Multi-View Representation Based Face Sketch Synthesis

Chunlei Peng[1,2], Jie Li[1,2], Nannan Wang[2], Xinbo Gao[2]
1 Xidian-Ningbo Information Technology Institute, Ningbo 315200, P.R. China
2 Xidian University, Xi'an 710071, P.R. China
clp.xidian@gmail.com, leejie@mail.xidian.edu.cn, nannanwang.xidian@gmail.com
xbgao@mail.xidian.edu.cn

ABSTRACT

Face photos and sketches are in two different modalities and synthesizing face sketches from photos plays an important role in law enforcement and digital entertainment. Although many face sketch synthesis methods have been proposed in recent years, most of these methods choose neighbor image patches based on image intensities. However, it is not appropriate to represent the similarity of two image patches merely according to image intensities, especially under uncontrolled environments such as varying illumination conditions. In this paper, we propose a multi-view representation of image patches for face sketch synthesis. It first constructs a high dimensional multi-view feature vector through a hierarchical framework including multiple filters and multiple local features. Then an unsupervised dimensionality reduction method is used to reduce the cost of computation and model storage. Finally traditional face sketch synthesis methods can be applied based on the proposed representation. Experimental results on the Chinese University of Hong Kong (CUHK) face sketch database and celebrity photos from the Internet illustrate that the proposed strategy improves the performance of the state-of-the-arts.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Application—*Computer vision*

General Terms

Algorithms

Keywords

Face sketch synthesis, image quality assessment, multi-view representation

1. INTRODUCTION

Face sketch synthesis has been extensively investigated in recent years [14]. It refers to synthesizing a face sketch according to the given photo. It has been widely applied in the circumstances in which only a modality (photo or sketch) of the face images is available. For example, when the police search for criminal suspects, a photo of the suspect is always unavailable, and the description of the eyewitness is the only clue. The artist can draw a sketch with the aid of eyewitness. However, directly conducting face recognition by face sketches on photo gallery is not feasible because they are in two different modalities with different geometry and texture characteristics. In order to reduce the discrepancies between face sketches and photos, face sketch synthesis can be applied to transform photos into the sketch modality. Subsequently traditional face recognition methods can be used to identify the suspect. In addition, face sketch synthesis can be applied in digital entertainment. For instance, people may choose to make a sketch as the profile in their social network websites such as Facebook and Twitter.

While a number of face sketch synthesis methods have been proposed [11, 6, 15, 19, 13, 12], most of them are constrained that the face images should be captured under controlled conditions, including frontal pose, normal illumination and neutral expression. Unfortunately, such constraints are very difficult to satisfy in practice. Most of the existing face sketch synthesis methods cannot work well on face images collected under uncontrolled environments.

In this paper, we illustrate that it is due to the mode of feature representation to calculate neighboring image patches that existing face sketch synthesis methods fail to work well on real world face photos. It is not appropriate to represent the similarity of two image patches merely according to image intensities, especially under uncontrolled environments. Recent literatures on face recognition have shown the effectiveness of low-level features (*e.g.*, Local Binary Patterns (LBP) [8], Scale-Invariant Feature Transform (SIFT) [7], and Speeded Up Robust Features (SURF) [1]) in recognizing face photos under lighting variations. Inspired by these observations, we propose a multi-view representation to measure the similarity of two image patches and apply it to traditional face sketch synthesis methods. The proposed strategy can be followed in three steps: firstly, a high dimensional multi-view feature vector is constructed through a hierarchical framework including multiple filters and multiple local features; subsequently, an unsupervised dimensionality reduction method is adopted to reduce the cost of computation and storage space with little accuracy sacrificed; finally, the generated multi-view feature representation can

ICIMCS'14, July 10–12, 2014, Xiamen, Fujian, China.

be combined into traditional face sketch synthesis methods. By switching the roles of sketches and photos, the face photo synthesis method induced aforementioned face sketch synthesis method. Therefore, we just introduce the procedure of face sketch synthesis from an input photo in the following sections.

The main contributions of this paper are:

- We propose an efficient and effective multi-view feature representation to measure the similarity of image patches;
- We apply the proposed approach in different traditional face sketch synthesis methods and demonstrate the effectiveness of the multi-view representation.

The rest of this paper is organized as follows: section 2 gives an introduction to related works and section 3 describes the proposed multi-view representation strategy. Experimental results and analysis are shown in section 4 and section 5 concludes this paper.

2. RELATED WORK

Tang and Wang [11] firstly proposed a linear subspace method based on Principal Component Analysis (PCA), which considers the sketch synthesis procedure as a linear process. Liu *et al.* [6] proposed a nonlinear face sketch synthesis algorithm by exploring local linear embedding (LLE) method [9]. By dividing face images into overlapping patches, it synthesizes a sketch patch instead of the whole sketch as in [11]. However, the method [6] synthesizes each image patch independently and does not consider the compatibility relationship between neighborhoods. Wang and Tang [15] introduced Markov Random Fields (MRF) to model the relationship of sketch-photo patch pairs as well as neighborhood relationships. Zhou *et al.* [19] proposed Markov Weight Fields (MWF), which is capable of synthesizing new sketch patches not existing in the training set. Considering that the weighted combination of the candidate patches may lose the high-frequency information, Wang *et al.* [3] proposed a two-step framework to enhance synthesized images. Wang *et al.* [13] introduced a novel transductive method to reduce the high losses resulted from inductive learning based synthesis methods.

Zhang *et al.* [18] extended the MRF model [15] by integrating shape prior information, local evidence and neighborhood compatibility into a new function, which is robust to lighting and pose variations. The robustness to lighting and pose variations in [18] is achieved by three parts: firstly, the shape prior information is introduced to reduce the distortions; secondly, the similarity of a photo patch and the corresponding sketch patch is considered by a dense SIFT descriptor; thirdly, the distance between dense SIFT descriptors of neighboring patches is explored to measure the gradient compatibility. Recently, Klare *et al.* [5] proposed a heterogeneous face recognition method which involved matching two face images from alternate imaging modalities, such as a photo to a viewed sketch (drawn according to a static photo) or a forensic sketch (drawn according to the description of a witness). They used three image filters and two local descriptors to represent a face image and conducted direct recognition of two face images in different modalities. It refers to image retrieval without face sketch synthesis in that paper. Considering the differences between heterogeneous face recognition and heterogeneous image transformation, the feature representation method in [5] cannot be directly used in face sketch synthesis, which may lead to the loss of face structures. Therefore, we improve the representation of image patches based on [5] to form a high dimensional feature vector. Recently, Chen *et al.* [2] revealed that the high-dimensional feature is crucial to the performance in the context of modern technology. In this paper, we propose to extract a multi-view high dimensional feature to represent an image patch and demonstrate the effectiveness of the multi-view representation through a number of experiments.

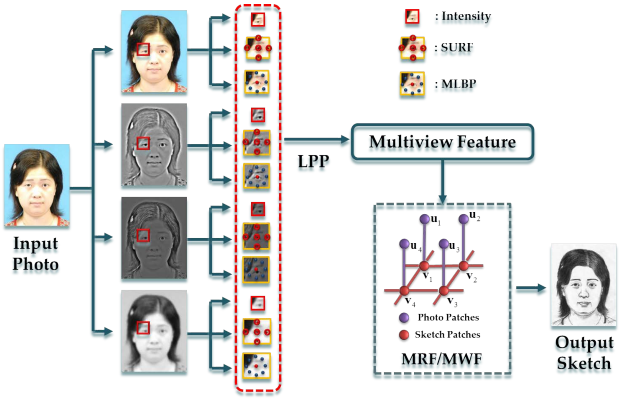


Figure 1: Framework of the proposed sketch synthesis method

3. MULTI-VIEW REPRESENTATION BASED METHOD

Considering the training data set with N face photo-sketch pairs denoted as $(\mathbf{p}^1, \mathbf{s}^1), \cdots, (\mathbf{p}^N, \mathbf{s}^N)$, we firstly divide each face image into M overlapping patches with the same patch size, *i.e.* photo patches $(\{\mathbf{p}_1^1, \cdots, \mathbf{p}_M^1\}, \cdots, \{\mathbf{p}_1^N, \cdots, \mathbf{p}_M^N\})$ and sketch patches $(\{\mathbf{s}_1^1, \cdots, \mathbf{s}_M^1\}, \cdots, \{\mathbf{s}_1^N, \cdots, \mathbf{s}_M^N\})$. The input test photo can also be divided into M overlapping patches. We then construct multi-view representation of each photo patch in the training set as well as the input test photo when synthesizing a face sketch from a photo.

The proposed framework consists of two components: constructing the multi-view representation and applying the generated multi-view feature to traditional face sketch synthesis methods, as shown in Fig. 1. We will discuss each part in the following two subsections, respectively.

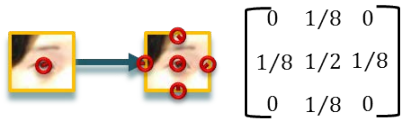
3.1 Constructing multi-view representation

The multi-view representation is constructed through a hierarchical framework. The input face image firstly passes through four filters in the filter level. We then divide the four filtered images into patches and extract three low-level descriptors for each image patch. Finally, the generated multiple features through multiple filters are concatenated to form a high dimensional feature vector to represent the input photo patch.

There are four different components in the filter level, including the original image, results of images filtered by difference of Gaussian (DoG) filter, Center-Surround Divisive Normalization (CSDN) filter and Gaussian smoothing filter respectively. A difference of Gaussian image filter is gener-

(a) (b) (c) (d)

Figure 2: Examples of a face photo (first row) and sketch (second row) after being filtered. (a) Original image; (b) Filtered by DoG filter; (c) Filtered by CSDN filter; (d) Filtered by Gaussian smoothing filter.



(a) (b)

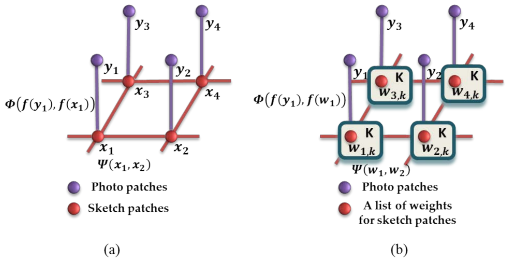
Figure 3: Illustration of the extracted SURF features with a template. (a) Illustration of the five SURF feature positions; (b) The weighting template.

ated by convolving the photo image with the difference of two Gaussian kernels with two standard deviations σ_0 and σ_1. In this paper, $\sigma_0 = 1$ and $\sigma_1 = 4$. The CSDN filter divides the intensity of each pixel by the mean of its neighborhood intensities. We set the neighborhood size $s = 16$. The Gaussian smoothing filter is designed with $\sigma = 2$.

In order to avoid the loss of facial structures during the synthesis procedure, we add the original image in the filter level. The DoG filter and CSDN filter are helpful for removing the effects of lighting variations as well as enhancing the facial details. The Gaussian smoothing filter is used to remove noise contained in high spatial frequencies. Fig. 2 demonstrates the filtered results of an example sketch-photo pair.

In the feature level, we deploy the image patch intensities together with two local feature descriptors to represent an image patch. The two local feature descriptors explored in this paper are Speeded Up Robust Features (SURF) [1] and Multi-scale Local Binary Pattern (MLBP) [8] since both of them have been shown to be very successful in describing face images in recent years. In this paper, instead of extracting only one SURF feature from an image patch, we exploit a weighting template to extract five positions' SURF features in order to obtain more information about the image patch, as shown in Fig. 3. MLBP is the concatenation of LBP feature descriptors with different radiuses. The radiuses in this paper are set as follows: $r = \{1, 2, 3, 4\}$.

The image patch intensities added in feature level is helpful to avoid the loss of face structures as well as improving the synthesized results. The SURF features obtained



(a) (b)

Figure 4: Graphical representations of multi-view representation based MRF (a) and multi-view representation based MWF (b).

through the weighting template can represent the structure in the image patch as well as the structure of the boundaries, which is crucial to the synthesis procedure. The MLBP feature is utilized to complement the multi-view representation.

For an image photo patch, we can generate twelve representations through the hierarchical framework. The simplest way to form the multi-view representation is to concatenate all the extracted features to form a high-dimensional feature and experimental results demonstrate that this simple strategy can perform well. Although the concatenated high dimensional multi-view vector can lead to better performance, it leads to a high cost of computation and storage requirements. Therefore, in this paper we use Locality Preserving Projections (LPP) [4] as the unsupervised dimensionality reduction method to map the high dimensional feature to a low dimensional subspace with a much lower computation and storage cost. The dimensionality of the final feature vector is set to 30 via cross validation. LPP as a common used unsupervised dimensionality reduction method is chosen in this paper and it can also be substituted by other unsupervised methods such as PCA. The generated low dimensional multi-view feature can then be applied in traditional face sketch synthesis methods.

3.2 Multi-view representation based face sketch synthesis

Let $\mathbf{y}_j$ be the photo patch intensities and $\mathbf{x}_j$ be the sketch patch intensities. The multi-view representation of photo patch $\mathbf{y}_j$ is represented as $\mathbf{f}(\mathbf{y}_j)$. The traditional face sketch synthesis methods (*e.g.*, MRF [15] and MWF [19]) can be modified by using the multi-view representation to measure the similarity of two photo patches, as shown in Fig. 4.

In the multi-view representation based MRF model (multi-view MRF), we compute the local evidence by the low dimensional multi-view representations:

$$\mathbf{\Phi}(\mathbf{f}(\mathbf{x}_j), \mathbf{f}(\mathbf{y}_j)) = exp\{\frac{-\|\mathbf{f}(\widetilde{\mathbf{y}}_j) - \mathbf{f}(\mathbf{y}_j)\|^2}{2\sigma_e^2}\} \quad (1)$$

where $\mathbf{f}(\widetilde{\mathbf{y}}_j)$ is the multi-view representation of the photo patch corresponding to the candidate sketch patch $\mathbf{x}_j$. The joint probability of multi-view MRF is given by

$$p(\mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{y}_1, \cdots, \mathbf{y}_N) = \prod_{j_1, j_2} \mathbf{\Psi}(\mathbf{x}_{j_1}, \mathbf{x}_{j_2}) \prod_j \mathbf{\Phi}(\mathbf{f}(\mathbf{x}_j), \mathbf{f}(\mathbf{y}_j)) \quad (2)$$

In the multi-view representation based MWF model (multi-

Figure 5: Synthesized sketches. (a) and (d) Input photos; (b) and (e) Results of the proposed multi-view MRF; (c) and (f) Results of the proposed multi-view MWF.

view MWF), the local evidence is given by

$$\mathbf{\Phi}(\mathbf{f}(\mathbf{w}_j), \mathbf{f}(\mathbf{y}_j)) = exp\{\frac{-\|\mathbf{f}(\mathbf{y}_j) - \sum_{k=1}^{K} w_{j,k}\mathbf{f}(\mathbf{p}_{j,k})\|^2}{2\sigma_D^2}\} \tag{3}$$

where $\mathbf{f}(\mathbf{p}_{j,k})$ is the multi-view representation of the kth candidate photo patch of $\mathbf{y}_j$. The joint probability of multi-view MWF is given by

$$p(\mathbf{w}_1, \cdots, \mathbf{w}_N, \mathbf{y}_1, \cdots, \mathbf{y}_N) = \prod_{j_1,j_2} \mathbf{\Psi}(\mathbf{w}_{j_1}, \mathbf{w}_{j_2}) \prod_j \mathbf{\Phi}(\mathbf{f}(\mathbf{w}_j), \mathbf{f}(\mathbf{y}_j)) \tag{4}$$

The above two optimization problems can be solved by belief propagation and quadratic programming methods respectively. The multi-view representation can also be applied to other face sketch synthesis methods [6, 13] in the similar way.

4. EXPERIMENTS

We conduct experiments on the CUHK database [15] which contains 188 photo-sketch pairs and celebrity photos [18]. In the experiments, we select 88 subjects as the training data and the rest 100 cases together with the celebrity photos as the testing data. The parameters are set as follows: the patch size is 10, the overlapping size is 5 and the neighborhood search region is 20×20. In multi-view MRF model we select $K = 15$ candidates for each test photo patch and in multi-view MWF model we set $K = 10$. When the neighborhood size K is small, it causes a great deal of distortions and artifaces. With the increase of K, the results become blurring. In dimension reduction procedure we select 5 nearest neighbors to construct the adjacency graph and then use heat kernel with $t = 5$ when choosing the weights in LPP.

4.1 Experiments on the CUHK Database

In this paper, we apply the proposed multi-view representation to two traditional face sketch synthesis methods: MRF [15] and MWF [19]. Examples of some synthesized sketches are shown in Fig. 5. It takes most of the time on extracting the multiple features. Once the features are extracted, it takes less than five minutes to synthesize one face sketch.

Fig. 6 evaluates the contribution of the original image in the filter level and the image patch intensities in the feature level. We take multi-view MRF results as examples. It can be seen that these two components are crucial to avoid the loss of facial image features. Fig. 7 evaluates the benefit of

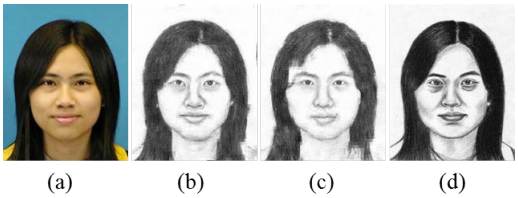


Figure 6: Effect of original image in filter level and patch intensities in feature level. (a) Input photos; (b) Results of the proposed multi-view MRF; (c) Results of multi-view MRF without the original image in filter level and patch intensities in feature level; (d) Sketches drawn by the artist.

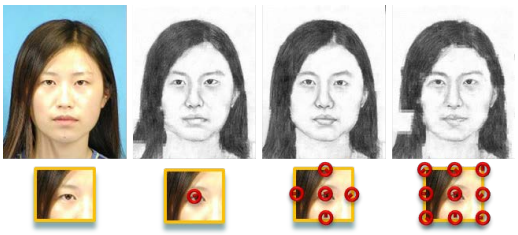


Figure 7: Comparison of three SURF feature extraction strategies. First row: input photos and synthesized results. Second row: graphical illustration of different SURF feature extraction strategies.

the weighting template used in extracting SURF features. We compare the result of the proposed weighting template with the result of another template in which nine positions' SURF features are extracted. From the multi-view MRF results we can see that the proposed SURF extraction strategy is the most effective way to describe the structure of an image patch.

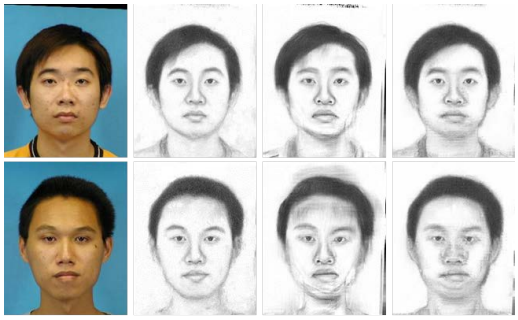
In Fig. 8, we compare our multi-view representation based MRF method with traditional MRF method [15]. It can be seen that with the assistance of the multi-view feature, our method can synthesize facial structures very well. One reason is that the multi-view feature can reduce the influence of different lighting conditions and can find more suitable neighboring patches from the training data. The other reason is that the proposed multi-view representation can measure the similarity of two image patches better and then contribute to the synthesized results. Fig. 9 shows the comparison of multi-view representation based MWF with the LLE-based method [6] and the traditional MWF method [19]. The LLE-based method neglected the relationship of neighborhoods and synthesized each image patch independently, which may result in incompatibility. The traditional MWF method still results in some losses of face structures and a great deal of distortions when the input test photo is under a different lighting condition.

By switching the roles of photos and sketches, our methods can also be used to synthesize photos from sketches. Fig. 10 shows some photo examples generated by multi-view representation based MWF and the traditional MWF [19]. It can be seen that the multi-view based method can generate better synthesized photos as well.

To give a quantitative evaluation, we compare the results

(a) (b) (c)

Figure 8: Comparison of the proposed multi-view MRF and the traditional MRF. (a) Input photos; (b) Results of the proposed multi-view MRF; (c) Results of MRF [15].



(a) (b) (c) (d)

Figure 9: Comparison of the proposed multi-view MWF with the LLE-based method and the traditional MWF. (a) Input photos; (b) Results of the proposed multi-view MWF; (c) Results of the LLE-based method [6]; (d) Results of MWF [19].

(a) (b) (c)

Figure 10: Comparison of the proposed multi-view MWF and the traditional MWF. (a) Input sketches; (b) Results of the proposed multi-view MWF; (c) Results of MWF [19].

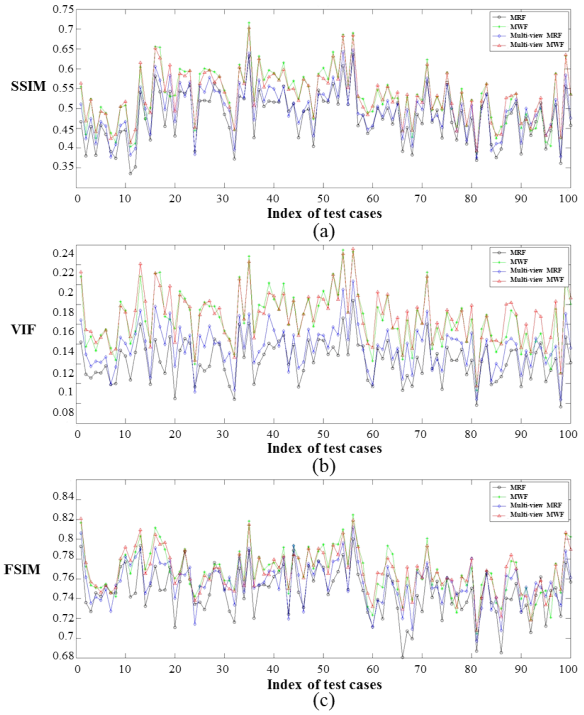


Figure 11: FR IQA values of the proposed multi-view MRF (colored blue), the proposed multi-view MWF (colored red), MRF [15] (colored black) and MWF [19] (colored green). (a) IQA values of SSIM [16] metric; (b) IQA values of VIF [10] metric; (c) IQA values of FSIM [17] metric.

Table 1: The comparison of average FR IQA values of two proposed methods, MRF [15] and MWF [19].

	SSIM [16]	VIF [10]	FSIM [17]
MRF [15]	0.4770	0.1234	0.7268
Multi-view MRF	**0.4949**	**0.1365**	**0.7362**
MWF [19]	0.5302	0.1649	0.7469
Multi-view MWF	**0.5309**	**0.1680**	**0.7471**

synthesized by two proposed methods with traditional MRF [15] and MWF [19] using three full reference (FR) image quality assessment (IQA) metrics, *i.e.* the Structural SIMilarity index (SSIM) [16], the Visual Information Fidelity index (VIF) [10] and the Feature SIMilarity index (FSIM) [17]. The sketches drawn by the artist are used as the reference images. Fig. 11 shows the FR IQA values of the 100 test cases. The average values are shown in Table 1. It can be seen that the FR IQA values of our proposed methods beat the other two methods. In addition, we have conducted face recognition experiments and these four methods achieve similar performance.

4.2 Experiments on the celebrity photos

We also test the proposed multi-view representation based face sketch synthesis methods on some photos of Chinese celebrities obtained from the Internet. These celebrity photos are quite different from the training images in lighting condition and backgrounds. As shown in Fig. 12, we com-

Figure 12: Synthesized results for celebrity photos. (a) Input photos; (b) Results of the proposed multi-view MRF; (c) Results of the proposed multi-view MWF; (d) Results of MRF [15]; (e) Results of the method [18].

pare the results of our proposed methods with the traditional MRF as well as the method proposed by Zhang et.al [18]. It can be seen that the traditional MRF method cannot synthesize well under this uncontrolled condition. The method in [18] can overcome the uncontrolled condition to some extent. However, the results of the proposed multi-view representation based methods look more natural. And more importantly, the proposed multi-view representation strategy can be widely applied to a variety of face sketch synthesis methods, face image super-resolution methods, facial animation methods, and so on.

5. CONCLUSION

In this paper, we proposed a multi-view representation of face image patches and combine it into two state-of-the-art face sketch synthesis methods. The multi-view representation is constructed through a hierarchical framework including multiple filters and multiple features. The Locality Preserving Projections algorithm is then used to help reduce the cost of computation and storage requirements. Finally we conducted two multi-view representation based face sketch synthesis methods. Experimental results showed that the proposed multi-view representation can achieve better synthesized results than traditional methods from both perceived quality (subjective) and image quality assessment values (objective). In the future, we will focus on the relationship of these multiview features because they may not just have equal weights. We would like to explore a more reasonable strategy to fuse these features and further improve the performance of the proposed multiview framework.

6. ACKNOWLEDGMENTS

We want to thank the helpful comments and suggestions from the anonymous reviewers. This research was supported partially by the Seed Fund from Xidian-Ningbo Information Technology Institute, the National Natural Science Foundation of China (Grant Nos. 61125204 and 61172146), the Fundamental Research Funds for the Central Universities (Grant Nos. K5051202048, BDZ021403 and JB149901), the Program for Changjiang Scholars and Innovative Research Team in University of China (No.IRT13088) and the Shaanxi Innovative Research Team for Key Science and Technology (No.2012KCT-02).

7. REFERENCES

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Gool. SURF: speeded up robust features. *CVIU*, 110:346–359, 2008.

[2] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: high-dimensional feature and its efficient compression for face verifcation. In *CVPR*, pages 3025–3032, 2013.

[3] X. Gao, N. Wang, D. Tao, and X. Li. Face sketch-photo synthesis and retrieval using sparse representation. *IEEE TCSVT*, 22:1213–1226, 2012.

[4] X. He and P. Niyogi. Locality preseerving projections. In *NIPS*, pages 1–8, 2003.

[5] B. Klare and A. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE TPAMI*, 35:1410–1422, 2013.

[6] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma. A nonlinear approach for face sketch synthesis and recognition. In *CVPR*, pages 1005–1010, 2005.

[7] D. Lowe. Distinctive image features from scale-invariant key-points. *IJCV*, 60:91–110, 2004.

[8] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE TPAMI*, 24:971–987, 2002.

[9] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[10] H. Sheikh and A. Bovik. Image information and visual quality. *IEEE TIP*, 15:430–444, 2006.

[11] X. Tang and X. Wang. Face photo recognition using sketches. In *ICIP*, pages 257–260, 2002.

[12] N. Wang, J. Li, D. Tao, X. Li, and X. Gao. Heterogeneous image transformation. *PRL*, 34:77–84, 2013.

[13] N. Wang, D. Tao, X. Gao, X. Li, and J. Li. Transductive face sketch-photo synthesis. *IEEE TNNLS*, 24:1364–1376, 2013.

[14] N. Wang, D. Tao, X. Gao, X. Li, and J. Li. A comprehensive survey to face hallucination. *IJCV*, 106:9–30, 2014.

[15] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE TPAMI*, 31:1955–1967, 2009.

[16] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13:600–612, 2004.

[17] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE TIP*, 20:2378–2386, 2011.

[18] W. Zhang, X. Wang, and X. Tang. Lighting and pose robust face sketch synthesis. In *ECCV*, pages 420–423, 2010.

[19] H. Zhou, Z. Kuang, and K. Wong. Markov weight fields for face sketch synthesis. In *CVPR*, pages 1091–1097, 2012.