2011 Sixth International Conference on Image and Graphics

Face Sketch-photo Synthesis under Multi-dictionary Sparse Representation Framework

Nannan Wang, Xinbo Gao
School of Electronic Engineering
Xidian University
Xi'an, P. R. China
wangnannan006@126.com; xbgao@mail.xidian.edu.cn

Dacheng Tao
Faculty of Engineering and Information Technology
University of Technology, Sydney
Sydney, Australia
Dacheng.Tao@uts.edu.au

Xuelong Li
Center for OPTical Imagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics
Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences
Xi'an, P. R. China
xuelong_li@opt.ac.cn

Abstract—**Sketch-photo synthesis is one of the important research issues of heterogeneous image transformation. Some available popular synthesis methods, like locally linear embedding (LLE), usually generate sketches or photos with lower definition and blurred details, which reduces the visual quality and the recognition rate across the heterogeneous images. In order to improve the quality of the synthesized images, a multi-dictionary sparse representation based face sketch-photo synthesis model is constructed. In the proposed model, LLE is used to estimate an initial sketch or photo, while the multi-dictionary sparse representation model is applied to generate the high frequency and detail information. Finally, by linear superimposing, the enhanced face sketch or photo can be obtained. Experimental results show that sketches and photos synthesized by the proposed method have higher definition and much richer detail information resulting in a higher face recognition rate between sketches and photos.**

Keywords-face recognition; multi-dictionary; sketch-photo synthesis; sparse representation

I. INTRODUCTION

With the development of ID authentication technique, face recognition attracts great attentions. Most face recognition techniques focus on face photo-photo matching, however, a face photo is not always available. In case-solving and suspect searching, photos of suspects are generally unobtainable. A sketch is usually used as a substitute, generated from the cooperation of witnesses and an artist. However, as recent studies illustrated [1], [2], direct sketch-photo identification achieves a low accuracy using traditional face recognition algorithms such as eigenfaces. This is due to the great differences between sketches and photos both in the geometry and texture. Thus, an alternative way to perform sketch-based face recognition is transforming sketches and photos into the same modality: either from a sketch to a photo or from a photo to a sketch. Therefore, sketch-photo synthesis is an urgent problem to address.

The research on sketch-photo synthesis is still on its initial stage. In [1], sketches are categorized into two classes: 1) simple line drawings and caricatures, and 2) complex sketches studied in this manuscript. The algorithms [3]-[7] generate simple line drawings or caricatures from a photo image. It is easy for human being to identify line drawings and caricatures. Nevertheless, it is really a hard task for a computer to perform face recognition using these simple sketches. Recently, B. Klar et al. have used forensic sketches (drawn by an artist according to the descriptions of a witness rather than a static photo as adopted in this manuscript) to do the matching in mug shot photos, with the help of multi-scale local binary pattern and SIFT feature descriptors [8]. In that paper, it does not refer to sketch-photo synthesis yet only image retrieval.

Recent achievements on sketch-photo synthesis can be mainly grouped into two categories: linear methods [2], [9], [10] and nonlinear methods [1], [11]-[15]. Owing to the fact that the relation between sketches and photos is not a simple linear mapping, these linear methods cannot obtain satisfying results, especially when hair regions are included. Using the idea of locally linear embedding (LLE) [16], a sketch synthesis algorithm is proposed in [11], where global nonlinear relation is approximated by locally linear embedding. Multi-scale Markov random field is applied to describe the neighbor relation in sketch-photo patch pairs in [12] and [13]. Inspired by the idea of embedded hidden Markov model, several approaches are presented in [1], [14] and [15]. These nonlinear methods can generally synthesize higher quality sketches or photos. However, a fixed number (K) of nearest neighbors are selected to construct the corresponding model for synthesis according to K-NN principle, resulting in a low definition and blurring effect for synthesized images.

In order to improve the definition and detail information of the synthesized image, a nonlinear face sketch-photo synthesis algorithm under multi-dictionary sparse representation framework is proposed in this manuscript. The proposed approach is composed of two steps: First,

978-0-7695-4541-7/11 $26.00 © 2011 IEEE
DOI 10.1109/ICIG.2011.112

IEEE computer society

using the idea of LLE, an initial image is generated; Subsequently, a multi-dictionary based sparse representation model is constructed to enhance the quality of the initial obtained image. Since the transformation from a sketch to a photo and that from a photo to a sketch are symmetric and reversible just by switching the roles of sketches and photos, we will introduce face sketch synthesis from a photo as an example in the following contents. The rest of this manuscript is arranged as follows. The proposed sketch-photo synthesis method is introduced in section 2. Some experimental results are shown in section 3 and section 4 concludes this paper.

II. The Proposed Method

According to neurobiologists' study [17], it is sparse for visual neurons responding to visual input. Moreover, high frequency information such as edges and contours attracts much more attention for a visual entry. Hence, high frequency information is assumed to be sparsely distributed in this manuscript. Due to the fact that great differences exist among different parts of a facial image such as hair and jowl, it is reasonable to learn different dictionaries to represent various parts or features of these parts.

Given there are N sketch-photo pairs in the training set, which are denoted as $\{S_1, S_2, \cdots, S_N\}$ and $\{P_1, P_2, \cdots, P_N\}$ respectively. Since the proposed method works at patch-level, each sketch and photo is divided into even patches with some overlapping, resulting in the set of sketch patches $\{S_{11}, \cdots S_{1M}, S_{21}, \cdots S_{NM}\}$ and the set of photo patches $\{P_{11}, \cdots, P_{1M}, P_{21}, \cdots P_{NM}\}$. Afterwards, these patches can be clustered into several classes using a clustering method such as *K*-means approach. It should be noticed that the sketch patch intensity subtracting the mean intensity of this patch is explored as the feature for clustering. Here let $\{C_1, \cdots, C_K\}$ represent K clusters consisting of sketch patches and their corresponding photo patches and $\{v_1, \cdots, v_K\}$ denotes their K cluster centers. K is set to 20 in our experiments.

A. Initial Estimate

To obtain an initial sketch from a photo, we can perform the following steps:

1) Input a testing photo, dividing it into M even areas $\{P^1, \cdots P^M\}$ with some overlapping, where P^i consists of patch intensities from the ith photo patch attached into a column, $i = 1, \ldots, M$;

2) $i = 1$;

3) For photo patch P^i, a feature vector f^i is formed by subtracting mean patch intensity from each intensity of this patch; Determine which cluster is nearest apart from f^i according to the following equation (1);

$$index_i = \arg\{\min_j \left\| f^i - v_j \right\|_2\} \tag{1}$$

Where $index_i$ is the index of the nearest cluster apart from f^i;

4) For P^i, k nearest photo patch neighbors P_l^i, $l = 1, \ldots k$, are found within the $index_i$th cluster based on the following criterion:

$$Dis_{ih} = \| P^i - P_h^i \|_2 \tag{2}$$

where P_h^i is the hth photo patch in $index_i$th cluster; This step can also be substituted by finding k photo patches in the whole photo patch training set around the same position as P^i;

5) Optimize (3), leading to k weights $w_l^i, l = 1, \ldots k$:

$$\varepsilon(w^i) = \left\| P^i - \sum_{l=1}^{k} w_l^i P_l^i \right\|_2^2 \text{ s.t. } \sum_{l=1}^{k} w_l^i = 1 \tag{3}$$

6) With $w_l^i, l = 1, \ldots k$ and k sketch patches S_l^i corresponding to the above k nearest photo patch neighbors, a synthesized sketch patch is calculated as in (4):

$$S_L^i = \sum_{l=1}^{k} w_l^i S_l^i \tag{4}$$

7) $i = i + 1$;

8) Iterate 3)-7) until $i = M + 1$; Fuse these M synthesized patches into an initial sketch S_L with overlapping areas averaged.

B. Multi-dictionary Based Sparse Representation

Suppose that $D \in \mathbb{R}^{u \times e}$ is an overcomplete dictionary consisting of e atoms, where $e > u$ and u is the dimension of each atom in D. A signal $y \in \mathbb{R}^u$ can be decomposed into a linear combination of these e atoms in D, i.e. $y = Dw_0$, where $w_0 \in \mathbb{R}^e$ is a column vector with few nonzero entries. This is the basic idea of sparse representation. Here our sparse representation model is mainly made up of two stages: dictionary learning stage and the synthesis stage.

A joint-learning strategy as in [18] is applied to each cluster for two dictionaries: a sketch patch feature dictionary and a photo patch feature dictionary. For any testing photo patch P^i, sparse representation coefficient vector w_0 is computed by projecting the feature of patch P^i to the photo patch feature dictionary learned from the $index_i$th cluster. After this, the high frequency or detail information of the synthesized sketch results with the help of w_0 and the sketch patch feature dictionary which is also learned from the $index_i$th cluster.

The ruse for extracting features of photo patch and sketch patch is as follows: for every photo patch, the first order and second order derivative (both vertical and horizontal) are linked into a column vector; for each sketch patch, the mean patch intensity subtracted by patch intensities are arranged into a column vector. Next the joint-learning strategy is introduced in the ith cluster as an example. Since the features used in the initial estimate stage and image enhancement stage were different, one can perform clustering algorithm twice for the two stages respectively which could be improve the results further.

However, in order to reduce the time costing, clustering on initial estimate stage is implemented only.

According to the basic idea of sparse representation, the sparse representation coefficient vector should be as sparse as possible. At the same time, the fidelity term should be as small as possible. Thus, the following equation is constructed:

$$D_p^{(i)} = \arg\left\{ \min_{\{D_p^{(i)},C^{(i)}\}} \left\| P^{(i)} - D_p^{(i)} C^{(i)} \right\|_2^2 + \alpha \left\| C^{(i)} \right\|_1 \right\} \tag{5}$$

$$D_s^{(i)} = \arg\left\{ \min_{\{D_s^{(i)},C^{(i)}\}} \left\| S^{(i)} - D_s^{(i)} C^{(i)} \right\|_2^2 + \alpha \left\| C^{(i)} \right\|_1 \right\} \tag{6}$$

where $D_p^{(i)}, D_s^{(i)}$ are the photo patch feature dictionary $D_p^{(i)} \in \mathbb{R}^{d_p \times e_i}$ and sketch patch feature dictionary $D_s^{(i)} \in \mathbb{R}^{d_s \times e_i}$ of the ith cluster, respectively, and e_i is the number of atoms in $D_p^{(i)}$ or $D_s^{(i)}$. $P^{(i)}$ and $S^{(i)}$ denote the matrix of photo patches and sketch patches from the ith cluster, where $P^{(i)} \in \mathbb{R}^{d_p \times n_i}$, $S^{(i)} \in \mathbb{R}^{d_s \times n_i}$, n_i is the number of sketch-photo patch pairs of ith cluster, and d_p and d_s are the dimensions of each atom in $D_p^{(i)}$ and $D_s^{(i)}$ respectively. $C^{(i)}$ is the sparse representation coefficient matrix for the ith cluster and α balances the tradeoff between the fidelity term and the sparse term in (5) and (6). These two equations can be combined into one formula as (7):

$$D^{(i)} = \arg\left\{ \min_{\{D_p^{(i)},D_s^{(i)},C^{(i)}\}} \left\| I^{(i)} - D^{(i)} C^{(i)} \right\|_2^2 + \beta \left\| C^{(i)} \right\|_1 \right\} \tag{7}$$

where $D^{(i)} = \left[\frac{(D_p^{(i)})^T}{\sqrt{d_p}}, \frac{(D_s^{(i)})^T}{\sqrt{d_s}} \right]^T$, $I^{(i)} = \left[\frac{(P^{(i)})^T}{\sqrt{d_p}}, \frac{(S^{(i)})^T}{\sqrt{d_s}} \right]^T$, and $\beta = \alpha \left(\frac{1}{d_p} + \frac{1}{d_s} \right)$. In all of our experiments, β is set to 0.05. The optimization problem in (7) is nonconvex for $D^{(i)}$ and $C^{(i)}$ together, however, it is a convex problem for $D^{(i)}$ or $C^{(i)}$. Thus, it can be addressed by optimizing $D^{(i)}$ and $C^{(i)}$ alternatively.

Implementing above learning process for all K clusters, we obtain K pairs of sketch-photo patch feature dictionary. For any testing photo patch P^i , $D_p^{(index_i)}$ is used to find the sparse representation coefficient vector w_0 :

$$w_0 = \arg\left\{ \min_w \lambda \left\| w \right\|_1 + \left\| D_p^{(index_i)} w - P^i \right\|_2^2 \right\} \tag{8}$$

where λ is set to 0.1 in our experiments. Afterwards, with w_0 and $D_s^{(index_i)}$, the high frequency information can be reached:

$$S_H^i = D_s^{(index_i)} w_0 \tag{9}$$

Adding these detail information to the initial synthesized sketch, the final sketch results.

III. Experimental Results and Analysis

In order to validate the effectiveness of the proposed method, experiments on two databases are executed: one is the Chinese University of Hong Kong (CUHK) student database [12] which is publicly available, and the other is our newly constructed database, namely the VIPSL database. There are 188 sketch-photo pairs in the CUHK student database, with one sketch-photo pair for each person. The VIPSL database contains 200 face photos and 1000 sketches in total, with one face photo and five sketches (drawn by five different artists) for each person. All face photos are collected from some public face databases such as the FERET database [19], FRAV2D database [20], and Indians Face database [21]. All face photos in these two databases are taken in a frontal pose with neural expression. However, faces selected in the VIPSL database are from different backgrounds and races in contrast to the same background and the same race in CUHK student database. Some examples are shown in Fig. 1.

A. Face Sketch-photo Synthesis

In this subsection, the proposed face sketch-photo synthesis algorithm is implemented with several other methods [11] [14] [15]. In all these approaches, images are cropped to 64×64 with the patch size of 9×9 and a 7×7 overlapping. For the method proposed in [11], the number of nearest neighbors is set to 5. Through experiments, the above setting can achieve better results than other settings. For the method [14] and [15], we adopt the same settings as in corresponding literatures. We randomly selected 88 sketch-photo pairs as the training set and the rest as the testing set on the CUHK student database. For the VIPSL database, sketches drawn by the same artist are grouped into a category

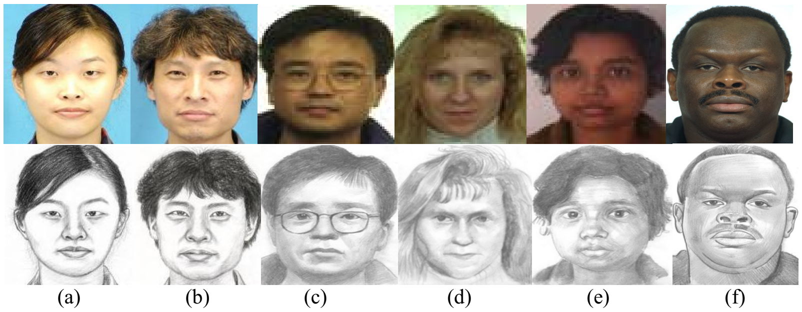
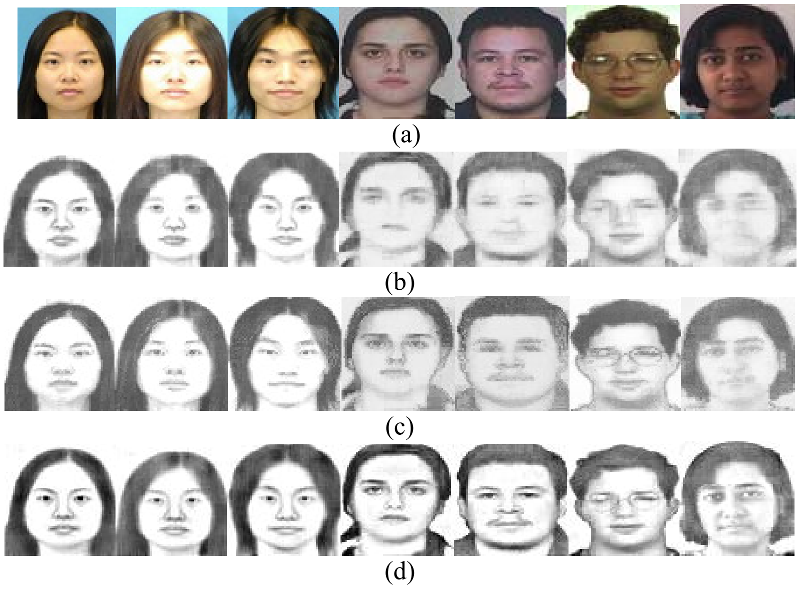


Fig. 1 Examples from CUHK Student database and VIPS database. (a) and (b) are examples from CUHK Student database. (c-f) are examples from VIPS database.



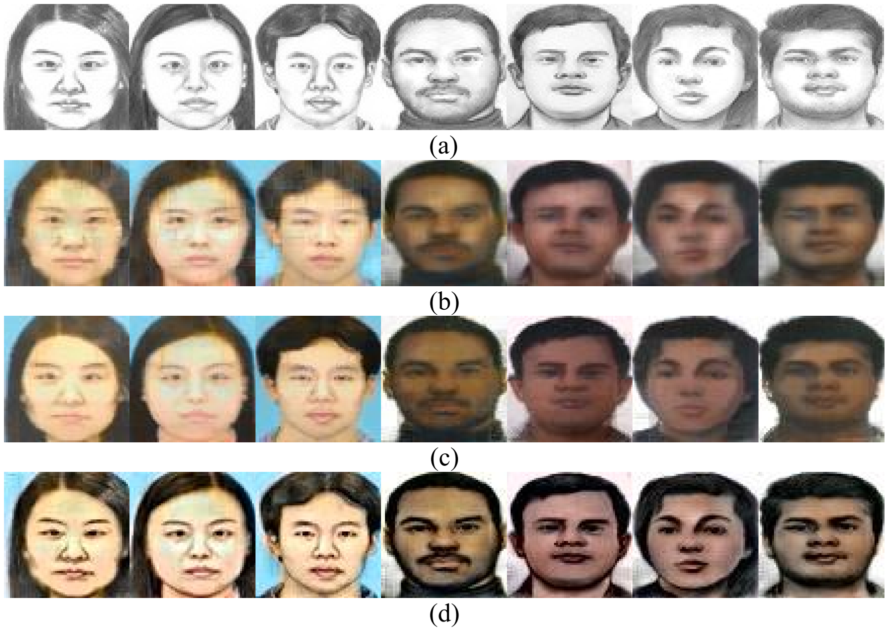
(a)

(b)

(c)

(d)

Fig. 2 The synthesized sketches. (a) Original photos; (b) Sketches generated by [11] (LLE initial estimate); (c) Sketches generated by [14]; (d) Sketches generated by the proposed method.



(a)

(b)

(c)

(d)

Fig. 3 The synthesized photos. (a) Original sketches; (b) Photos generated by [11] (LLE initial estimate); (c) Photos generated by [15]; (d) Photos generated by the proposed method.

with their corresponding face photos. Thus, there are five categories in total. After that, 100 sketch-photo pairs are randomly chosen as the training set and the rest for testing set. Some synthesized sketches and photos are shown in Fig. 2 and Fig. 3 respectively.

It should be noticed that the method in [11] is just the same as the initial estimate in this paper. So, from the comparisons between [11] and the proposed method, the effectiveness of the proposed sparse representation based image enhancement could be illustrated. Furthermore, from Fig. 2 and Fig. 3, we can see that sketches and photos synthesized by the proposed method even have a higher definition and richer detail information than embedded hidden Markov model based methods [14],[15]. Considering time costing, it takes about 2.5 minutes to synthesize a sketch running on a 3 GHZ CPU computer with the sketch size of 64×64.

B. *Face Sketch-photo Recognition*

In order to further illustrate the efficacy of the proposed synthesis method, we perform face recognition [21] using synthesized images (sketches and photos). Two situations

TABLE I. FACE RECOGNITION USING SKETCHES/PHOTOS GENERATED BY DIFFERENT ALGORITHMS ON THE CUHK STUDENT DATABASE

Methods	[11] (LLE-based)	[14]	[15]	The proposed
sketch	99	100	---	**100**
photo	82	---	96	**98**

TABLE II. FACE RECOGNITION USING SKETCHES/PHOTOS GENERATED BY DIFFERENT ALGORITHMS ON THE VIPSL DATABASE

Methods	[11] (LLE-based)	[14]	[15]	The proposed
sketch	89.4	90.6	---	**91.4**
photo	87	---	88.1	**88.5**

TABLE III. MOS VALUES FOR SYNTHESIZED IMAGES BY DIFFERENT ALGORITHMS

Methods	[11] (LLE-based)	[14]	[15]	The proposed
CUHK Student-sketch	2.4	3.3	---	**4.05**
CUHK Student-photo	2.1	---	3.4	**3.9**
VIPS-sketch	1.75	2.7	---	**3.95**
VIPS-photo	1.8	---	2.95	**3.2**

Notice：“---” in above tables means that the algorithm in the corresponding column is not used to synthesize the image in the corresponding row

for face recognition are considered: First, take the original sketch drawn by the artist as the input testing image and retrieve in the synthesized sketch database; Secondly, the synthesized photo are considered as the input testing photo with the photo set retrieved. For the VIPSL database, since there are five categories of sketch-photo pairs, the final recognition rate is calculated by averaging five obtained recognition rates. Experimental results are shown in TABLE I. and TABLE II. respectively. From these two tables, the recognition rate is well reserved and even increases to some extent.

C. Subjective Image Quality Assessment

In the last subsection, face recognition has been applied to indicate the proposed method's validity, where face recognition can be seen as an objective image quality assessment rule. In this subsection, we will assess the synthesized images' quality from the view of subjective. We invite 20 volunteers to score the images generated by [11], [14], and [15]. Here the principle we used is Mean Opinion Score (MOS) which is applied to the field of multimedia service (ITU-T p.910) as a standard. The score ranges from 1 to 5 and is some times of 0.5. The final result is computed as (10):

$$MOS(l)=\frac{1}{20}\sum_{i=1}^{20}A(i,l) \qquad (10)$$

where $A(i,l)$ denotes the score of the lth image from the ith person. The final result is shown in TABLE III.

From this table, we can see that the proposed method achieve a much higher MOS value than the other three methods, which also implies that sketches and photos synthesized by the proposed method is more acceptable for human beings' visual input.

IV. CONCLUSIONS

In order to improve the definition and the richness of detail information, a face sketch-photo synthesis algorithm under multi-dictionary sparse representation framework is presented in this manuscript. The proposed method consists of two stages: First, synthesize an initial image using the idea of LLE; Then, with multi-dictionary sparse representation model, enhance the quality of the initial image obtained by the first stage. Experimental results demonstrate the efficacy of the proposed approach. In the future, we will focus on how to use multi-sketches to improve the retrieval result further. Furthermore, we will try to use idea of image denoising to enhance the quality of the synthesized images.

ACKNOWLEDGMENT

We want to thank the helpful comments and suggestions from the anonymous reviewers. This project is supported by the National Basic Research Program of China (973 Program) (Grant No. 2011CB707100). This research was supported partially by the National Natural Science Foundation of China (Grant No. 60832005, No. 41031064, No. 61072093), the Ph.D. Programs Foundation of Ministry of Education of China under Grant 20090203110002, the Public Science and Technology Research Funds Projects of Ocean, State Oceanic Administration of the P. R. China under Grant 201005017, the Key Science and Technology Program of Shaanxi Province of China under Grant 2010K06-12, the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2009JM8004, and the Open Project Foundation of State Key Laboratory of Industrial Control Technology(Zhejiang University, Grant No. ICT1105).

REFERENCES

[1] X. Gao, J. Zhong, J. Li, and C. Tian, “Face Sketch Synthesis Algorithm Based on E-HMM and Selective Ensemble,” *IEEE Trans. Circuits Syst. Video Technol.*, April, 2008, vol. 18, no. 4, pp. 487-496.

[2] X. Tang and X. Wang, "Face Sketch Recognition," *IEEE Trans. Circuit Syst. Video Technol.*, January, 2004, vol. 14, no. 1, pp. 50-57.

[3] S. Brennan, "Caricature Generator", *M.S. thesis*, MIT, Cambridge, MA, 1982.

[4] M. Tominaga, S. Fukuoka, K. Murakami, and H. Koshimizu, "Facial Caricaturing with Motion Carica- turing in PICASSO System," in Proc. *IEEE/ASME Int. Conf. Advanced Intelligent Mechatronics*, Tokyo, Japan, Jun. 1997, pp. 30-37.

[5] Y. Li and H. Kobatake, "Extraction of Facial Sketch Based on Morpho-logical Processing," in Proc. *IEEE Int. Conf. Image Process*., Washington DC, USA, Oct. 1997, pp. 316-319.

[6] H. Chen, N. Zheng, Y. Xu, and H. Shum, "An Example-based Facial Sketch Generation System," *J. Softw.*, 2003, vol. 14, no. 2, pp. 202-208.

[7] J. Wang, H. Bao, and W. Zhou, "Automatic Image-based Pencil Sketch Rendering," J. Comput. Sci. Technol., 2002, vol. 17, no. 5, pp. 347-356.

[8] B. Klar, Z. Li, and A. Jain, "Matching Forensic Sketches to Mug Shot Photos", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2011, vol. 33, no. 3, pp. 639-646.

[9] Y. Li, M. Savvides, and V. Bhagavatula, "Illumination Tolerant Face Recognition Using a Novel Face from Sketch Synthesis Approach and Advanced Correlation Filters," in Proc. IEEE *Int. Conf. Acoustics, Speech, and Signal Processing*, Toulouse, France, 15-19 May 2006, pp. 357-360.

[10] W. Liu, X. Tang, and J. Liu, "Bayesian Tensor Inference for Sketch-based Facial Photo Hallucination," in Proc. *Int. Joint Conf. Artificial Intelligence*, Hyderabad, India, Jan. 2007, pp. 2141-2146.

[11] Q. Liu and X. Tang, "A Nonlinear Approach for Face Sketch Synthesis and Recognition," in Proc. *IEEE Int. Conf. Computer Vision and Pattern Recognition*, San Diego, CA, Jun. 2005, pp. 1005-1010.

[12] X. Wang and X. Tang, "Face Photo-Sketch Synthesis and Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2009, vol. 31, no. 11, pp. 1955-1967.

[13] W. Zhang, X. Wang, and X. Tang, "Lighting and Pose Robust Face Sketch Synthesis," *In Proc. European Conference on Computer Vision (6)*, Hersonissos, Heraklion, Crete, Greece, 6-9 Sep. 2010, pp. 420-433.

[14] X. Gao, J. Zhong, D. Tao, and X. Li, "Local Face Sketch Synthesis Learning," *Neurocomputing*, 2008, vol. 71, no. 10-12, pp. 1921-1930.

[15] B. Xiao, X. Gao, X. Li, and D. Tao, "A New Approach for Face Recognition by Sketches in Photos," *Signal Processing*, 2009, vol. 89, no. 8, pp. 1531-1539.

[16] S. Roweis and L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, 2000, vol. 290, no. 5500, pp. 2323-2326.

[17] B. Olshausen, D. Field, "Emergence of Simple-cell Receptive Field Properties by Learning a Sparse Code for Natural Images," *Nature*, 1996, vol. 381, no. 13, pp. 607-609.

[18] J. Yang, J. Wright, T. Huang and Y. Ma, "Image Super-resolution via Sparse Representation," *IEEE Trans. Image Processing*, 2010, vol. 19, no. 11, pp. 2861-2873.

[19] P. Phillips, H. Moon, S. Rizvi, P. Rauss, "The FERET Evaluation Methodology for Face Recognition Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2000, vol. 22, no. 10, pp. 1090-1104.

[20] Serrano, I. Diego, C. Conde, E. Cabello, L. Shen and L. Bai, "Influence of wavelet frequency and orientation in an SVM-based parallel Gabor PCA face verification system," in *Proc. Conference on Intelligent Data Engineering and Automated Learning*. vol. 4881, Springer-Verlag Birmingham, UK, 16-19 Dec. 2007, pp. 219-228.

[21] Indian Face Database: http://vis-www.cs.umass.edu/~vidit/IndianFaceDatabase/,2002

[22] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2009, vol. 31, no. 2, pp. 210-227.