Face Sketch-to-Photo Synthesis from Simple Line Drawing

Yang Liang* , Mingli Song* , Lei Xie† , Jiajun Bu* and Chun Chen*
* Zhejiang Provincial Key Laboratory of Service Robot
College of Computer Science, Zhejiang University, Hangzhou, China
{liangyang,brooksong,bjj,chenc}@zju.edu.cn
† School of Computer Science, Northwestern Polytechnical University, Xi'an, China
lxie@nwpu.edu.cn

***Abstract*—Face sketch-to-photo synthesis has attracted increasing attention in recent years for its useful applications on both digital entertainment and law enforcement. Although great progress has been made, previous methods only work on face sketches with rich textures which are not easily to obtain. In this paper, we propose a robust algorithm for synthesizing a face photo from a simple line drawing that contains only a few lines without any texture. In order to obtain a robust result, firstly, the input sketch is divided into several patches and edge descriptors are extracted from these local input patches. Afterwards, an MRF framework is built based on the divided local patches. Then a series of candidate photo patches are synthesized for each local sketch patch based on a coupled dictionary learned from a set of training data. Finally, the MRF is optimized to get the final estimated photo patches for each input sketch patch and a realistic face photo is synthesized. Experimental results on CUHK database have validated the effectiveness of the proposed method.**

I. Introduction

As a successful application of computer vision in industry, automatic face recognition systems have been widely used in law enforcement [15] and security in recent years. Most traditional face recognition techniques [18][19] focus on face photo-photo matching. However in law enforcement, face photos of suspects are generally unobtainable and a simple sketch drew by an artist according to the description of the witnesses is usually used as a substitute. Direct sketch-photo recognition algorithms achieve low accuracy because of the great differences between sketches and photos. Transforming the sketches and photos to the same modality is an optional solution. The sketches of suspects are often unavailable in police database. So transform from sketches to photos is a better substitute (i.e., sketch-to-photo synthesis).

Several works have been done to address the issues of synthesizing photos from sketches [1][2][3] and sketch-based face recognition [1][2]. By assuming that a face photo and its associated sketch have the same coordinate in the learned subspace, Tang and Wang [1] proposed a global linear model for face photo-sketch synthesis based on principle component analysis (PCA). However, owing to the complicated nonlinear relationship between photo and sketch spaces, such a linear method does not obtain satisfying results, especially when the hair region is included.

Liu et. al. [2] proposed Bayesian Tensor Inference for style transformation between photo images and sketch images of human faces. In their approach the local relations between photos and sketches spaces are extracted by explicitly establishing the connection between two feature spaces formed by a patch-based tensor model. There are three steps in their approach. Firstly, the initial result is obtained by applying the local geometry reserving method [3]. Then, given input image derivatives extracted from the test face sketch, the image derivatives of the face photo is inferred using the Bayesian tensor inference. Finally, gradient correction is conducted to hallucinate the realistic face photo.

Inspired by the idea of embedded hidden Markov model (EHMM) which can learn the nonlinearity of sketch-photo pair with less training samples, Xiao el. at. [4] made use of an EHMM in sketch-to-photo synthesis. The nonlinear relationship of each sketch-photo pair in training set is learnt by a pair of EHMMs. For a given face sketch, several intermediate pseudo-photos are generated based on the selected trained EHMM pairs. And these intermediate pseudo-photos are fused together to obtain the expected pseudo-photo finally. However, the fusion process lacks a global optimal constraint the results usually include mosaics.

Wang and Tang [5] employed a multi-scale Markov random fields (MRFs) model to synthesize face photo-sketch and do recognition in the same modality. They use a multi-scale MRFs model to learn face structures at different scales. Thus local patches in different regions and scales are learnt jointly instead of independently as in [2]. The approach achieved state-of-the-art performance under well controlled conditions. And it significantly reduced the artifacts, such as the blurring and aliasing effects, which commonly exist in the results of previous methods.

Although great progress has been made in recent years as mentioned above, all these methods use the sketches with both contours and shading textures as training and test sketches. An example is showed in Figure 1. Such sketches with rich details often can be drawn only by a professional artist and the artist may need ample time to closely observe the subject. In practice, in a very short time, witnesses can only remember the general outline of suspects, such as having a square face, a

Fig. 1. Examples of a face photo, a sketch with rich detail and a simple line drawing.

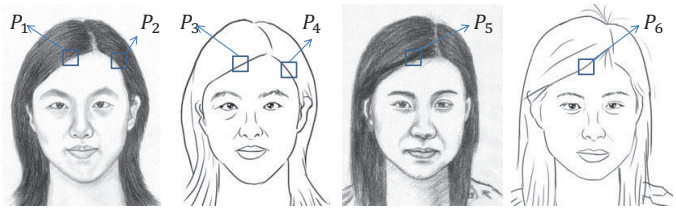

Fig. 2. Example patches for matching measure. The Euclidean distance between P_1 and P_5 is 90.3, the Euclidean distance between P_2 and P_5 is 152.3, the Euclidean distance between P_3 and P_6 is 139.2 and the Euclidean distance between P_4 and P_6 is 142.3

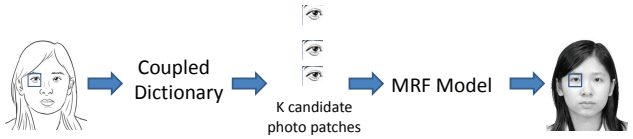

Fig. 3. Illustration of proposed face sketch-to-photo synthesis.

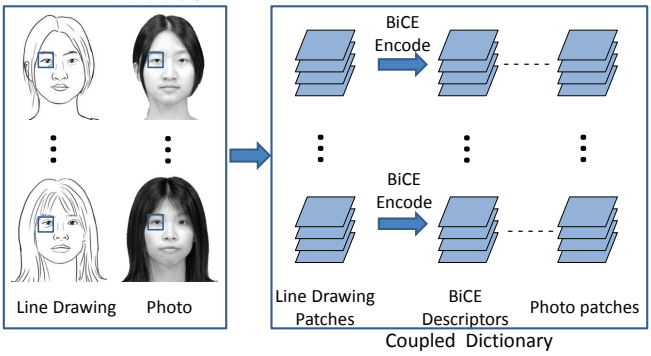

Fig. 4. Illustration of coupled dictionary learning from training photo and simple line drawing set.

big nose. Therefore, only simple line drawing can be obtained in the real situation, such as the example in Figure 1.

If we use the simple line drawings as the training and input test sketches, one problem is met. The previous sketch-to-photo synthesis methods use the Euclidean distance between intensities of two sketches or two sketch patches as the matching measure. Compared with sketches, simple line drawings only contain a few of lines and don't have shading textures. So when the line in drawing moves a little, the distance between the two patches is so large that may be not regard as the similar ones, such as the example patches in Figure 2. The patch P_1 is similar to the patch P_5 in appearance while the patch P_2 is greatly different to the patch P_5. The relation between corresponding patches P_3, P_4 and P_6 in simple line drawings is same to the patches P_1, P_2 and P_5. The Euclidean distance between P_2 and P_5 is much bigger than the Euclidean distance between P_1 and P_5. However, the Euclidean distance between P_3 and P_6 is close to the Euclidean distance between P_4 and P_6. So the matching measure is not suitable for simple line drawing any more.

In this paper, an MRF framework is used to synthesis face photo from a simple line drawing. Wang and Tang [5] have proved that the MRF framework is effective for sketch-to-photo synthesis which can well synthesis complicated face structures, such as hair and significantly reduces artifacts such as the blurring and aliasing effects. Take account of the special characteristic of the simple line drawing, improvements have been made on the two following points:

1) For each simple line drawing patch, an edge descriptor is extracted and the Euclidean distance between edge descriptors of two patches are used as the matching measure instead of the intensities of the patches.
2) A learning based candidate photo patch generation is imported. Wang and Tang [5] find K patches best matching the input sketch patch in appearance from the training set and use their corresponding photo patches as candidates for estimation of the synthesized photo patch corresponding to input sketch patch. This approach may lead to limit the search space for input sketch patch and fail if the test sketch is very different from the sketches in training data. Here we introduce a learning based candidate photo patch generation which can extend the candidate search space.

The remainder of the paper is organized as follows: Section 2 introduces the problem statement and overview of our approach. Section 3 describes the detail of the approach including patch matching and candidate photo patch generation. In Section 4, comprehensive experiment results and comparison are given. Finally, conclusions are made in Section 5.

II. Problem Statement and Overview

Compared with the sketch, simple line drawings are less expressive without any shading textures. Synthesis from simple line drawing to realistic photo is a particularly challenging work. For a successful face sketch-to-photo synthesis, the synthesized face photo should preserve identity characteristics and avoid mosaics and blurs. Coupled dictionary learning is carried out to preserve the identity characteristics of input simple line drawing which captures the nonlinear relationship between simple line drawing patches and photo patches. In order to avoid mosaics and blurs, a MRF model is used to keep the neighboring patches compatible and smooth.

A graphical illustration of our MRF framework is shown in Figure 3, and the detail of coupled dictionary is shown in Figure 4.

The input is a simple line drawing S and the output is a synthesized photo P.

1) Divide the simple line drawings and photos in training data into overlapping patches with the same size, and encode each patch using Binary Coherent Edge descriptors(BiCE) descriptor [6] which will be introduced in next section. A coupled dictionary $\{D^s, D^p\}$ is learned by concatenating the encoded BiCE descriptors and their corresponding photos.
2) Given a simple line drawing S, it is divided into N overlapping patches S_i with the same size of training patches. For each patch, a series of candidate photo patches $\{P_i^l\}_{l=1}^K$ are synthesized based on the learned coupled dictionary.
3) Build a MRF network based on the local input patches and conduct belief propagation to optimize the MRF.
4) Synthsize the face photo P based on the estimated sketch patches.

Compared with the MRF model in [5], there are two major differences. One is that we encode each patch of input simple line drawing using the BiCE descriptors [6] instead of using intensities directly. The other is that we synthesize a series of candidate photo patches for each input local patches based on a coupled dictionary. Details will be explained in the following sections.

III. Face sketch-to-photo synthesis from simple line drawing

In this section, we describe the detail of our algorithm for face photo synthesis, including BiCE descriptor, candidate photo patch generation and the implementation details of the MRF.

A. BiCE descriptor

BiCE descriptor [6] is designed to encode the histogram of edge locations, orientations and local linear lengths. Unlike previous works, the descriptor doesn't use the relative gradient magnitudes. It achieves state-of-the-art results with significant increases in accuracy over traditional approaches such as SIFT [7], GLOH [8] and variants of Daisy descriptors [9] [10]. The simple line drawing patches only contain lines. Hence BiCE descriptor is applicable to describe the characteristic of the simple line drawing.

Since the simple line drawings are typically less dense than natural images, a low dimensional version of BiCE is used here. Just like [11], a three dimensional histogram is defined, with 4 discrete edge orientations, 18 positions perpendicular to the edge, and 6 positions tangent to the edge. And the histogram is binaried by assigning a value of 1 to the top 20% of the bins, and 0 to others. For each simple line drawing patch S_i, the final descriptor of the patch is a vector containing 432 binary bits, hereinafter termed D_i.

B. Candidate photo patch generation

In order to optimize the MRF, for each input sketch patch S_i, K photos patches $\{P_i^l\}_{l=1}^K$ are need to be collected as candidates for the possible states.

In [5], Wang and Tang assume that if a patch S_j found on a training sketch is similar to the patch S_i on the input sketch in appearance, the photo patch P_j corresponding to S_j is considered as one of the good candidates for S_i. For each input sketch patch S_i, they select K patches best matching S_i in appearance from the training data, and use the K photo patches corresponding to the K selected sketch patches as candidates for the possible status of S_i. One candidate sketch patch is chose as the estimate finally. Although this can reduce blurring effect, the search space is limited. It may fails for the test sketch which is great different with the training samples.

In this paper, a learning based candidate photo patch generation is proposed to synthesis K patch photos for each input simple line drawing patch. The sampled facial patches have a certain local nonlinearity and sparsity. The locality is essential for linear extraction from high-dimensional manifold as mentioned in LCC algorithm [12] and a fast implementation of LCC called Locality-constrained Linear Coding(LLC) is proposed by Wang [13], which has an analytical solution. Hence, LLC is employed to capture the nonlinear relationship between simple line drawing patches and photo patches.

Let X be a set of D-dimensional local patches extracted from an image, i.e. $X = [x_1, x_2, \cdots, x_N] \in \mathbb{R}^{D\times N}$. Given a dictionary with M bases, $B = [B_1, B_2, \cdots, B_M] \in \mathbb{R}^{D\times M}$, local-constrained linear coding finds a best coding $c \in \mathbb{R}^{M\times N}$ for a sample x that minimizes the reconstruction error and the violation of the locality constraint. This process can be formulated as optimize the following objective function:

$$\begin{gathered}\min_c \left|\left|x - Bc\right|\right|^2 + \lambda \sum_{i=1}^M Dist_i * c_i, \\ s.t. \sum_{i=1}^M c_i = 1\end{gathered} \tag{1}$$

where $Dist_i = \exp\left(\frac{\left|\left|x-B_i\right|\right|^2}{\sigma}\right)$ and it is the locality adaptor that gives different freedom for each basis vector proportional to its similarity to the input descriptor x. Wang proposes an analytical solution in [13] as follows:

$$c^* = Norm\Big(C_i + \lambda * diag\big(\exp(\frac{\left|\left|x - B_i^c\right|\right|^2}{\sigma})\big)\Big)$$

where $C_i = (B - 1x_i)(B - 1x_i)^T$ denotes the data covariance matrix.

In (1), the dictionary B is assumed to be known. Given a set of training samples, we use the LLC coding criteria to learn a dictionary that is adapted to the distribution of the samples. An obvious approach is to minimize the summed objective functions of all data samples over X and c simultaneously as following:

$$\arg\min_{C,B} \left|\left|x_i - Bc_i\right|\right|^2 + \lambda \sum_{j=1}^M Dist_i * c_i^j \tag{2}$$

where c_i is the coding coefficient corresponding to the ith sample x_i. The optimization algorithm for (2) can be found in [13].

Here we concatenate each BiCE descriptor of the training simple line drawing patch and photo patch pair together as x_i in (2). By optimizing (2), we can obtain a coupled dictionary B. Then each dictionary entry B_i is separated as the photo's one B_i^p and the sketch's one B_i^s, forming B_p and B_s.

Assuming that small patches in the sketch images and its associated photo images have the same coordinate in the learned subspace, we can synthesize the photo patches using the coefficients of the input simple line drawing patch in the learned subspace with respect to the photo dictionary B_p. For a BiCE descriptor D_i of input simple line drawing patch S_i, the coefficients c with respect to B_s can be obtained by the equation (1). Then the corresponding synthesized photo patch P_i^* can be obtained as

$$P_i^* = \sum B_p * c. \tag{3}$$

If we simple choose the P_i^* as the estimate of the photo patch for input simple line drawing patch x_i, the synthesized photo image maybe obtain mosaic effect. Inspired by [20], we relax the (1) by adding constraint on the objective function instead of minimizing the objective function. Then (1) is rewritten as

$$\begin{gathered} \left|\left|x - Bc\right|\right|^2 + \lambda \sum_{i=1}^{M} Dist_i * c_i < \delta, \\ s.t. \sum_{i=1}^{M} c_i = 1. \end{gathered} \tag{4}$$

So we can choose K approximate coefficients c for the input simple line drawing patch x_i, and K corresponding candidate photo patches can be generated by (3).

C. Implementation details

Given a simple line drawing S, we firstly divide the sketch into N patches S_i. Each node on the network is a sketch patch. Let S_j and P_j be the input sketch patch and the photo patch to be estimated at face patch j. For each input sketch patch S_j, the BiCE descriptor D_j is extracted. Then we code the descriptor using the learned couple dictionary and get the coefficients c. So the approximation of D_j can be formulated as follows,

$$D_j^* = \sum B_s * c \tag{5}$$

and the local evidence for S_j can be computed as follows:

$$E_L^j = \exp\{-\left|\left|D_j^* - D_j\right|\right|^2 / 2\sigma_e^2\}. \tag{6}$$

The approximate candidates of the synthesized photo patches can be obtained by replacing the sketch dictionary B_s with the corresponding photo dictionary B_p,

$$P_j^* = \sum B_p * c. \tag{7}$$

With the candidate photo patch P_j^*, the neighboring compatibility function is computed as

$$E_C^j = \exp\{-\left|\left|\partial P_j^* - \partial P_j\right|\right|^2 / 2\sigma_c^2\} \tag{8}$$

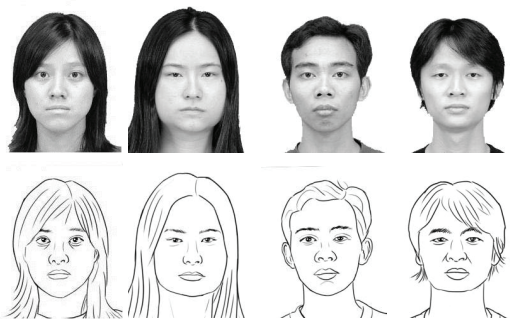

Fig. 5. Examples of face photos and simple line drawings from the CUHK database.

where ∂P_j^* is the overlapped region of the candidate photo patch and ∂P_j is the overlapped region of the neighborhood patches.

So the joint probability of the input sketch and its photo can be written as:

$$p(S_1, \cdots, S_N, P_1, \cdots, P_N) = \prod_j E_L^j \prod_j E_C^j. \tag{9}$$

We conduct belief propagation to optimize the MRF , and the detail description of the Belief propagation can be found in [14][15]. Based on the estimated photo patches selected by MRF, a photo image is stitched for the input simple line drawing.

IV. Experimental results

In order to validate the effectiveness of the proposed method, a simple line sketch-photo database is built. The face photos of the database are from CUHK face sketch database, which includes 188 faces from the CUHK student database. For each face photo in the database, a simple line drawing is drawn. 88 faces are selected for testing and the remaining 100 faces are selected for testing. The size of the face photos and simple line drawings is 200×250. Some examples are shown in Figure 5.

In our experiments, the size of the local image patch is set to 20×20 with 8 pixels overlapped. And the λ in equation (1) and (2) is set as 0.6. We compare our method with the method in [5] which achieves state-of-the-art performance.

The results are shown in Figure 6. From the results, we can find that the photos synthesized by the proposed method are acceptability and are similar to the group truth, while the photos synthesized by [5] are not smooth, especially the hair region. Since the simple line drawing only contains a few lines, the matching measure using the Euclidean distance between the intensities of the patches in [5] is not suitable. So the candidate photo patches selected from data set are not proper and the final results are not smooth. We modify the method in [5] to employ BiCE descriptor and use the Euclidean distance between the BiCE descriptor of the patches as the matching measure. The results of the modified method are listed on the third column, which are close to our results. Hence the BiCE

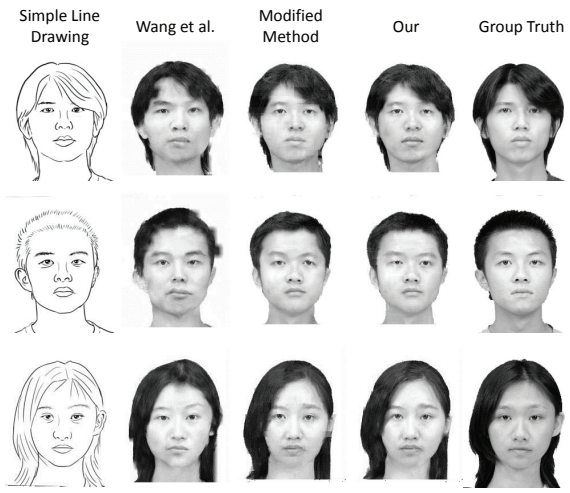

Fig. 6. Comparison between Wang and Tang [5], modified method of [5] and our approach.

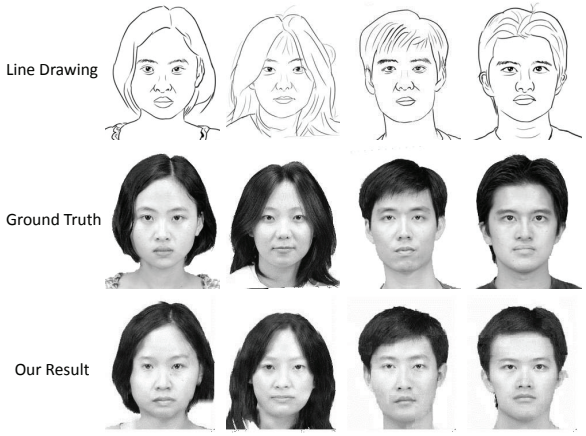

Fig. 7. More results from the proposed approach.

descriptor is essential for synthesizing the face photo from the simple line drawing.

To further validate the proposed approach, we show some more results from our approach in Figure 7.

V. Conclusions

In order to synthesize a face photo from a simple line drawing, we proposed a sketch-to-photo synthesis framework based on coupled dictionary learning and MRF. Take account of the special characteristic of the simple line drawing, BiCE descriptor is employed to encode the local patches of the simple line drawings. Firstly, a coupled dictionary of the encoded BiCE descriptors and their corresponding photos is learned by local-constrained linear coding. Then given a simple line drawing, it is divided into N overlapping patches. And for each local patches, K candidate photo patches are synthesized by the coupled dictionary. Afterward, a multi MRF is built to estimate the photo patches for the input simple line drawing patches. Finally, a face photo is synthesized based on the photo patches selected by MRF. Experimental results demonstrate the effectiveness of the proposed approach.

Acknowledgment

This work is supported by National Natural Science Foundation of China(61170142), National Key Technology R&D Program (2011BAG05B04),the Zhejiang Province Key S&T Innovation Group Project (2009R50009), and the Fundamental Research Funds for the Central Universities(2012FZA5017).

References

[1] Xiaoou Tang, Xiaogang Wang, "Face sketch recognition", *IEEE Transactions on Circuits and Systems for Video Technology (Trans. CSVT)*, 2004, 14 (1) 50-57.

[2] Wei Liu, Xiaoou Tang, Jianzhuang Liu, "Bayesian tensor inference for sketch-based facial photo hallucination", *International Conference on Artificial Intelligence (ICAI)*, 2007, pp. 2141-2146.

[3] Qingshan Liu, Xiaoou Tang, Hongliang Jin, Hanqing Lu, SongdeMa, "A nonlinear approach for face sketch synthesis and recognition", *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 1005-1010.

[4] Bing Xiao, Xinbo Gao, Dacheng Tao, Xuelong Li, "A new approach for face recognition by sketches in photos", *Signal Process*, 2009, 89 (8) 1576-1588.

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

[6] C. Lawrence Zitnick, "Binary Coherent Edge Descriptors", *European Conference on Computer Vision (ECCV)*, 2010 pages(170–182).

[7] Lowe, D.G., "Distinctive image features from scale-invariant keypoints". *International Journal of Computer Vision (IJCV)*, 2004, 60 91-110.

[8] Mikolajczyk, K., Schmid, C., "A performance evaluation of local descriptors". *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2005, 27 1615-1630.

[9] Winder, S.A.J., Brown, M., "Learning local image descriptors". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[10] Winder, S., Hua, G., Brown, M., "Picking the best daisy". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , 2009, 178-185.

[11] Yong Jae Lee, C. Lawrence Zitnick, Michael F. Cohen, "ShadowDraw: Real-Time User Guidance for Freehand Drawing", *ACM Transactions on Graphics (TOG)*, 2011.

[12] Kai Yu, Tong Zhang, Yihong Gong,"Nonlinear Learning using Local Coordinate Coding". *In Advances in Neural Information Processing Systems (NIPS)*, 2009, Vol. 22, Pages 2223-2231.

[13] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, Yihong Gong,"Locality-constrained Linear Coding for Image Classification". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, Pages 3306.

[14] J. S. Yedidia, W. T. Freeman, and Y. Weiss. "Understanding Belief Propagation and Its Generalizations". *Exploring Artificial Intelligence in the New Millennium* , 2003.

[15] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. "earning low-level vision". *International Journal of Computer Vision (IJCV)*, 2000, 40:25-27.

[16] Stephen Mancusi, "The Police Composite Sketch", *Human Press*, 2010.

[17] Dacheng Tao, Xuelong Li, Xindong Wu, Stephen J. Maybank,"Geometric mean for subspace selection", *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* , 2009, 31 (2) 260-274.

[18] Ying-hui Wang, Xiao-juan Ning, Chun-xia Yang, Qiong-fang Wang, "A method of illumination compensation for human face image based on quotientimage", *Information Sciences*, 2008, 178 (12) 2705-2721.

[19] Hujun. Yin, Weilin. Huang, "Adaptive nonlinear manifolds and their applications to pattern recognition", *Information Sciences*,2010, 180 (14) 2649-2662.

[20] Mingli Song, Chun Chen, Jiajun Bu, Teng Sha, "Image-based facial sketch-to-photo synthesis via online coupled dictionary learning", *Information Sciences*, 2012.