Multiple Representations-Based Face Sketch–Photo Synthesis

Chunlei Peng, Xinbo Gao, *Senior Member, IEEE*, Nannan Wang, Dacheng Tao, *Fellow, IEEE*, Xuelong Li, *Fellow, IEEE*, and Jie Li

***Abstract*— Face sketch–photo synthesis plays an important role in law enforcement and digital entertainment. Most of the existing methods only use pixel intensities as the feature. Since face images can be described using features from multiple aspects, this paper presents a novel multiple representations-based face sketch–photo-synthesis method that adaptively combines multiple representations to represent an image patch. In particular, it combines multiple features from face images processed using multiple filters and deploys Markov networks to exploit the interacting relationships between the neighboring image patches. The proposed framework could be solved using an alternating optimization strategy and it normally converges in only five outer iterations in the experiments. Our experimental results on the Chinese University of Hong Kong (CUHK) face sketch database, celebrity photos, CUHK Face Sketch FERET Database, IIIT-D Viewed Sketch Database, and forensic sketches demonstrate the effectiveness of our method for face sketch–photo synthesis. In addition, cross-database and database-dependent style-synthesis evaluations demonstrate the generalizability of this novel method and suggest promising solutions for face identification in forensic science.**

***Index Terms*— Face recognition, face sketch–photo synthesis, forensic sketch, multiple representations.**

Manuscript received May 6, 2014; revised April 9, 2015 and July 29, 2015; accepted July 30, 2015. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2012CB316400, in part by the National Natural Science Foundation of China under Grant 61125204, Grant 61172146, Grant 61432014, and Grant 61501339, in part by the Fundamental Research Funds for the Central Universities under Grant JB149901 and Grant XJS15049, in part by the Program for Changjiang Scholars and Innovative Research Team in University of China under Grant IRT13088, in part by the Shaanxi Innovative Research Team for Key Science and Technology under Grant 2012KCT-02, in part by the Key Research Program of the Chinese Academy of Sciences under Grant KGZDEW-T03, and in part by the Australian Research Council Projects under Grant FT-130101457, Grant DP-140102164, and Grant LP-140100569.

C. Peng and J. Li are with the School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: clp.xidian@gmail.com; leejie@mail.xidian.edu.cn).

X. Gao is with the State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: xbgao@mail.xidian.edu.cn).

N. Wang is with the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: nnwang@xidian.edu.cn).

D. Tao is with the Centre for Quantum Computation and Intelligent Systems, and the Faculty of Engineering and Information Technology, University of Technology, Sydney, Ultimo, NSW 2007, Australia (e-mail: dacheng.tao@uts.edu.au).

X. Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), the State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xuelong_li@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2015.2464681

I. Introduction

A NUMBER of face sketch–photo-synthesis methods have recently been developed due to their widespread utility in law enforcement and digital entertainment [1]–[5]. For example, in law enforcement, a photo of the suspect is not always available, and an eyewitness description can be the only clue to assist in identifying suspects.[1,2] However, direct face recognition of the suspect is difficult because of the variety of geometric and textural differences between the sketches and the photographs [6]–[9]. To reduce discrepancies between the face sketches and the photos, the face sketch–photo synthesis can be applied to transform them into the same modality [2], [3], [10], [11]. Furthermore, the face sketch–photo synthesis has been applied to digital entertainment and facial animation [12].

However, most existing methods are constrained by the fact that face images need to be captured under controlled conditions, such as even illumination, similar backgrounds, and using faces of the same ethnic origin. Moreover, these approaches have only been successful when the training data and the test data originate from the same data set. Unfortunately, these constraints are difficult to meet in practice. These limitations occur, because existing approaches only consider single representation (the image patch intensities), while in real-world scenarios, images can be described by features acquired from multiple complimentary representations. This is because different features explain distinct aspects of the visual characteristics present in face images, and features from multiple representations can represent face images more accurately and robustly. Although some approaches utilize multiple representations, most of these are applied to discriminative problems, such as classification [12], [13], image annotation [14], reranking [15], super-resolution [16], and recognition [17]–[19]. Zhang *et al.* [20] proposed a lighting and pose robust method to use multiple representations in a Markov random field (MRF) model. However, they manually set the weights of multiple representations, and the weights were fixed on the whole face image. The best way to adaptively take advantage of multiple representations for image synthesis remains an unresolved problem.

One simple solution is to combine multiple representations by directly concatenating all the feature vectors into a single

[1] http://www.askaforensicartist.com/composite-sketch-leads-to-arrest-in-virginia-highland-robbery/.

[2] http://www.askaforensicartist.com/phoenix-police-sketch-leads-to-arrest-of-kidnapper/.

vector, and then applying existing methods to the single vector. However, this concatenation is suboptimal, because each single representation has a specific statistical property, and directly concatenating the feature vectors ignores the diversity of multiple representations, and the complementary information in an image patch is insufficiently explored.

In this paper, we propose a novel multiple representations-based face sketch–photo-synthesis (MrFSPS) method to address these problems. First, all face sketches and photos are evenly divided into patches with overlap between the neighboring patches. Second, image patch intensities are extracted along with two low-level features from each image patch from the original face image and face images processed by three filters as multiple representations, i.e., 12 feature types in total. Third, MrFSPS is modeled based on Markov networks, and distinct features are adaptively combined using weighting. Fourth, the aforementioned framework is optimized using an alternating optimization strategy. Finally, a minimum error boundary cut algorithm [21] is adopted to stitch the overlapping areas. By switching the roles of sketches and photos, our proposed method can handle both the face sketch synthesis and the face photo synthesis. To illustrate our method, we, therefore, only perform the face sketch synthesis from an input photo.

The main contributions of this paper are as follows.

1) Multiple representations are used to describe a face image patch in the face sketch–photo-synthesis problem.
2) An effective and efficient framework based on Markov networks is proposed to adaptively learn the combination weights of multiple representations.
3) A database-dependent style-synthesis strategy is designed, which achieves promising performance on forensic sketch–photo synthesis and recognition.
4) Perceptive and quantitative experiments are used to illustrate the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section II outlines the existing literature on face sketch–photo synthesis. Section III introduces the proposed MrFSPS method. The experimental results and analysis are presented in Section IV. Finally, the conclusion is drawn in Section V.

II. Related Work

Existing face sketch–photo-synthesis methods can be divided into three main categories [1]: 1) subspace learning-based; 2) sparse representation-based; and 3) Bayesian inference-based approaches.

Subspace learning-based methods include the linear subspace method, which is based on the principal component analysis [22], and the nonlinear subspace method, which is based on local linear embedding (LLE) [23]. Tang and Wang [10], [24] first considered the face-sketch-synthesis procedure as a linear process and proposed an eigensketch transformation method. The input photo was projected onto the training photo set to obtain projection coefficients, and the target sketch was then synthesized by linearly combining the corresponding training sketches with the previously obtained projection coefficients. They subsequently proposed separating both the shape and the texture and the conducted eigentransformation [11]; the target sketch was then obtained by fusing the synthesized shape and the texture. 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.* [2] proposed a nonlinear face sketch–photo-synthesis algorithm by dividing the face image into overlapping patches. For an input test photo, each photo patch was reconstructed using a linear combination of photo patches selected from the training set. The corresponding sketch patches in the training set were selected as candidates, and the synthesized sketch patches obtained from the linear combination of candidate sketch patches generated the target sketch. This approach had the drawback that each patch was independently synthesized and thus neglected compatible relationships between the neighboring image patches. The global structural information, therefore, had the potential to be poorly synthesized.

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.* [25] assumed that the face photo patch and the corresponding sketch patch could be decomposed on the photo patch dictionary and the sketch patch dictionary with the same sparse coefficients. By building a coupled dictionary with the photo and sketch patch pairs using sparse coding [26], 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. Finally, a target sketch could be created by fusing the obtained sketch patches. In order to improve the quality of the synthesized images, Wang *et al.* [27] constructed a multidictionary sparse representation-based face sketch–photo-synthesis model. High frequency and detailed information were hallucinated to enhance the synthesized results. Gao *et al.* [28] also utilized sparse representation for neighbor selection, and proposed a two-step framework to further improve the quality of the synthesized image. They adaptively selected neighbors by sparse neighbor selection and proposed a sparse-representation-based enhancement strategy to enhance the quality of the synthesized photos and sketches. Chang *et al.* [29] built a local regression model using a multivariate output regression method to enforce structural constraints on the synthesized results. Zhang *et al.* [30] proposed a support vector regression-based approach to reduce the loss of high-frequency information that limits existing methods.

Bayesian inference-based methods can be further divided into the embedded hidden Markov model (EHMM)-based methods and the MRFs-based methods. Hidden Markov models have widely been applied to speech-recognition problems [31]. Considering the large amount of 2-D spatial information in face images, Gao *et al.* [32] employed EHMM to model the nonlinear relationship between the sketches and their photo counterparts. A face image was decomposed into five superstates that included the forehead, eye, nose,

mouth, and chin. Each superstate was then decomposed into several embedded states to extract the local features in the face image. A series of synthesized sketches was generated and fused using the selective ensemble strategy to obtain the target result. Xiao *et al.* [33] extended EHMM to face the photo synthesis and the recognition. By taking the relationship between the face image patch and its neighboring patches into consideration, Wang and Tang [3] applied a multiscale MRF model to face the sketch–photo synthesis and the recognition. A local evidence function provided the dependence between the input test photo patch and the target sketch patch, and a compatibility function provided the constraints between the target sketch patch and its neighboring patches, with the assumption that the neighboring patches should be consistent in their overlapping regions. Zhang *et al.* [20] extended the model in [3] to a lighting- and pose-robust method. Robustness was achieved using three steps. First, the prior shape information was introduced to reduce distortion. Second, the similarity of a photo patch to its corresponding sketch patch was measured by a weighted combination of the Euclidean distances between the patches, filtered by a difference of Gaussian (DoG) filter and the distance between the dense scale-invariant feature transform (SIFT) [34] descriptors of two image patches. Third, the distance between the dense SIFT descriptors of neighboring patches, together with the Euclidean distance of the overlapping areas, were explored to measure gradient compatibility. Considering that the aforementioned MRF model-based methods generate some facial deformations and have only limited synthesis ability when the training set is small, Zhou *et al.* [4] proposed a Markov weight fields (MWFs) model capable of synthesizing new patches that do not exist in the training set. They formulated their model into a convex quadratic programming (QP) problem and proposed a cascade decomposition method (CDM) to solve it. However, the MWF model did not perform well when the input face images were not produced under controlled conditions.

III. Multiple Representations-Based Face Sketch Synthesis

Considering a training set with M face photo–sketch pairs $(\mathbf{p}^1, \mathbf{s}^1), \ldots, (\mathbf{p}^M, \mathbf{s}^M)$, we first divide each face image into N overlapping patches. An input test photo t can also be divided into N overlapping patches. Let $\mathbf{y}_i$ be the ith test photo patch, where $i = 1, 2, \ldots, N$. We extract L representations to represent a photo patch $\mathbf{y}_i$, denoted by $\mathbf{f}_l(\mathbf{y}_i)$, $l = 1, 2, \ldots, L$. Here, L is set to 12. For each test photo patch, we find K candidate photo patches from the training set in terms of the distance of the multiple representations. The corresponding K candidate sketch patches can then be obtained from the training set. We assume that the weights of multiple representations at various locations in photo patches are different, and adaptively learn the weights of multiple representations and candidate patches by alternating optimization. In Section III-A, we introduce the multiple representations, and in Sections III-B and III-C, we present the multiple representations-based face sketch synthesis model and its optimization.

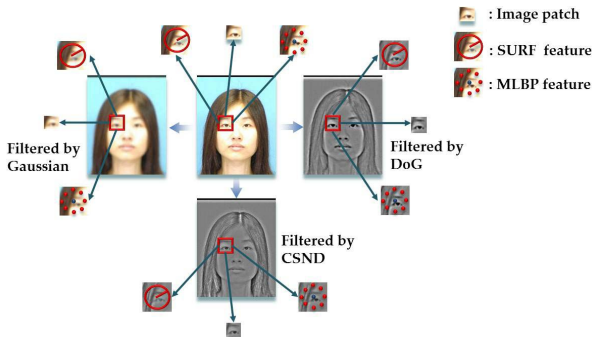


Fig. 1. Example of the multiple representations extracted from an input photo.

A. Multiple Representations Extracted for Face Sketch–Photo Synthesis

In this paper, we utilize image patch intensities, speeded up robust features (SURFs) [35], and multiscale local binary pattern (MLBP) [36] as the features to represent a face image patch. Furthermore, each face image is filtered using three filters: 1) the DoG filter; 2) the center-surround divisive normalization (CSDN) filter [37]; and 3) the Gaussian smoothing filter. Finally, we extract three features from the original face image and three filtered face images to obtain the multiple representations used in this paper. This method is shown in Fig. 1. Except for the features defined on the original pixels and patches, the filters and features used in this paper are also used for forensic sketch matching [17]. Note that there are many other features, such as SIFT [34], histogram of oriented gradient (HOG) [38], and the facial feature designed for photo–sketch recognition [39], which can also be used in this paper. In order to evaluate the influence when different feature descriptors are used, we replaced the SURF and MLBP feature used in this paper with the SIFT and HOG features, respectively, and conducted face sketch–photo synthesis on the forensic sketch database. Based on the visual comparisons of the synthesized results, little influence is introduced when different features were used in our method. We further conducted face-recognition experiments and the accuracies when different features were used and were exactly the same under the same experimental setting (detail settings are given in Section IV-F).

Two low-level features, together with the image patch intensities, are deployed to represent an image patch. The SURF has previously been successfully applied to face recognition and object detection in [40] and [41]. The SURF extraction process consists of a detection phase and a descriptor construction phase; however, in this paper, we skip the detection phase and directly extract the SURF descriptors on the center of image patches, each of which describes an image patch with a 64-dimensional vector. Ojala *et al.* [36] designed the LBPs feature as a texture descriptor, which has successfully been applied to face recognition and heterogeneous face recognition [17], [42]. The LBP feature is created at a particular pixel location by thresholding the 3×3 neighborhood with the center pixel value and considering the subsequent pattern as a binary number. The exact image patch intensities are ignored and only the relationship between the center intensity and the neighbors are used. Therefore, LBP is less sensitive to variations in

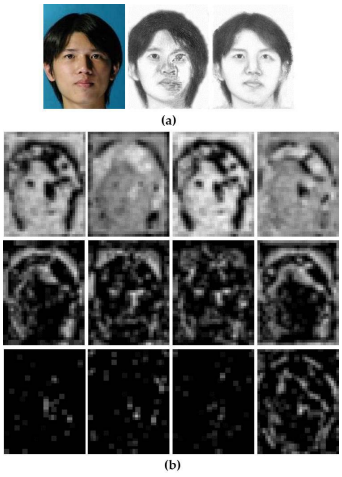


Fig. 2. Weight distribution of multiple representations learned by the proposed approach. (a) Input photo, result of the MWF method, and result of the proposed method. (b) Weight distribution images of 12 features used in this paper. First row to third row: image patch intensities, SURF, and MLBP feature. First column to fourth column: original image, DoG filter, CSDN filter, and Gaussian filter.

illumination and it extracts facial features while considering the shape and the textural information; it is, therefore, robust to different facial expressions and illumination conditions. A uniform LBP is then designed, such that the binary pattern contains, at most, two transitions from 0 to 1 (or vice versa) when the pattern is considered circular. In this paper, we utilize the MLBP, which is the concatenation of uniform LBP feature descriptor vectors with different radiuses. Here, the radii are set as follows: $r = (1, 2, 3, 4)$. The uniform LBP descriptor yields a 59-dimensional vector, resulting in a 236-dimensional MLBP feature to describe each image patch.

When the features are extracted from an image patch, the image patch intensities and the original images contain the most important and useful information. To help preserve the information extracted from a photo image, the original images are filtered using three different image filters. DoG is an edge enhancement algorithm, similar to the architecture of the retina's visual receptive field, which can be used to remove the effect of lighting variations [20]. The DoG filter is applied by convolving the photo image with the difference of two Gaussian kernels with two standard deviations, σ_0 and σ_1; here, $\sigma_0 = 1$ and $\sigma_1 = 4$. CSDN eliminates intensity gradients caused by shadows and can be seen as complementary to the DoG filter; it is analogous to the center-on/surround-off processing that occurs in the retina. The CSDN filter divides the intensity of each pixel by the mean of its neighborhood intensities. In this paper, we set the neighborhood size s to 16. The Gaussian smoothing filter is designed with $\sigma = 2$, and it is used to remove the noise contained in high spatial frequencies in the input photo image. All the above extracted features are normalized with their corresponding squared sums equaling one.

The weight distribution of multiple representations learned by the proposed approach is shown in Fig. 2. The size of weight images shown in Fig. 2 is the same to the size of input images, which is 200×250 in this paper. The weight of the feature in the lighter area is closer to 1, and the weight of the feature in the darker area is closer to 0. It can be seen that the weight of the image patch intensities is higher in plain areas, such as the background and hair regions, while the weight of the SURF is higher in the facial structure areas. The MLBP feature after Gaussian filtering is most helpful in those areas depicting facial structures, and further improves the details. Comparing the synthesized result of the MWF [4] method and the proposed method in Fig. 2, the most challenging region of the input photo is around the nose. In this challenging subregion, the weights of the image patch intensities in original image, after CSDN filtering and Gaussian filtering are close to 0, while the weight of the image patch intensities in DoG filter is close to 1. This weight distribution is consistent with the characteristic of DoG filter that the DoG filter is helpful for removing the effect of lighting variations [20]. The weights of SURF after DoG filtering and CSDN filtering are close to 1 around the nose, which is helpful for enhancing the edge structure of the synthesized sketch. The MLBP feature after Gaussian filtering contributes to the synthesis of the nose region, too. These weight distributions are consistent with the characteristics of the features and filters, and illustrate the rationale from the proposed multiple representation-based approach. Multiple representations can help increase the information obtained from the input images. However, more features will lead to a higher cost of computation and storage requirements. Therefore, features that are robust to illuminations and sensitive to the edge structures should be used, while irrelevant features may bring interferences to the synthesized result.

B. Multiple Representations-Based Face Sketch Synthesis

In order to synthesize the sketch patch $\mathbf{x}_i$ corresponding to the input photo patch $\mathbf{y}_i$, where $i = 1, 2, \ldots, N$, we find K candidate photo patches $\{\mathbf{y}_{i,1}, \mathbf{y}_{i,2}, \ldots, \mathbf{y}_{i,K}\}$, where $\mathbf{y}_{i,k}$ represents the kth candidate photo patch for the ith input photo patch $\mathbf{y}_i$ within the search region around the location of $\mathbf{y}_i$. The target sketch patch $\mathbf{x}_i$ $(i = 1, 2, \ldots, N)$ can then be obtained by the linear combination of the corresponding K candidate sketch patches $\{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,K}\}$ weighted by the K-dimensional vector $\mathbf{w}_i$, where $w_{i,k}$ represents the weight of the kth candidate sketch patch as follows:

$$\mathbf{x}_i = \sum_{k=1}^{K} w_{i,k} \mathbf{x}_{i,k} \tag{1}$$

where $\sum_{k=1}^{K} w_{i,k} = 1$.

To estimate the weights of the candidate sketch patches and generate a target sketch of high quality, we need to jointly model all the patches using the Markov network framework, similar to [3]–[5] and [20]. Since the target sketch patches are only dependent on the weights, as shown in (1), the joint probability of the input photo patches and the target sketch patches is equal to that of the input photo patches and the weights. It is defined as

$$\begin{aligned} &p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{y}_1, \ldots, \mathbf{y}_N) \\ &\quad = p(\mathbf{w}_1, \ldots, \mathbf{w}_N, \mathbf{y}_1, \ldots, \mathbf{y}_N) \\ &\quad = \prod_i \Phi(\mathbf{f}(\mathbf{y}_i), \mathbf{f}(\mathbf{w}_i)) \prod_{(i,j) \in \Xi} \Psi(\mathbf{w}_i, \mathbf{w}_j) \end{aligned} \tag{2}$$

where $(i, j) \in \Xi$ means that the ith image patch and the jth image patch are neighbors, $\Phi(\mathbf{f}(\mathbf{y}_i), \mathbf{f}(\mathbf{w}_i))$ is the local evidence function, and $\Psi(\mathbf{w}_i, \mathbf{w}_j)$ is the neighboring compatibility function. In the local evidence function, $\mathbf{f}(\mathbf{y}_i) = [\mathbf{f}_1(\mathbf{y}_i), \mathbf{f}_2(\mathbf{y}_i), \ldots, \mathbf{f}_L(\mathbf{y}_i)]$, where $\mathbf{f}_l(\mathbf{y}_i)$ means the lth representation of the photo patch $\mathbf{y}_i$, and $\mathbf{f}(\mathbf{w}_i) = [\mathbf{f}_1(\mathbf{w}_i), \mathbf{f}_2(\mathbf{w}_i), \ldots, \mathbf{f}_L(\mathbf{w}_i)]$, in which $\mathbf{f}_l(\mathbf{w}_i) = \sum_{k=1}^{K} w_{i,k}\mathbf{f}_l(\mathbf{y}_{i,k})$.

$\Phi(\mathbf{f}(\mathbf{y}_i), \mathbf{f}(\mathbf{w}_i))$ can be representated as

$$\Phi(\mathbf{f}(\mathbf{y}_i), \mathbf{f}(\mathbf{w}_i)) \propto \exp\left\{-\sum_{l=1}^{L}\mu_{i,l}\left\|\mathbf{f}_l(\mathbf{y}_i) - \sum_{k=1}^{K} w_{i,k}\mathbf{f}_l(\mathbf{y}_{i,k})\right\|^2 \Big/ 2\delta_\Phi^2\right\} \quad (3)$$

where $\mu_i = [\mu_{i,1}, \mu_{i,2}, \ldots, \mu_{i,L}]$. $\mu_{i,l}$ represents the weight of the distance of the lth representation between the ith photo patch $\mathbf{f}_l(\mathbf{y}_i)$ and the combination of its candidates. The rationale behind the local evidence function is that if $\sum_{k=1}^{K} w_{i,k}\mathbf{x}_{i,k}$ is a good estimation of $\mathbf{x}_i$, $\mathbf{f}_l(\sum_{k=1}^{K} w_{i,k}\mathbf{y}_{i,k})$ should be similar to $\mathbf{f}_l(\mathbf{y}_i)$. In order to exploit the relationship between the target image patch and its candidate image patches represented by multiple representations, we simply assume that $\sum_{k=1}^{K} w_{i,k}\mathbf{f}_l(\mathbf{y}_{i,k})$ should also be similar to $\mathbf{f}_l(\mathbf{y}_i)$ here, which is easier to be optimized, too.

The neighboring compatibility function $\Psi(\mathbf{w}_i, \mathbf{w}_j)$ is defined as

$$\Psi(\mathbf{w}_i, \mathbf{w}_j) \propto \exp\left\{-\left\|\sum_{k=1}^{K} w_{i,k}\mathbf{o}_{i,k}^{j} - \sum_{k=1}^{K} w_{j,k}\mathbf{o}_{j,k}^{i}\right\|^2 \Big/ 2\delta_\Psi^2\right\} \quad (4)$$

where $\mathbf{o}_{i,k}^{j}$ represents the overlapping area of the candidate sketch patch $\mathbf{x}_{i,k}$ with the jth patch. This term is utilized to guarantee that neighboring patches have compatible overlaps. In (3) and (4), δ_Φ and δ_Ψ are two parameters balancing the local evidence function and the neighboring compatibility function separately.

To avoid the weight of multiple representations overfitting to one representation [43], a regularization term $\exp\{-\lambda_i\|\mu_i\|^2\}$ is added to the joint probability in (2)

$$\begin{aligned} &p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{y}_1, \ldots, \mathbf{y}_N) \\ &= p(\mathbf{w}_1, \ldots, \mathbf{w}_N, \mathbf{y}_1, \ldots, \mathbf{y}_N) \\ &= \prod_{(i,j)\in\Xi}\Psi(\mathbf{w}_i, \mathbf{w}_j)\prod_i \Phi(\mathbf{f}(\mathbf{w}_i), \mathbf{f}(\mathbf{y}_i))\prod_i \exp\{-\lambda_i\|\mu_i\|^2\} \end{aligned} \quad (5)$$

where $\mu_i = [\mu_{i,1}, \mu_{i,2}, \ldots, \mu_{i,L}]$ and λ_i balances the regularization term with the other two terms. To obtain the optimal weights for sketch synthesis, we need to maximize the joint probability in (5). Optimization details are introduced in Section III-C, and the proposed approach is shown in Fig. 3.

C. Optimization

Substituting (3) and (4) into (5), we find that maximizing the joint probability is equivalent to minimizing the following problem:

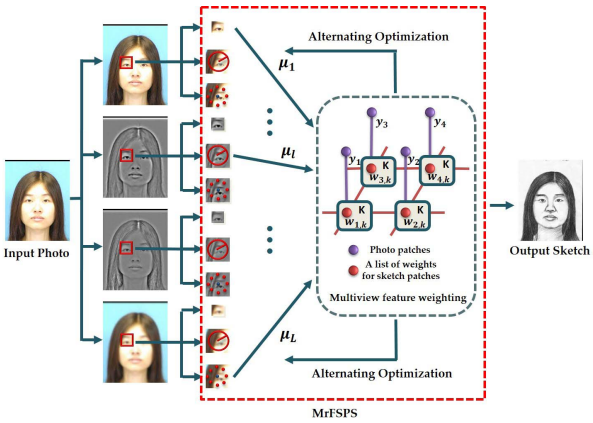


Fig. 3. Framework of the proposed multiple representations-based face sketch synthesis.

$$\begin{aligned} \min_{\mathbf{w},\mu}\ &\frac{1}{2\delta_\Psi^2}\sum_{(i,j)\in\Xi}\left\|\sum_{k=1}^{K} w_{i,k}\mathbf{o}_{i,k}^{j} - \sum_{k=1}^{K} w_{j,k}\mathbf{o}_{j,k}^{i}\right\|^2 \\ &+\frac{1}{2\delta_\Phi^2}\sum_{i=1}^{N}\sum_{l=1}^{L}\mu_{i,l}\left\|\mathbf{f}_l(\mathbf{y}_i) - \sum_{k=1}^{K} w_{i,k}\mathbf{f}_l(\mathbf{y}_{i,k})\right\|^2 + \sum_{i=1}^{N}\lambda_i\|\mu_i\|^2 \\ \text{s.t.}\ &\sum_{k=1}^{K} w_{i,k} = 1, \quad 0 \le w_{i,k} \le 1 \\ &\sum_{l=1}^{L}\mu_{i,l} = 1, \quad 0 \le \mu_{i,l} \le 1 \end{aligned} \quad (6)$$

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

We apply the alternating optimization strategy to solve this problem via the following two steps.

1) For a fixed μ, we can ignore the regularization term. It is equivalent to minimizing

$$\begin{aligned} \min_{\mathbf{W}}\ &\alpha\sum_{(i,j)\in\Xi}\|\mathbf{O}_i^j\mathbf{W} - \mathbf{O}_j^i\mathbf{W}\|^2 \\ &+\sum_{i=1}^{N}\sum_{l=1}^{L}\mu_{i,l}\|\mathbf{f}_l(\mathbf{y}_i) - \mathbf{F}_{i,l}\mathbf{W}\|^2 \end{aligned} \quad (7)$$

where $\alpha = \delta_\Phi^2/\delta_\Psi^2$. $\mathbf{F}_{i,l}$ and $\mathbf{O}_i^j$ are two matrices, with the $((i-1)K+k)$th column being $\mathbf{f}_l(\mathbf{y}_{i,k})$ and $\mathbf{o}_{i,k}^{j}$, respectively, and the other columns being zero vectors. $\mathbf{W}$ is an NK-dimensional vector, with the $((i-1)K+k)$th element being $w_{i,k}$. Equation (7) can be formulated into the following problem:

$$\begin{aligned} \min_{\mathbf{W}}\ &\mathbf{W}^T\mathbf{Q}\mathbf{W} + \mathbf{W}^T\mathbf{C} + \mathbf{b} \\ \text{s.t.}\ &\sum_{k=1}^{K} w_{i,k} = 1, \quad 0 \le w_{i,k} \le 1 \\ &i = 1, 2, \ldots, N, \quad k = 1, 2, \ldots, K \end{aligned} \quad (8)$$

where

$$\begin{aligned}\mathbf{Q} &= \alpha \sum_{(i,j)\in\Xi} \left(\mathbf{O}_i^j - \mathbf{O}_j^i\right)^T \left(\mathbf{O}_i^j - \mathbf{O}_j^i\right) \\ &\quad + \sum_{i=1}^{N}\sum_{l=1}^{L} \mu_{i,l}\mathbf{F}_{i,l}^T\mathbf{F}_{i,l} \\ \mathbf{C} &= -2\sum_{i=1}^{N}\sum_{l=1}^{L} \mu_{i,l}\mathbf{F}_{i,l}^T\mathbf{f}_l(\mathbf{y}_i) \\ \mathbf{b} &= \sum_{i=1}^{N}\sum_{l=1}^{L} \mu_{i,l}\mathbf{f}_l^T(\mathbf{y}_i)\mathbf{f}_l(\mathbf{y}_i). \end{aligned} \tag{9}$$

The last term $\mathbf{b}$ has no effect on the optimization result, and we can further simplify the optimization problem into

$$\begin{aligned} &\min_{\mathbf{W}} \ \mathbf{W}^T\mathbf{Q}\mathbf{W} + \mathbf{W}^T\mathbf{C} \\ &\text{s.t.} \sum_{k=1}^{K} w_{i,k} = 1, \quad 0 \le w_{i,k} \le 1 \\ &\qquad i = 1, 2, \ldots, N, \quad k = 1, 2, \ldots, K. \end{aligned} \tag{10}$$

This can be solved using the CDM proposed in [4].

2) For a fixed $\mathbf{w}$, we can ignore the compatibility function and we obtain the following problem:

$$\begin{aligned} &\min_{\mu} \ \frac{1}{2\delta_\Phi^2}\sum_{i=1}^{N}\sum_{l=1}^{L}\mu_{i,l}\left\|\mathbf{f}_l(\mathbf{y}_i) - \sum_{k=1}^{K} w_{i,k}\mathbf{f}_l(\mathbf{y}_{i,k})\right\|^2 \\ &\qquad + \sum_{i=1}^{N}\lambda_i\|\mu_i\|^2 \\ &\text{s.t.} \sum_{l=1}^{L}\mu_{i,l} = 1, \quad 0 \le \mu_{i,l} \le 1 \\ &\qquad i = 1, 2, \ldots, N, \quad l = 1, 2, \ldots, L. \end{aligned} \tag{11}$$

Actually, each μ_i can be solved separately in (11). Letting

$$s_{i,l} = \frac{1}{2\delta_\Phi^2}\left\|\mathbf{f}_l(\mathbf{y}_i) - \sum_{k=1}^{K} w_{i,k}\mathbf{f}_l(\mathbf{y}_{i,k})\right\|^2 \tag{12}$$

the optimization problem (11) can be transformed into the following problem:

$$\begin{aligned} &\min_{\mu_i} \ \sum_{l=1}^{L}\mu_{i,l}s_{i,l} + \lambda_i\|\mu_i\|^2 \\ &\qquad \Rightarrow \min_{\mu_i} \ \lambda_i\mu_i^T\mu_i + \mathbf{s}_i^T\mu_i \end{aligned} \tag{13}$$

where $\mathbf{s}_i = (s_{i,1}, s_{i,2}, \ldots, s_{i,L})^T$. The parameter λ_i is set in a data driven manner similar to [43]. Geng *et al.* [43] investigated that the middle range value performed best, i.e., a balance between a unanimous weighting and a single manifold. We adaptively set

$$\lambda_i = \frac{\frac{1}{2\delta_\Phi^2}\sum_{l=1}^{L}\left\|\mathbf{f}_l(\mathbf{y}_i) - \sum_{k=1}^{K} w_{i,k}\mathbf{f}_l(\mathbf{y}_{i,k})\right\|^2}{L}. \tag{14}$$

Algorithm 1 Multiple Representations-Based Face Sketch Synthesis

Input: Training photo–sketch pairs $(\mathbf{p}^1, \mathbf{s}^1), \cdots, (\mathbf{p}^M, \mathbf{s}^M)$, test photo $\mathbf{t}$, the number of neighboring patches K, the number of multiple representations L, α, the size of an image patch, the size of the overlapping region and the search region;

Initialize: $\mu_i = [\mu_{i,1}, \mu_{i,2}, \cdots, \mu_{i,L}] = [1/L, 1/L, \cdots, 1/L]$, $\mathbf{w}_i = [w_{i,1}, w_{i,2}, \cdots, w_{i,K}] = [1/K, 1/K, \cdots, 1/K]$, and $i = 1, 2, \cdots, N$;

***Step* 1**: All the face images are evenly divided into some patches with overlap existing between the neighboring patches. For each test photo patch $\mathbf{y}_i$, find its K candidate neighboring patches from the training photo patches within the search region around the location of $\mathbf{y}_i$ according to the Euclidean distance of their multiple representations;

***Step* 2**: Optimize the problem (10) to compute the weights of candidate patches $\mathbf{w}$;

***Step* 3**: For every μ_i, compute $\mathbf{s}_i$ and λ_i according to (12) and (14). Optimize the problem (13) to compute the weights corresponding to multiple representations;

***Step* 4**: Iterate ***Step* 2** and ***Step* 3** until convergence;

***Step* 5**: Generate all the synthesized sketch patches with the weighted combination of candidate patches. Stitch them together through the algorithm [21] to obtain the target sketch.

Output: The target sketch.

Equation (13) is a standard convex QP problem and can efficiently be optimized.

We initialize every $\mu_{i,l}$ to $1/L$ and every $w_{i,k}$ to $1/K$, and then alternately perform 1) and 2) until convergence. In our experiments, it always converges after five iterations. Once the weight matrix $\mathbf{w}$ is obtained, the target sketch patches can be synthesized by the linear combination of the weighted candidate sketch patches. Finally, we apply a minimum error boundary cut method to stitch the neighboring synthesized sketches [21].

In reality, the proposed method can be simplified into the MWF method [4] when only a single representation, the image patch intensity from the original image, is explored. The proposed method is summarized in Algorithm 1.

IV. Experimental Results

We conducted our experiments on the Chinese University of Hong Kong (CUHK) face sketch database [3], a set of Chinese celebrity photos obtained from the Web [20], the CUHK Face Sketch FERET Database (CUFSF) [39], the IIIT-D Viewed Sketch Database [44], and our newly collected forensic sketch database. The CUHK face sketch database contains 606 photo-sketch pairs collected from three databases: 1) the CUHK student database (188 photos); 2) the A. Marinez-R. Benavente (AR) database (123 photos) [45]; and 3) the XM2VTS database (295 photos) [46]. All the photos are frontal, with a neutral expression, and taken under normal lighting conditions,

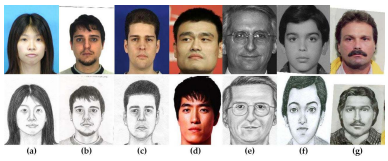


Fig. 4. Examples of the databases used in this paper. (a)–(c) Photo–sketch pairs from the CUHK database. (d) Chinese celebrity photos. (e) Photo–sketch pair from the CUFSF database. (f) Photo–sketch pair from the IIIT-D Viewed Sketch Database. (g) Photo–sketch pair from the forensic sketch database.

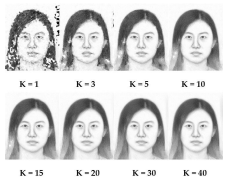


Fig. 5. Effect of different numbers of neighboring patches K.

and the sketches are drawn by an artist based on these photos. The photos of Chinese celebrities were collected from the Web and, therefore, have a variety of backgrounds, and the lighting and pose are variable. The CUFSF database includes 1194 photo–sketch pairs collected from the FERET database [47], with photos containing lighting variations and sketches containing the shape exaggeration. The IIIT-D Viewed Sketch Database contains 238 photo–sketch pairs collected from different sources. The forensic sketch database contains 171 real-world forensic sketches and their corresponding mug shot photos. The forensic sketches were drawn by sketch artists using the descriptions of eyewitnesses or victims. The forensic sketch database originates from a collection of images from the forensic sketch artist Gibson [6], the forensic sketch artist Taylor [48], and other Internet sources. Examples from the databases are shown in Fig. 4. It can be seen that the photos are database-dependent; there is a variety of different lighting conditions, backgrounds, skin colors, and ethnic origins.

A. Experimental Settings

The parameters used were as follows: the size of the images used in this paper was 200×250, the image patch size was 10, and the overlapping region size was 5. Therefore, there were 1911 patches per image. The neighborhood search region was 20×20. We did not need to set δ_Φ or δ_Ψ, and instead α was set to 0.025, where $\alpha = \delta_\Phi^2/\delta_\Psi^2$. The number of neighboring patches K was set to 20. The number of multiple representations was 12, as shown in Fig. 1, and the effect of different numbers of neighboring patches K is shown in Fig. 5. When the neighborhood size is small, there is a large number of distortions and artifacts. With increasing K, the target sketch becomes blurred, and it was, therefore, set to 20 in our experiments.

Fig. 6 shows the comparison between using fixed equal weights and learning region-adaptive weights of multiple representations for synthesizing sketches. The usage of fixed equal weights ignores the diversity of the multiple representations and cannot explore the information in an image patch sufficiently. With the help of learning region-adaptive weights, the proposed method can overcome the influence of lighting and background differences between the training and the test images. We performed face-recognition experiments on the CUHK database to justify the contribution of region adaptive weights in compared with fixed equal weights. The same protocol, as used in [3] and [5], was applied, and the Fisherface [49] was taken as the classifier. The proposed region adaptive weights-based method achieved a rank-1 accuracy of 98.3% on synthesized sketches, while a rank-1 accuracy of 95.7% was achieved using fixed equal weights. We further performed face-recognition experiments on forensic sketch database here. In order to keep the same division ratio with [17], we randomly selected 114 synthesized results to train the random sampling linear discriminant analysis (LDA) (RS-LDA) classifier, and the rest 57 pairs were taken as the test data together with 10 000 face photos from Labeled Faces in the Wild-a (LFW-a) to augment the scale of the gallery. The experiments were conducted 10 times, and the average accuracies are listed here. The proposed region-adaptive weights-based method achieved a rank-50 accuracy of 62.81%, while the method using fixed equal weights achieved a rank-50 accuracy of 53.51%. Both the visual comparison in Fig. 6 and the face-recognition experiments show that the region-adaptive weights-based method outperforms the fixed weight-based method.

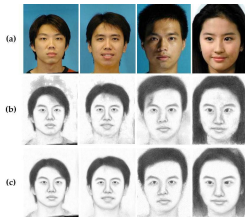


Fig. 6. Comparison between using fixed equal weights and learning region-adaptive weights of multiple representations for synthesized sketches. (a) Input photos. (b) Results of using fixed equal weights. (c) Results of the proposed method.

The most time-consuming part of our proposed method lies in three phases: 1) the feature extraction phase; 2) the neighbor-searching phase; and 3) the optimization phase. The operations in filtering and feature extraction on the training set can be finished offline. When we input a test photo, it takes <1 min to extract the multiple representations from this test photo. Around the search region for given input photo patch, we first find the best match patch from each training photo. Then, we select top K most similar photo patches and the corresponding K sketch patches in the training set as the candidates. The complexity of this process is $O(cMN\sum_{l=1}^{L} p_l)$. Here, c is the number of candidates in the search region around one patch. M is the number of patches on each image, and N is the number of sketch–photo pairs in the training set. L is the number of the multiple representations, and p_l is the dimension of the lth feature. The optimization phase mainly depends on the number of iterations. When the iteration number is five, it takes ~10 min to synthesize a sketch by our method running on an Intel Core i5-3470 3.20 GHz

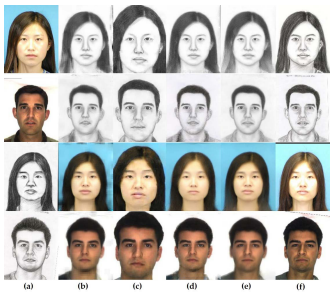


Fig. 7. Comparison of the proposed method with the LLE-based method [2], the MRF method [3], and the MWF method [4] on the CUHK database. Top two rows: results of synthesized sketches. Bottom two rows: results of synthesized photos. (a) Input images. (b) Results of the LLE-based method. (c) Results of the MRF method. (d) Results of the MWF method. (e) Results of the proposed method. (f) Ground-truth images.

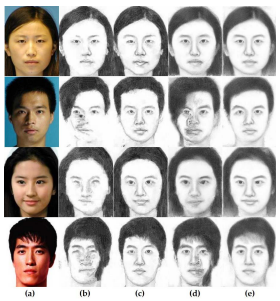


Fig. 8. Synthesized results of photos under different lighting conditions and celebrity photos. Top two rows: synthesized results of photos under different lighting conditions. Bottom two rows: results of celebrity photos. (a) Input photos. (b) Results of the MRF method [3]. (c) Results of the method in [20]. (d) Results of the MWF method. (e) Results of the proposed method.

PC under MATLAB R2012b environment, when the multiple representations are extracted offline.

B. Face Sketch–Photo Synthesis

Face sketch–photo synthesis was performed on the three databases separately. On the CUHK student database, 88 photo–sketch pairs are selected as the training data, and the remaining 100 photo–sketch pairs are the test data. On the AR database, we randomly select 23 photo–sketch pairs as the test data, and the remaining 100 photo–sketch pairs are set as the training data. This is repeated until the entire photo–sketch pairs are selected once as the test data. It is similar to the leave-one-out strategy with 23 pairs left out. On the XM2VTS database, we select 100 photo–sketch pairs as the training data, and the remaining 195 photo–sketch pairs are the test data.

Fig. 7 shows a comparison of the proposed method with the LLE-based method [2], the MRF method [3], and the MWF method [4] for synthesizing sketches and photos.[3] The sketches and photos synthesized by MWF originate from the authors, while the results of the LLE-based method are from our own implementations. Since the LLE-based method neglects the relationships between neighborhoods and synthesizes each image patch independently, there may be incompatibility. The MRF method [3] finds an optimal image patch in the training database to synthesize the target image patch, which can result in deformation around the mouth, eyes, and chin, because these parts of the test faces may be distinct from the training images and thus the MRF cannot find an appropriate image patch. Our method partially overcomes this effect, because the candidate image patches are linearly combined, which generates image patches that do not exist in the training data set. The MWF method only considers single representation information, which may have resulted in loss of facial structure and distortion, especially when the input photo was taken under different lighting conditions in the training data set. By taking advantage of multiple representations, the proposed method synthesizes facial structures very well. This is because our proposed method adaptively learns the weights of the multiple representations to measure the similarity between two image patches and can obtain the most suitable combination of candidate patches.

We next performed sketch-synthesis experiments on photos captured under different lighting conditions [20] (different frontal lighting and different side lighting conditions). To make accurate comparisons and to validate the benefits of using multiple representations, the proposed approach was also performed using only a single representation (the image patch intensities from original image), i.e., the MWF method [4]. Examples of these synthesis results are shown in top two rows of Fig. 8. MRF with luminance remapping [3] and MWF poorly synthesize photos under different lighting conditions, and the images are of poor quality and contain incorrect blocks. Although the method presented in [20] overcomes variable lighting conditions to some extent, our results look more natural, perhaps because the method in [20] selects the best candidate for the input test patch. It is sometimes difficult to find suitable patches in the training data, especially when the training data set is small. Using a weighted combination of candidate patches, our method synthesizes new patches that do not exist in the training data set to represent the input test photo patches. In addition, the method in [20] defines the local evidence function and the neighboring compatibility function with specified balance weights, while our method adaptively learns these weights to combine multiple representations from different locations, which is helpful when synthesizing sketches.

We next tested our method on photos of Chinese celebrities, which vary in lighting conditions and backgrounds between the test and the training sets. Examples of the synthesis results are shown in bottom two rows of Fig. 8. Our method outperforms the other methods, because it utilizes multiple representations to adaptively measure the distance between the image patches and it effectively decreases lighting and background differences between the training and the test data.

We further performed face sketch–photo synthesis on the CUFSF [39]. A total of 250 sketch–photo pairs are randomly selected as the training data, and the rest 944 sketch–photo pairs are used for testing. Fig. 9 shows a comparison of the proposed method with the transductive method in [5].

[3]All sketches and photos for the MRF method [3] are download from the Web Site: http://www.ee.cuhk.edu.hk/~xgwang/sketch_multiscale.html.

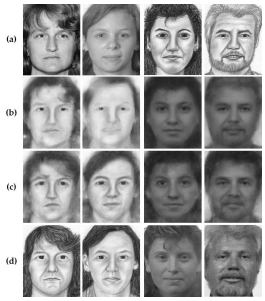


Fig. 9. Comparison of the proposed method with the transductive method [5] on the CUHK FERET database. Left two columns: results of synthesized sketches. Right two columns: synthesized photos. (a) Input images. (b) Results of the transductive method. (c) Results of the proposed method. (d) Ground-truth images.

Fig. 10. Example results on the IIIT-D Viewed Sketch Database. Left two columns: results of synthesized sketches. Right two columns: synthesized photos. (a) Input images. (b) Results of the proposed method. (c) Ground-truth images.

Utilizing multiple representations to overcome the influence of complex lighting conditions, the proposed method achieves better results on this challenge database.

In addition, we conducted some experiments on IIIT-D Viewed Sketch Database [44]. There are 238 sketch–photo pairs, where the photos are collected from different sources. We randomly select 119 photo-sketch pairs as the training data, and the rest 119 photo–sketch pairs are taken as the test data. Then, we switch the role of the training data and the test data aforementioned, and the entire photo–sketch pairs are selected once as the test data. Some example results are shown in Fig. 10. The proposed method achieves promising results on this challenge database.

C. Cross-Database Evaluation

Cross-database evaluations were conducted to verify whether the proposed method is sensitive to different databases. The photos in the different databases are variable in lighting, background, and ethnic origin of the subject (Fig. 4). Taking the 88 sketch–photo pairs from CUHK student database as the training data, our method was performed on the AR and XM2VTS databases for face-sketch-synthesis experiments. To do the cross-database bidirectionally, we performed face-photo-synthesis experiments with input sketches from the CUFSF [39] and the Sketch Face Database [44]. For equitable comparison, the analysis was performed using a single representation (image patch intensities from the original image), the same as in the MWF method [4]. As shown in Fig. 11, the results of cross-database evaluation experiments are promising, suggesting that it is efficient and generalizable. We further constructed a mixed database containing 100 photo–sketch pairs (34 from the CUHK student database, 33 from the AR database, and 33 from the XM2VTS database) as training data and the remainder as test data. The synthesis results are shown in Fig. 12; it can be seen that the proposed method is, to some extent, database-independent, i.e., once a training database is selected, the proposed algorithm can be applied to images in other databases. This is particularly useful for law enforcement, since when local police obtain a face sketch of a foreign suspect, the proposed method can be applied to synthesize a corresponding photo based on the local gallery.

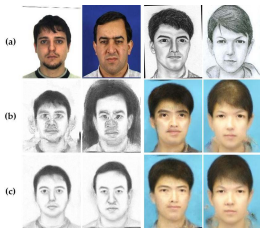


Fig. 11. Synthesized results of cross-database evaluation trained on the CUHK student database. Left two columns: results of synthesized sketches. Right two columns: synthesized photos. (a) Input photos. (b) Results of the MWF method. (c) Results of the proposed method.

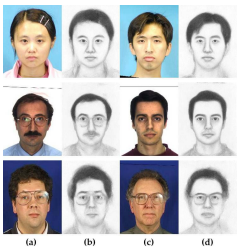


Fig. 12. Synthesized results of cross-database evaluation on the mixed database. (a) and (c) Input photos. (b) and (d) Results of the proposed method.

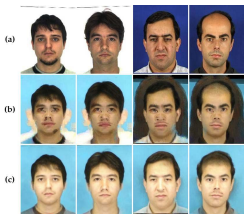


Fig. 13. Synthesized results of database-dependent style synthesis on the CUHK database. (a) Input photos. (b) Results of the MWF method. (c) Results of the proposed method.

D. Database-Dependent Style Synthesis

Noticing that the photos in different databases have variable lighting conditions, backgrounds, and skin colors, we performed database-dependent style synthesis using our method by replacing the training sketch data set with the training photo data set in the sketch-synthesis process. The same 88 CUHK student database photos were assigned as the training data, and the AR database, XM2VTS database, and Chinese celebrity photos were used as the test data. Our method was used for synthesizing photos in the same style as those in the CUHK student database. As shown in Figs. 13 and 14, our method outperforms MWF, further confirming the robustness of the proposed method, which can be used to synthesize the photos of suspects in the same style as the police's photo gallery.

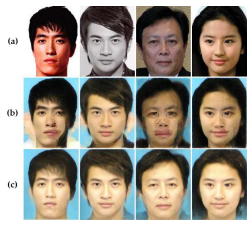


Fig. 14. Synthesized results of database-dependent style synthesis on celebrity photos. (a) Input photos. (b) Results of the MWF method. (c) Results of the proposed method.

TABLE I
AVERAGE FR-IQA VALUES USING THE DIFFERENT FACE SKETCH–PHOTO-SYNTHESIS METHODS AND THE IMAGE QUALITY ASSESSMENT METHODS ON THE CUHK DATABASE

	SSIM [50]	VIF [51]	FSIM [52]
Nonlinear approach_SS [2]	0.4811	0.1264	0.7064
Nonlinear approach_SP [2]	0.5753	0.1747	0.7831
MWF_SS [4]	0.4996	0.1298	0.7121
MWF_SP [4]	0.6057	0.1799	0.7996
MrFSPS_SS	0.5130	0.1402	0.7339
MrFSPS_SP	0.6326	0.2012	0.8031

E. Quantitative Evaluation

For quantitative evaluation, we performed full reference image quality assessment (FR-IQA) and face sketch recognition on the three databases.

At the FR-IQA stage, we conducted experiments on both the synthesized sketches and the synthesized photos. When conducting FR-IQA on the synthesized sketches, the artist-drawn sketches were used as reference images, and when conducting FR-IQA on the synthesized photos, the original photos were used as the reference images. We evaluated performance with three FR-IQA metrics: 1) the structural similarity (SSIM) index [50]; 2) the visual information fidelity (VIF) index [51]; and 3) the feature similarity (FSIM) index [52]. The FR-IQA values for synthesized sketches and synthesized photos using the nonlinear approach [2], MWF [4], and our method are shown in Fig. 15; they show the percentage of images, whose FR-IQA values are higher than the values on the horizontal axis. Our method achieves higher FR-IQA values than the other two methods. The average FR-IQA values of different sketch–photo-synthesis methods are compared in Table I, which shows that our method outperforms the others and achieves the highest average FR-IQA values for both the synthesized sketches and the synthesized photos.

For the face sketch recognition, we evaluated two strategies: 1) transforming all face photos in the gallery to sketches using our method—a query sketch was then matched with the synthesized sketches and 2) transforming a query sketch to a photo—the synthesized photo was then matched with the photo gallery. Three face-recognition methods were evaluated: 1) Fisherface [49]; 2) null-space LDA (NLDA) [53]; and 3) RS-LDA [54]. The 606 photo–sketch pairs were divided into three subsets, as in [3] and [5].

The rank-1 recognition accuracies using the sketch–photo-synthesis methods [3], [5], [11] and the mean accuracies and standard deviations of the proposed method after ten random splits are shown in Table II.[4] In the table, MMRF denotes the multiscale MRF method in [3], and TFSP denotes the transductive method in [5]. Our method achieves superior results to the other methods, and the proposed method combined with Fisherface achieves the highest recognition rate (98.3%). Since the proposed method is mainly used for image-synthesis issue, we simply used three classic-recognition methods (Fisherface, NLDA, and RS-LDA) for face sketch recognition to validate the quality of the synthesized results to some extent. The proposed method has the advantage that any existing face-recognition methods can be applied to the synthesized results for further improving the recognition performance. For example, we implemented a state-of-the-art face-recognition method [17] for face sketch recognition on the synthesized results of the CUHK database, and achieved a rank-1 accuracy of 99.45%, which is comparative to the state-of-the-arts (>99%). In contrast, directly matching original sketches with photos using [17] achieved a rank-1 accuracy of 95.14%. The proposed method is also preferred in the scenarios where the intermedia output (a synthesized sketch or photo) is required; for example, in entertainment applications, sketch portraits are usually taken as the avatar. Note that we also proposed an effective strategy for real-world forensic sketch–photo recognition in Section IV-F and achieved superior performance in comparison with the state-of-the-art matching strategies.

TABLE II
RANK-1 RECOGNITION ACCURACIES USING THE DIFFERENT FACE SKETCH–PHOTO-SYNTHESIS METHODS AND THE MEAN ACCURACIES ± STANDARD DEVIATIONS OF THE PROPOSED METHOD ON THE CUHK DATABASE (PERCENTAGE)

	Fisherface [49]	NLDA [53]	RS-LDA [54]
Eigentransformation[11]	79.7	84.0	90.0
MMRF_SS [3]	89.3	90.7	93.3
MMRF_SP [3]	93.3	94.7	96.3
TFSP_SS [5]	91.3	93.7	95.7
TFSP_SP [5]	96.3	96.3	97.7
MrFSPS_SS	98.3±0.46	97.7±0.89	97.3±0.57
MrFSPS_SP	97.7±0.70	96.7±0.51	97.7±0.92

The rank-1–rank-10 face-recognition accuracies of the proposed method are compared with four other methods [2], [3], [5], [11] in Table III. The multiscale MRF method in [3] had a first match rate of 96.3% and a tenth match rate of 99.7%, while the transductive method in [5] had a first match rate of 97.7% and a tenth match rate of 99.7%. Our method is highly promising, with a first match rate of 97.7% and a tenth match rate of 100%.

We conducted face-sketch-recognition experiments on the CUFSF database. The 1194 sketch–photo pairs in the CUFSF database were separated into three subsets: the 250 sketch–photo pairs were randomly selected for synthesis

[4]Since the methods [3], [5], [11] do not publish statistical results; here, we just provide the standard deviations of the proposed method.

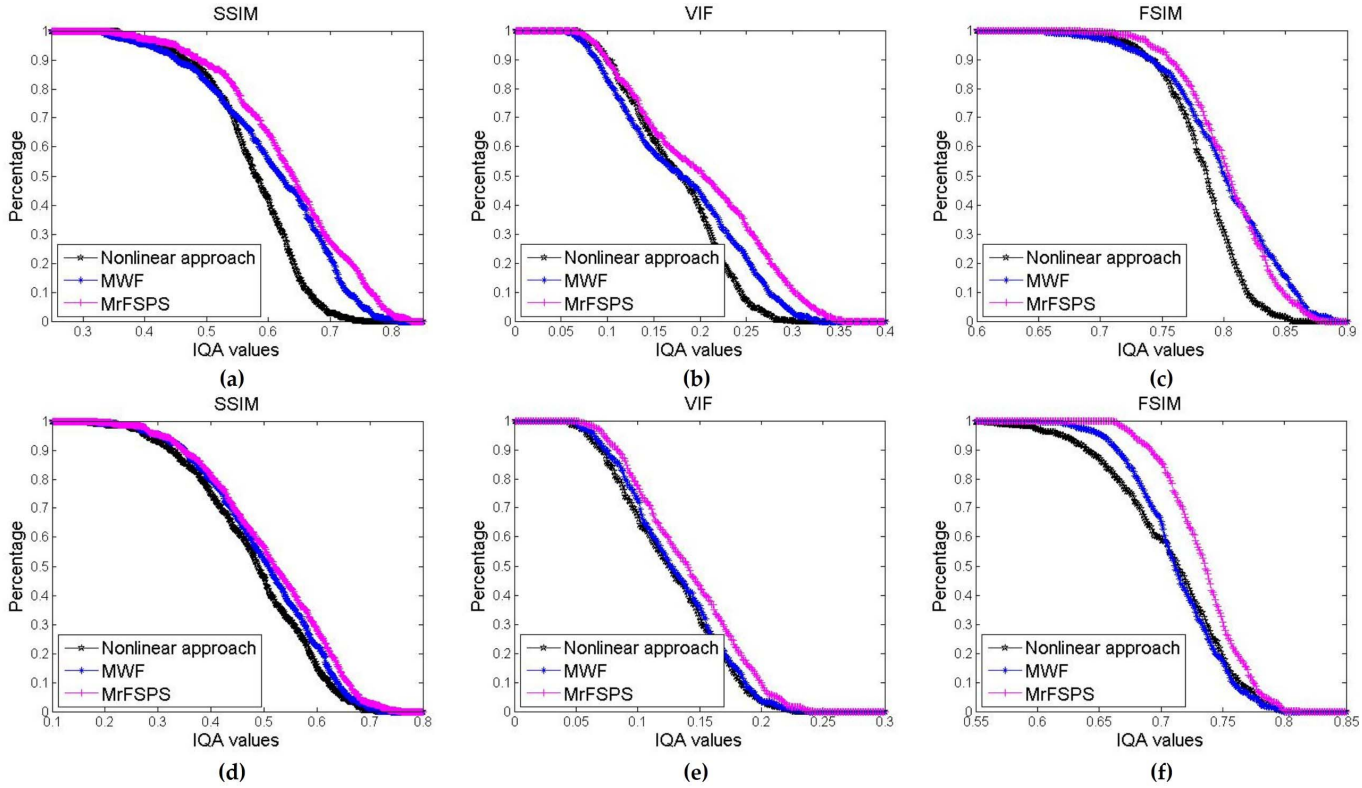


Fig. 15. FR-IQA values of synthesized results using different sketch–photo-synthesis methods on the CUHK database. (a) and (d) SSIM values of synthesized photos and sketches, respectively. (b) and (e) VIF values of synthesized photos and sketches, respectively. (c) and (f) FSIM values of synthesized photos and sketches, respectively.

TABLE III

RANK-1–RANK-10 ACCURACIES USING THE DIFFERENT FACE-SKETCH-RECOGNITION METHODS ON THE CUHK DATABASE (PERCENTAGE)

	1	2	3	4	5	6	7	8	9	10
Eigen transformation [11]	90.0	94.0	96.7	97.3	97.7	97.7	98.3	98.3	99.0	99.0
Nonlinear approach [2]	87.7	92.0	95.0	97.3	97.7	98.3	98.7	99.0	99.0	99.0
MMRF_SS + RS-LDA [3]	93.3	94.6	97.3	98.3	98.3	98.3	98.3	99.0	99.0	99.0
MMRF_SP + RS-LDA [3]	96.3	97.7	98.0	98.3	98.7	98.7	99.3	99.3	99.7	99.7
TFSP_SS + RS-LDA [5]	95.7	97.3	98.0	98.3	99.0	99.0	99.0	99.0	99.3	99.3
TFSP_SP + RS-LDA [5]	97.7	98.0	98.3	98.7	98.7	99.0	99.0	99.7	99.7	99.7
MrFSPS_SS + RS-LDA	97.3	98.7	98.7	98.7	98.7	98.7	99.0	99.0	99.0	99.0
MrFSPS_SP + RS-LDA	97.7	98.3	98.7	98.7	99.7	99.7	99.7	100	100	100

training, the 250 sketch–photo pairs were randomly selected to train the RS-LDA classifier [54], and the rest 694 sketch–photo pairs were used for testing. The CUFSF database is much larger than CUHK database. The lighting variation in face photos and the shape exaggeration in sketches both increase the difficulty of face sketch–photo synthesis and recognition. The proposed method utilizes multiple representations to decrease the influence of lighting variation. Therefore, we achieved an accuracy of 75.36% (with a standard deviation 0.0256) using the synthesized sketches and 59.37% (with a standard deviation 0.0287) using the synthesized photos, which outperformed the transductive method [5] (72.62% and 43.7% separately).

We conducted face-sketch-recognition experiments on the IIIT-D Viewed Sketch Database. A total of 95 synthesized sketch–photo pairs were selected to train the RS-LDA classifier [54], and the rest 143 pairs are set as the test data. The sketches in this database are drawn by an artist for photos collected from the face and gesture recognition research network (FG-NET) aging database, the LFW database, and the IIIT-D student and staff database. The recognition performance of the proposed method is 80.15% (with a stan-

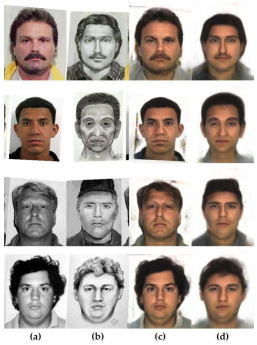


Fig. 16. Synthesized results of the forensic sketches and mug shot photos on the forensic sketch database, taking the CUHK AR database as the training data. (a) Mug shot photos. (b) Forensic sketches. (c) Synthesized photos by taking mug shot photos in (a) as the test images. (d) Synthesized photos by taking forensic sketches in (b) as the test images.

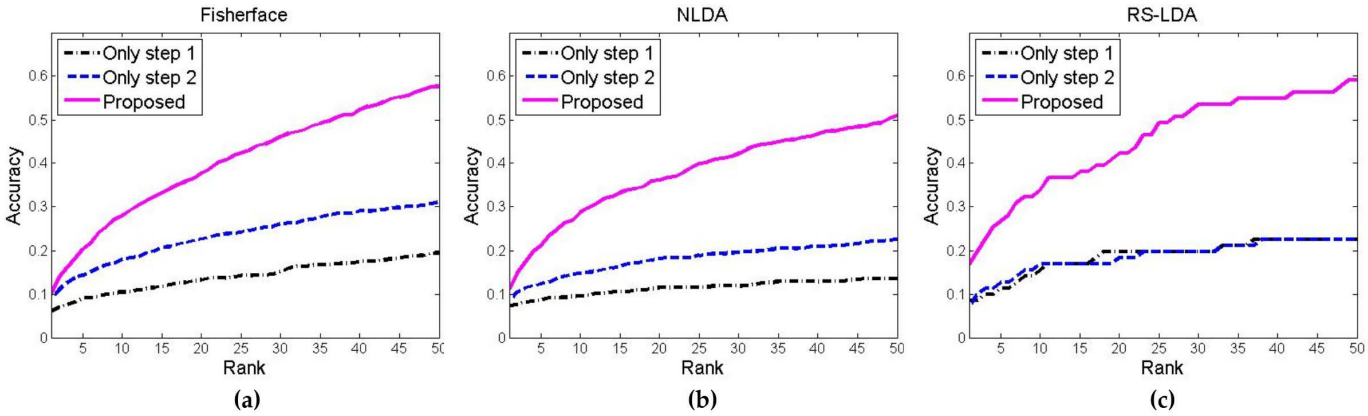


Fig. 17. Cumulative match score results of the proposed three-step strategy compared with using only Step 1 or Step 2 on the forensic sketch database using (a) Fisherface as the classifier, (b) NLDA as the classifier, and (c) RS-LDA as the classifier.

dard deviation 0.0152) and 78.06% (with a standard deviation 0.0180), using synthesized photos and synthesized sketches separately. Bhatt *et al.* [44] reported an accuracy of 84.24% utilizing modified Weber's local descriptor and memetic optimization.

F. Forensic Sketch–Photo Synthesis and Recognition

Matching forensic sketches to mug shots is difficult, because the forensic sketches are drawn by the police artist according to the descriptions of eyewitnesses or victims; an eyewitness's face perception and the artist's perceptual experience when drawing the details of a sketch lead to differences between the forensic sketches and the mug shots. In this paper, we proposed a three-step strategy to perform face recognition between the forensic sketches and the mug shots in law enforcement. For instance, we take the CUHK AR database, including 123 photo–sketch pairs as the training data. In Step 1, in order to decrease the influence of different lighting, pose, and background conditions, the photos in mugshot databases can be transformed into synthesized photos by performing the aforementioned database-dependent style synthesis. The mug shots in the gallery are transformed into the same style as the photos in the AR database. In Step 2, the law enforcement agencies generate a forensic sketch according to the eyewitness's description, and the forensic sketch can be transformed into synthesized photos by the proposed MrFSPS method. In Step 3, the face-recognition methods (Fisherface [49], NLDA [53], and RS-LDA [54]) can be applied to perform face recognition on the photos synthesized in Step 1 and Step 2. We have added 10 000 face photo images from the LFW-a data set [55] to increase the size of the gallery. Forensic sketch to mug shots matching experiments with a large gallery can present results more close to real-world face retrieval in law enforcement agencies.

Fig. 16 shows the synthesis results for Step 1 and Step 2. These photos further demonstrate the effectiveness and robustness of the proposed method. Although the mug shots and the forensic sketches are in different modalities, and have different lighting and pose conditions and backgrounds, after transformation to the same modality, they are in the same style as the photos in the AR database. This is likely to facilitate suspect identification.

We randomly selected 100 synthesized photo pairs in Step 1 and Step 2 to train the classifiers, and the rest together with 10 000 face photos from the LFW-a were taken as the test data. Each of the following experiments was conducted ten times, and the average accuracies are demonstrated here. To evaluate the advantage of the proposed three-step strategy, we compared the proposed strategy using only Step 1 or Step 2. As shown in Fig. 17, the proposed three-step strategy performs best in all the three face-recognition methods. The recognition performance proved the effectiveness of the three-step strategy, and it can be seen that the proposed three-step strategy with RS-LDA as the classifier achieves the best matching performance. Thus, we utilize RS-LDA as the classifier and compared the proposed strategy with three recent methods [17], [56], [57]. The feature-based heterogeneous face-recognition method in [17] used 106 subjects of forensic sketches and mug shot photos for training and 53 subjects plus 10 000 mug shot images for testing. They achieved a rank-50 accuracy of 17.4%. We used the same division ratio with [17]. A total of 114 synthesized results were randomly selected to train the RS-LDA classifier, and the rest 57 pairs together with 10 000 face photos from the LFW-a were taken as the test data. Our proposed method achieves a rank-50 accuracy of 62.81% under this protocol. The method in [56] utilized modified Weber's local descriptor and memetic optimization. They used 140 sketch–photo pairs from the IIIT-D Semiforensic Sketch database for training, and 190 forensic sketches were used as probe. A total of 599 face images plus 6324 photos were used as gallery. They achieved a rank-50 accuracy of 28.52%. We performed face sketch–photo synthesis on the IIIT-D Semiforensic Sketch database, and the synthesized results were used to train the classifier. The same protocol, as in [56], was followed, and our method achieves a rank-50 accuracy of 37.13%. The component-based approach in [57] used 123 composite sketches as the probe set and 123 photos from the CUHK AR database plus 10 000 mug shots were used as the gallery. Note that the methods in [57] and [58] conducted composite sketch to photo matching, which is quite different from the forensic sketch to the photo matching. The 123 composite sketches generated using Identi-Kit [59] are available in the PRIP Viewed Software-Generated Composite (PRIP-VSGC) database [60]. We per-

TABLE IV
RANK-50 ACCURACIES USING THE DIFFERENT FACE-SKETCH-RECOGNITION METHODS AND THE MEAN ACCURACIES ± STANDARD DEVIATIONS OF THE PROPOSED METHOD (PERCENTAGE)

(a) Comparison Methods	Results of (a)	Results of MrFSPS
P-RS [17]	17.4	62.81±5.47
MCWLD [56]	28.52	37.13±2.09
Component-based [57]	<5	21.14

formed face sketch–photo synthesis on these 123 composite sketches. The 123 synthesized results of CUHK AR database were used to train the classifier, and the 123 synthesized results of the composite sketches were taken as the test set together with 10 000 face photos from the LFW-a used to augment the scale of the gallery. Our method achieved a rank-50 accuracy of 21.14% on the composite sketch database generated by Identi-Kit. The method in [57] only conducted composite sketch to photo matching of different facial components on the composite sketches generated using Identi-Kit. All the rank-50 accuracies of different facial components in [57] were lower than 5%. Our proposed method outperforms existing methods following the same protocols. The mean accuracies and standard deviations achieved by our method are shown in Table IV. Because the training and testing sets are fixed on the PRIP-VSGC database, there is no standard deviation on this composite sketch database. A limitation of our study is the small size of the forensic sketch database; it is unfortunately not easy to obtain a large number of the forensic sketches and the corresponding mug shot photos. While this does not influence the synthesis stage, it does limit the training of the classifiers during recognition. We believe that with a larger number of forensic sketch–photo pairs, the recognition performance could be further improved.

V. CONCLUSION

In this paper, we proposed a novel MrFSPS method. Unlike existing methods, our method adaptively combines multiple representations to represent an image patch. To effectively and efficiently combine these features for face sketch–photo synthesis, we developed a Markov networks-based framework to adaptively learn the weights of multiple representations and candidate patches for the target image patch, and used an alternating optimization strategy to optimize the proposed framework. Our experimental results showed that the proposed method outperforms existing synthesis methods. In addition, the method achieves superior face recognition and image quality assessment performance on multiple databases. Using cross-database evaluation and database-dependent style synthesis, we demonstrate that our method is database-independent. The proposed database-dependent style-synthesis strategy was also used to perform forensic sketch-based face recognition, with promising results.

ACKNOWLEDGMENT

The authors would like to thank the Editor-In-Chief Prof. D. Liu, the handling Associate Editor, and all anonymous reviewers for their constructive and helpful comments and suggestions.

REFERENCES

[1] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *Int. J. Comput. Vis.*, vol. 106, no. 1, pp. 9–30, 2014.
[2] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 1005–1010.
[3] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.
[4] H. Zhou, Z. Kuang, and K.-Y. K. Wong, "Markov weight fields for face sketch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 1091–1097.
[5] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "Transductive face sketch-photo synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 9, pp. 1364–1376, Sep. 2013.
[6] L. Gibson, *Forensic Art Essentials: A Manual for Law Enforcement Artists*. Waltham, MA, USA: Academic, 2010.
[7] R. G. Uhl, Jr., N. da Vitoria Lobo, and Y. H. Kwon, "Recognizing a facial image from a police sketch," in *Proc. 2nd IEEE Workshop Appl. Comput. Vis.*, Sarasota, FL, USA, Dec. 1994, pp. 129–137.
[8] R. G. Uhl, Jr., and N. da Vitoria Lobo, "A framework for recognizing a facial image from a police sketch," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 1996, pp. 586–593.
[9] H. Koshimizu, M. Tominaga, T. Fujiwara, and K. Murakami, "On KANSEI facial image processing for computerized facial caricaturing system PICASSO," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Tokyo, Japan, Oct. 1999, pp. 294–299.
[10] X. Tang and X. Wang, "Face sketch recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 50–57, Jan. 2004.
[11] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Nice, France, Oct. 2003, pp. 687–694.
[12] J. Yu, M. Wang, and D. Tao, "Semisupervised multiview distance metric learning for cartoon synthesis," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4636–4648, Nov. 2012.
[13] X. Wang, W. Bian, and D. Tao, "Grassmannian regularized structured multi-view embedding for image classification," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2646–2660, Jul. 2013.
[14] W. Liu and D. Tao, "Multiview Hessian regularization for image annotation," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2676–2687, Jul. 2013.
[15] C. Deng, R. Ji, D. Tao, X. Gao, and X. Li, "Weakly supervised multi-graph learning for robust image reranking," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 785–795, Apr. 2014.
[16] K. Zhang, D. Tao, X. Gao, X. Li, and Z. Xiong, "Learning multiple linear mappings for efficient single image super-resolution," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 846–861, Mar. 2015.
[17] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.
[18] H. Li, F. Zhang, and S. Zhang, "Multi-feature hierarchical topic models for human behavior recognition," *Sci. China Inf. Sci.*, vol. 57, no. 9, pp. 1–15, 2014.
[19] Y. Li, L. Meng, J. Feng, and J. Wu, "Downsampling sparse representation and discriminant information aided occluded face recognition," *Sci. China Inf. Sci.*, vol. 57, no. 3, pp. 1–8, 2014.
[20] W. Zhang, X. Wang, and X. Tang, "Lighting and pose robust face sketch synthesis," in *Proc. 11th Eur. Conf. Comput. Vis.*, Heraklion, Greece, Sep. 2010, pp. 420–433.
[21] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, Los Angeles, CA, USA, Aug. 2001, pp. 341–346.
[22] I. T. Jolliffe, *Principal Component Analysis*. Hoboken, NJ, USA: Wiley, 2002.
[23] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[24] X. Tang and X. Wang, "Face photo recognition using sketch," in *Proc. IEEE Int. Conf. Image Process.*, Rochester, NY, USA, Sep. 2002, pp. I-257–I-260.
[25] M. Chang, L. Zhou, Y. Han, and X. Deng, "Face sketch synthesis via sparse representation," in *Proc. 20th Int. Conf. Pattern Recognit.*, Istanbul, Turkey, Aug. 2010, pp. 2146–2149.
[26] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. 21st Annu. Conf. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2007, pp. 801–808.

[27] N. Wang, X. Gao, D. Tao, and X. Li, "Face sketch-photo synthesis under multi-dictionary sparse representation framework," in *Proc. 6th Int. Conf. Image Graph.*, Hefei, China, Aug. 2011, pp. 82–87.

[28] X. Gao, N. Wang, D. Tao, and X. Li, "Face sketch–photo synthesis and retrieval using sparse representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 8, pp. 1213–1226, Aug. 2012.

[29] L. Chang, M. Zhou, X. Deng, Z. Wu, and Y. Han, "Face sketch synthesis via multivariate output regression," in *Proc. 14th Int. Conf. Human-Comput. Interact., Design Develop. Approaches*, Orlando, FL, USA, Jul. 2011, pp. 555–561.

[30] J. Zhang, N. Wang, X. Gao, D. Tao, and X. Li, "Face sketch-photo synthesis based on support vector regression," in *Proc. 18th IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 1125–1128.

[31] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[32] 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.*, vol. 18, no. 4, pp. 487–496, Apr. 2008.

[33] B. Xiao, X. Gao, D. Tao, and X. Li, "A new approach for face recognition by sketches in photos," *Signal Process.*, vol. 89, no. 8, pp. 1576–1588, 2009.

[34] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[35] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[36] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[37] E. Meyers and L. Wolf, "Using biologically inspired features for face processing," *Int. J. Comput. Vis.*, vol. 76, no. 1, pp. 93–104, 2008.

[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 886–893.

[39] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 513–520.

[40] P. Dreuw, P. Steingrube, H. Hanselmann, and H. Ney, "SURF-face: Face recognition under viewpoint consistency constraints," in *Proc. Brit. Mach. Vis. Conf.*, London, U.K., Sep. 2009, pp. 1–11.

[41] J. Li and Y. Zhang, "Learning SURF cascade for fast and accurate object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3468–3475.

[42] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[43] B. Geng, D. Tao, C. Xu, L. Yang, and X.-S. Hua, "Ensemble manifold regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1227–1233, Jun. 2012.

[44] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetic approach for matching sketches with digital face images," Indraprastha Inst. Inf. Technol., New Delhi, India, Tech. Rep. TR-2011-006, Oct. 2011.

[45] A. Martínez and R. Benavente, "The AR face database," CVC, Barcelona, Spain, Tech. Rep. #24, Jun. 1998.

[46] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio Video-Based Biometric Person Authentication*, Washington, DC, USA, Apr. 1999, pp. 72–77.

[47] P. J. Phillips, H. Moon, P. J. Rauss, and S. A. Rizvi, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[48] K. T. Taylor, *Forensic Art and Illustration*. Boca Raton, FL, USA: CRC Press, 2000.

[49] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[51] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[52] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[53] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognit.*, vol. 33, no. 10, pp. 1713–1726, 2000.

[54] X. Wang and X. Tang, "Random sampling for subspace face recognition," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 91–104, 2006.

[55] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1978–1990, Oct. 2011.

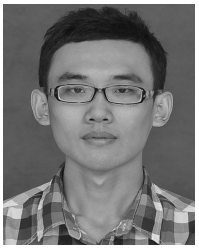
[56] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetically optimized MCWLD for matching sketches with digital face images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1522–1535, Oct. 2012.

[57] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching composite sketches to face photos: A component-based approach," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 1, pp. 191–204, Jan. 2013.

[58] S. J. Klum, H. Han, A. K. Jain, and B. F. Klare, "Sketch based face recognition: Forensic vs. composite sketches," in *Proc. Int. Conf. Biometrics*, Madrid, Spain, Jun. 2013, pp. 1–8.

[59] *Identi-Kit*. [Online]. Available: http://www.identikit.net/, accessed Aug. 12, 2015.

[60] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain, "The FaceSketchID system: Matching facial composites to mugshots," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2248–2263, Dec. 2014.



Chunlei Peng received the B.Sc. degree in electronic and information engineering from Xidian University, Xi'an, China, in 2012, where he is currently pursuing the Ph.D. degree in intelligent information processing with the VIPS Laboratory, School of Electronic Engineering.

His current research interests include computer vision, pattern recognition, and machine learning.

Xinbo Gao (M'02–SM'07) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively.

He was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan, from 1997 to 1998. From 2000 to 2001, he was a Post-Doctoral Research Fellow with the Department of Information Engineering, Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor with the Ministry of Education, a Professor of Pattern Recognition and Intelligent Systems, and the Director of the State Key Laboratory of Integrated Services Networks, Xi'an. He has authored five books and around 200 technical articles in refereed journals and proceedings. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications.

Prof. Gao is a fellow of the Institution of Engineering and Technology. He is on the Editorial Boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier). He served as the General Chair/Co-Chair, the Program Committee Chair/Co-Chair, or a PC Member for around 30 major international conferences.

Nannan Wang received the B.Sc. degree in information and computation science from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2009, and the Ph.D. degree in information and telecommunications engineering, in 2015.

He was a Visiting Ph.D. Student with the University of Technology at Sydney, Sydney, NSW, Australia, from 2011 to 2013. He is currently with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an. He has authored over ten papers in refereed journals and proceedings, including the *International Journal of Computer Vision*, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. His current research interests include computer vision, pattern recognition, and machine learning.

Dacheng Tao (F'15) is currently a Professor of Computer Science with the Centre for Quantum Computation and Intelligent Systems, and the Faculty of Engineering and Information Technology, University of Technology at Sydney, Sydney, NSW, Australia. He mainly applies statistics and mathematics to data analytics problems. His research results have expounded in one monograph and over 100 publications at prestigious journals and prominent conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the *Journal of Machine Learning Research*, the *International Journal of Computer Vision*, the Conference on Neural Information Processing Systems, the International Conference on Machine Learning, the Conference on Computer Vision and Pattern Recognition, the International Conference on Computer Vision, the European Conference on Computer Vision, the International Conference on Artificial Intelligence and Statistics, the International Conference on Data Mining (ICDM), and the ACM Special Interest Group on Knowledge Discovery and Data Mining. His current research interests include computer vision, data science, image processing, machine learning, and video surveillance.

Dr. Tao received several best paper awards, such as the Best Theory/Algorithm Paper Runner Up Award in the IEEE ICDM in 2007, the Best Student Paper Award in the IEEE ICDM in 2013, and the ICDM 10 Year Highest-Impact Paper Award in 2014.

Xuelong Li (M'02–SM'07–F'12) is currently with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), the State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.

Jie Li received the B.Sc. degree in electronics engineering, the M.Sc. degree in signal and information processing, and the Ph.D. degree in circuit and systems from Xidian University, Xi'an, China, in 1995, 1998, and 2004, respectively.

She is currently a Professor with the School of Electronic Engineering, Xidian University. She has authored around 50 technical articles in her research areas in refereed journals and proceedings, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *Information Sciences*. Her current research interests include image processing and machine learning.