicate and distinctive details on the facial components including eyebrows, eyes, nose and mouth are generated. Compared with the state-of-the-arts, the superior performance of our method on CUFS demonstrates the effectiveness of the coarse-to-fine face sketch synthesis. The qualitative results further verify the significance of the fine stage in our method.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (under Grants 61501339, 61772402, 61671339, 61432014, U1605252, 61601158, 61602355, 61301287 and 61301291), in part by Young Elite Scientists Sponsorship Program by CAST (under Grant 2016QNRC001), in part by Natural Science Basic Research Plan in Shaanxi Province of China (under Grant 2017JM6085 and 2017JQ6007), in part by Young Talent fund of University Association for Science and Technology in Shaanxi, China, in part by CCF-Tencent Open Fund (under Grant IAGR 20170103), 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 Grants 2015M580818 and 2016T90893, and in part by the Shaanxi Province Post-Doctoral Science Foundation, in part by the 111 Project (B08038).

References

Chang, M.; Zhou, L.; Han, Y.; and Deng, X. 2010. Face sketch synthesis via sparse representation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2146–2149.

Gao, X.; Zhong, J.; Li, J.; and Tian, C. 2008a. Face sketch synthesis using e-hmm and selective ensemble. *IEEE Trans. Circuits Syst. Video Technol.* 18(4):487–496.

Gao, X.; Zhong, J.; Tao, D.; and Li, J. 2008b. Local face sketch synthesis learning. *Neurocomputing.* 71(10-12):1921–1930.

Gao, X.; Wang, N.; Tao, D.; and Li, X. 2012. Face sketch-photo synthesis and retrieval using sparse representation. *IEEE Trans. Circuits Syst. Video Technol.* 22(8):1213–1226.

Gao, F.; Tao, D.; and Li, X. 2015. Learning to rank for blind image quality assessment. *IEEE Trans. Neural Netw. Learn. Syst.* 26(10):2275–2290.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Proc. Int. Conf. Neural Information Proc. Syst.*, 2672–2680.

Jolliffe, I. 2002. Principal component analysis. *New York, NY, USA:Springer-Verlag*.

Ledig, C.; Theis, L.; Husza, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; and Z., W. 2017. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv Preprint:1609.04802*.

Liu, Q.; Tang, X.; Jin, H.; Lu, H.; and Ma, S. 2005. A nonlinear approach for face sketch synthesis and recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1005–1010.

Liu, W.; Tang, X.; and Liu, J. 2007. Bayesian tensor inference for sketch-based face photo hallucination. In *Proc. IEEE Conf. Artif. Intell.*, 2141–2146.

Martinez, A., and Benavente, R. 1998. The AR face database. *CVC Technical Report* 24.

Messer, K.; Matas, J.; Kittler, J.; and Luettin, J.and Maitre, G. 1999. XM2VTSDB: the extended M2VTS database. In *Proc. Int. Conf. Audio and Video-Based Biometric Person Authentication*, 72–77.

Mirza, M., and Osindero, S. 2017. Conditional generative adversarial nets. *arXiv Preprint:1411.1784*.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2536–2544.

Peng, C.; Gao, X.; Wang, N.; and Li, J. 2015. Superpixel-based face sketch-photo synthesis. *IEEE Trans. Circuits Syst. Video Technol.* 1–12.

Peng, C.; Gao, X.; Wang, N.; Tao, D.; Li, X.; and Li, J. 2016. Multiple representations based face sketch-photo synthesis. *IEEE Trans. Neural Netw. Learn. Syst.* 1–15.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention* 9351(1):234–241.

Sheikh, H., and Bovik, A. 2006. Image information and visual quality. *IEEE Trans. Image Process.* 15(2):430–444.

Song, Y.; Bao, L.; Yang, Q.; and Yang, M. H. 2014. Real-time exemplar-based face sketch synthesis. In *Proc. Eur. Conf. Comput. Vis.*, 800–813.

Tang, X., and Wang, X. 2002. Face photo recognition using sketch. In *Proc. IEEE Int. Conf. Image Process.*, 257–260.

Wang, X., and Tang, X. 2009. Face photo-sketch synthesis and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 31(11):1955–1967.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13(4):600–612.

Wang, N.; Gao, X.; Tao, D.; and Li, X. 2011. Face sketch-photo synthesis under multi-dictionary sparse representation framework. In *Proc. 6th Int. Conf. Image Graph.*, 82–87.

Wang, S.; Zhang, L.; Liang, Y.; and Pan, Q. 2012. Semi-coupled dictionary learning with applications to image super-resolution and photo- sketch synthesis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2216–2223.

Wang, N.; Tao, D.; Gao, X.; Li, X.; and Li, J. 2013. Transductive face photo-sketch synthesis. *IEEE Trans. Neural Netw. Learn. Syst.* 24(9):1364–1376.

Wang, N.; Gao, X.; L., S.; and J., L. 2017. Bayesian face sketch synthesis. *IEEE Trans. Image Process.* 26(3):1264–1274.

Xiao, B.; Gao, X.; Tao, D.; Yuan, Y.; and Li, J. 2010. Photo-sketch synthesis and recognition based on subspace learning. *Neurocomputing.* 73(4-6):840–852.

Zhang, S.; Gao, X.; Wang, N.; Li, J.; and Zhang, M. 2015. Face sketch synthesis via sparse representation-base greedy search. *IEEE Trans. Image Process.* 24(8):2466–2477.

Zhang, L.; Lin, L.; Wu, X.; Ding, S.; and Zhang, L. 2017. End-to end photo-sketch generation via fully convolutional representation learning. *arXiv Preprint:1508.06576*.

Zhou, H.; Kuang, Z.; and Wong, K. 2012. Markov weight fields for face sketch synthesis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1091–1097.