MULTI-VIEW FACE HALLUCINATION BASED ON SPARSE REPRESENTATION

Zhuo Hui, Kin-Man Lam

Centre for Signal Processing, Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, Kowloon, Hong Kong

ABSTRACT

In this paper, we propose a novel method to generate the hallucinated multi-views of faces using the sparse-representation model. In order to render a faithful virtual view, we introduce centralized constraints into a variation framework for optimization. The constraints are formulated based on an attempt to minimize the difference between the sparse-coding coefficients derived for two distinct views. In our algorithm, sift optical-flow method is employed to formulate the constraints. An input face is firstly sparsely coded over a given dictionary, and then the sparse-coding coefficients for the input face are refined through an optimization framework with the centralized constraints. Intensive experimental results demonstrate that our proposed method can perform well in terms of both reconstruction accuracy and visual quality..

Index Terms— Multi-view, Face Hallucination, Sparse Representation

1. INTRODUCTION

Multi-view hallucination (MVH) aims to reconstruct high-resolution (HR) facial images of a certain pose based on low-resolution (LR) faces in different views. In real life, facial images are usually captured by cameras at different viewpoints, and exhibit large variations. Hence, this technique can be adopted as the pre-processing step to pose-invariant face recognition, which needs to generate faces of different views to improve the recognition rate. In recent years, many related algorithms have been proposed based on 3D human face models. Their underlying assumption is that a novel view can be generated by rotating a 3D model reconstructed at a certain angle, as illustrated in [1, 2, 3, 4]. As the input images used are basically arrays of numbers or pixels, and sets of object classes in the image space cannot be defined as vectors, 3D-based algorithms are therefore employed. This approach separates faces into shape and texture vectors, which encode the differences in terms of the intensities and the displacement of each point with respect to those reference images selected. Based on the shape and texture vectors generated, pixel-wise correspondence is expected to be established between faces captured at different viewpoints. This is because the components at the same position in a vector space can refer to the same types of features [5]. Hence, a 3D-based framework can be characterized as the process for deriving the correspondence between the novel features to be reconstructed and features extracted in the training set. However, the reconstruction performance of the 3-D based algorithms is highly dependent on the estimation of the pixel-to-pixel correspondence, which is still an open issue. Moreover, the construction of a 3D human face model and the generation of texture and shape vectors greatly increase the computational complexity.

2. RELATED WORK

Unlike the 3D-based framework, another approach reconstructs a virtual view in the 2D domain via learning methods without the use of any 3D human face model. In [6], Jia and Gong proposed a two-stage patch-based learning algorithm, which employs a hierarchical tensor space to represent facial images across multiple modalities. [7] applied locally linear regression (LLR) to overlapped local patches between faces with different views. [8] employed a similar method, which directly applied the LLR method to the patches at the same positions in faces with different views. Compared with the 3D-based methods, the 2D domain approaches can greatly reduce computational complexity and render a relatively good performance. However, when the pose needs to be reconstructed across a large change in viewpoint with respect to the input pose, the correspondence relation established based on patches becomes weak, and thus the reconstruction performance will degrade greatly.

In order to overcome the inherent weakness of patch-based method, we propose a 2D domain-based learning method which can deal with MVH when the change of viewpoint is large. The contributions of this paper are twofold. Firstly, we employ sparse representation instead of patches to encode the information on facial images. The aim of sparse representation is often to reveal certain structures of a signal and to represent these structures in a compact and sparse representation [9, 10], which is suitable to code the intrinsic features. Moreover, compared with those previous works which apply subspace analysis methods to derive the weights, sparse representation has shown its robustness in information extraction when the correspondence relationship is relatively weak [11, 12]. Secondly, we introduce a novel

framework based on the sift optical-flow method to reduce the differences between the sparse-coding coefficients (SPC) used to encode the input facial image and the unknown target HR face. Different from previous methods which consider the SPC for both the input face and the target face to be similar, our method takes their difference into consideration and embed a new term in the objective function to iteratively refine the initial estimated results. Moreover, we introduce the warped images which are generated via sift optical flow, and we further employ the warped samples to refine the SPC.

3. PROPOSED METHOD

3.1. Linear mapping relationship

Based on the assumption that the 3D face surface is Lambertian [13, 14], we denote the intensity function at the point (x, y, z) in the 3D space as $\Gamma(x, y, z)$. The 2D images with two different view indices (denoted as s and f) can be generated through the following equations:

$$I_s = P_s\Gamma, \tag{1}$$

$$I_f = P_f\Gamma, \tag{2}$$

where P represents the orthographic projection matrix which projects the 3D prototype to the 2D domain, and indices s and f represent the corresponding poses in the 2D domain, respectively.

As illustrated in [7], a linear mapping relationship can be established between I_s and I_f as follows:

$$I_s = (P_s{P_f}^T + P_s\kappa)I_f, \tag{3}$$

where κ is the matrix that characterizes the operation to estimate the missing points from its 3D prototype. Denote $C = P_s{P_f}^T + P_s\kappa$, Eq. 3 becomes

$$I_s = CI_f. \tag{4}$$

The matrix C can be regarded as the linear mapping relationship between faces in different poses. When the input is of low resolution with respect to the target image, we interpolate the input face to the same size as the target one and denote the interpolated face as I_s.

3.2. Generation of SPC

As stated previously, the linear relationship is difficult to estimate because it depends on individual geometry. Therefore, it is desirable that the HR faces can be generated without requiring the estimation of the linear mapping relationship. In our method, dictionaries are trained based on training faces at different views. With the dictionaries, the relation between the SPC of the same person at different views are learned. Denoting I_f as the target face, we have $I_f^i = R_iI_f$, which represents an image patch of size $s_1 \times s_2$ at location i. R_i is the transform matrix to extract patch I_f^i from I_f at location i. Based on the dictionary, defined as $\Phi_f \in R^{S\times M}$, where $S = s_1 \times s_2$ and M is the number of patches extracted from a facial image, the patch at location i can be sparsely represented as $I_f^i \approx \Phi_f\alpha_i$ via the sparse-coding algorithm [15]. Patches are of a small size, and are selected by one-pixel shifting each time to avoid over-complete.

We need to estimate the SPC α_f so that we can reconstruct a HR faces in frontal view. Based on the interpolated input face I_s and the linear mapping relationship in Eq. 4, we can approximately represent I_s as $I_s = \widehat{C} \cdot I_f$. The SPC of I_s can be obtained by solving the following equation:

$$\alpha_s = \arg\min_a\{\|I_s - C \cdot (\Phi_f \circ \alpha)\|_2^2 + \lambda\|\alpha\|_1\}. \tag{5}$$

The initially estimated results can be obtained by $\widehat{I}_f = \Phi_f \circ \alpha_s$. It is expected that the difference between α_s and α_f should be minimized. We define the difference between α_s and α_f as $g_\alpha = \alpha_s - \alpha_f$, and thus we want to minmize the value of g_α. First, we express α_f based on faces with the view index f,

$$\alpha_f = \arg\min_a\{\|I_f - \Phi_f \circ \alpha\|_2^2 + \lambda\|\alpha\|_1\}. \tag{6}$$

Given the learned dictionary Φ_f, the difference between the two sets of SPC will be small if the training facial images are very similar to the target view image I_f. Moreover, unlike natural scene images, facial images possess strong structural similarity, and thus similar samples can effectively infer the information for the target SPC α_f. If we can select samples which are the most similar to the input face, then we can derive the SPC of the selected samples and employ these SPC to generate a reasonable estimation of α_f. Hence, we can incorporate $l_1 norm$ g_a in Eq. 5. Denote the estimation of α_f based on the selected samples as $\widehat{\alpha}_f$, the optimization framework becomes Eq. 7, where $\circ$ denotes synthesis opertaion.

$$\alpha_s = \arg\min_a\{\|I_s - C\cdot(\Phi_f\circ\alpha)\|_2^2 + \lambda_1\|\alpha\|_1 + \lambda_2\|\alpha - \widehat{\alpha}_f\|_1\}. \tag{7}$$

3.3. Evaluation of of $\widehat{\alpha}_f$

As similar faces possess similar sparse-coding coefficients with a given dictionary, we use the input face with its view index denoted as s to search a training set for K similar samples, based on Principal Component Analysis (PCA), and denote the space containing the selected faces as $S = \{I_{s,1}, I_{s,2} \cdots I_{s,K}\}$. Then, the corresponding samples with the same view as the target face are identified, and denote the corresponding space as $F = \{I_{f,1}, I_{f,2} \cdots I_{f,K}\}$. Although using the similar images can help to make a good estimation, the sub-pixel misalignment will still affect the performance. Hence, in our method, we employ the optical-flow method to warp the similar faces selected in order to

reduce the sub-pixel misalignment. The optical-flow method used in our algorithm is based on the SIFT flow framework in [16]. Given the samples selected with respect to the input face I_s , we derive the relative displacement (u, v) for each selected sample by optimizing the following equation:

$$E_{s,j} = \arg \min_{u,v} \|I_{s,i}(x + u_i, y + v_i) - I_s(x, y)\|^2 \quad i = 1, 2 \cdots K \tag{8}$$

In our method, we predict $\widehat{\alpha}_f$ using a two-stage coarse-to-fine approach. Initially, given the selected samples in view s, we derive the spatial displacement with respect to the input face. Then, we employ the relative spatial displacement derived based on view index s to initially estiamte the corresponding spatial displacement for the selected counterparts with view index f. Hence, we warp the samples in space F based on the displacement vector (u, v) derived in Eq. 8, and the warped space is denoted as $W = \{W(I_{f,1}u_1, v_1), W(I_{f,2}, u_2, v_2) \cdots W(I_{f,K}, u_K, v_K)\}$. Hence, we minmize the warping error that can be expressed as follows:

$$E_{f,i} = \|W(I_{f,i}, u_i, v_i) - \widehat{I}_f\|^2. \tag{9}$$

where $\widehat{I}_f$ is initially set as $= \Phi_f \alpha_s$, and α_s is the SPC of the input face.

However, minimizing Eq. 9 only will easily cause overfit as no regularization terms are introduced. On the other hand, we expect that the estimation should be more dependent on the sample which has small warping errors. Hence, in the second stage, we define a penalty factor used to balance the weights that each warped sample contributes to the target face as follows:

$$\chi_i = \frac{E_{f,i}^{-\beta}}{\sum_i (E_{i,f} + \varepsilon)^{-\beta}}, \tag{10}$$

where χ_i represents the penalty factor assigned to the i^{th} selected frontal sample, β controls the penalty power, and ε is a small value that makes the denominator not be zero. We define the penalty-factor matrix in the form of $\Upsilon = \text{diag}(\chi_1, \chi_2 \cdots \chi_K)$ and denote the concatenation of the SPC $\alpha_{f,i}$ as $\alpha_f^r = \{\alpha_{f,1}, \alpha_{f,2} \cdots \alpha_{f,K}\}^T$, where $W(I_{f,i}, u_i, v_i) = \Phi_f \circ \alpha_{f,i}$ and the concatenation of K numbers of α_s is denoted as $\alpha_s^r = \{\alpha_{s,1}, \alpha_{s,2} \cdots \alpha_{s,K}\}^T$. The weights for the SPC of each warped sample contributed to the target face are treated according to the penalty factor derived in Eq. 10. Inspired by our previous methods [17, 18], we characterize the distribution of the weights with Gaussian Mixture Model (GMM) as follows:

$$w_f \propto \frac{1}{Z} \exp\{-\frac{1}{2\sigma^2}(\alpha_f^r - \alpha_s^r)^T \cdot \Upsilon \cdot (\alpha_f^r - \alpha_s^r)\}, \tag{11}$$

where w_f denotes the weights that SPC of warped selected samples contributed to the target face, Z is a normalization constant, and σ^2 evaluates the variance of the assumed estimation error. The weights $w_{f,i}$ for the SPC of the i^{th} warped sample can be calculated as:

$$w_{f,i} = \frac{1}{Z}\chi_i \exp\{-\frac{1}{2\sigma^2}(\alpha_f^r - \alpha_s^r)^T(\alpha_f^r - \alpha_s^r)\}. \tag{12}$$

Given the weights of SPC of each sample, the estimated SPC $\widehat{a}_f$ can be estimated.

$$\widehat{a}_f = \sum_{i=1}^{K} w_{f,i}\alpha_{f,i}. \tag{13}$$

In our method, we generate the final results based on an iterative reconstruction. Denoting i as the iterative index, Eq. 7 can be rewritten as:

$$\alpha_s^{(i)} = \arg \min_a \{\|I_s - C \cdot (\Phi_f \circ \alpha)\|_2^2 + \lambda_1\|\alpha\|_1 + \lambda_2\|\alpha - \widehat{\alpha}_f^{(i-1)}\|_1\}, \tag{14}$$

where initially $\widehat{\alpha}_f^0 = 0$. Note that α_s, used in Eq. 9 and 12, needs to be updated according to Eq. 14. The main procedures of this iterative reconstruction algorithm are summarized in Fig. 1.

Algorithm 1. Iterative reconstruction

Initialization: $\widehat{\alpha}_f = 0$ and $\widehat{I}_f = \Phi_f \circ \alpha_s$
for $i = 1, 2 \cdots iter$ **do**
1. Use Eq. 9 to update the warping errors for each warped sample.
2. Use Eq. 10 and 12 to update the weights of SPC.
3. Use Eq. 13 to calculate $\widehat{\alpha}_f$.
4. Solve the optimization equation Eq. 14 to update the value of $\alpha_s^{(i)}$
5. Update the value of $\widehat{I}_f = \Phi_f \circ \alpha_s^{(i)}$

end

Fig. 1. The iterative-reconstruction algorithm.

In our method, we adopt the local dictionary learning method for the patches at the same position, and we apply PCA to each cluster which contains patches at the same positions. As faces possess strong structural similarity, and thus patches at the same position can refer to the same type of facial features. Moreover, as indicated in [19, 20, 21], local cluster-based dictionary is robust in preserving local structure information.

4. EXPERIMENTS

In the experiments, five datasets of face images with different poses are used, and each of the datasets contains 68 samples selected from the CMU PIE [22] database. The poses of the five datasets are $\pm 45°$, $\pm 22.5°$, and frontal view $0°$, where $+$ denotes the right-side view and $-$ denotes the left-side view. The facial images are cropped based on the coordinates of the two eyes, normalized to a size of 100×100 pixels using the method in [23]. The down-sampling factor is 4 for the input facial images, i.e. the input face is of 25×25.

First, in order to verify the robustness of sparse representation, we compare our algorithm with two patch-based methods, namely the tenor patch-based method [6] and the LLR based method [7]. The reconstructed HR frontal faces are shown in Fig. 2. based on distinct view points (i.e. $-45°$, $-22.5°$, $22.5°$, $45°$). To evaluate the respective methods quantitatively, the mean squared error (MSE) and the structural similarity index (SSIM) [24] are measured and tabulated in Tables 1 and 2 .

Second, to evaluate the effectiveness of the regularization term $\|\alpha - \widehat{\alpha}_f\|_1$ used and the sift optical-flow method applied in our algorithm, we compare the results obtained by refined SPC with the results by regarding $\alpha_s = \alpha_f$. We select three representative samples (i.e. a male, a female and a man with beard) in the training set to demonstrate the performance of sparse-coding refinement, and the results are shown in Fig. 3.

Methods	L45	L22.5	R22.5	R45
Jia[6]	454.01	412.85	411.98	471.00
Chai[7]	420.84	322.31	321.74	416.44
Our	347.91	303.16	299.29	345.33

Table 1. The MSE of the respective methods when the input face images are of different poses.

Methods	L45	L22.5	R22.5	R45
Jia[6]	0.55	0.56	0.56	0.54
Chai[7]	0.59	0.62	0.62	0.59
Our	0.65	0.67	0.67	0.65

Table 2. The SSIM of the respective methods when the input face images are of different poses.

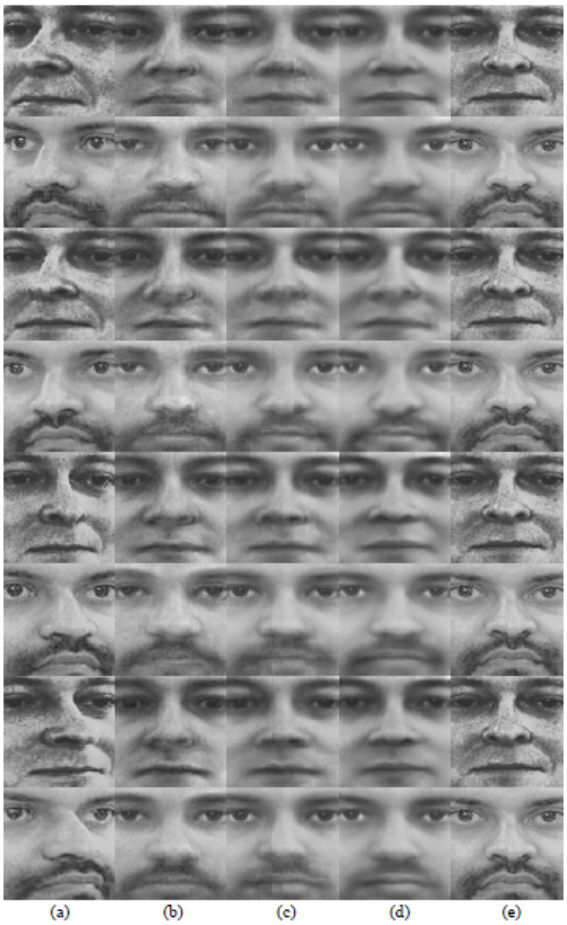

Fig. 2. Virtual facial images generated using different algorithms with the poses at $-45°$, $-22.5°$, $22.5°$,and $45°$: (a) the input LR faces, (b) Jia's method [6], (c) Chai's method [7], (d) our proposed method, and (e) the target HR images.

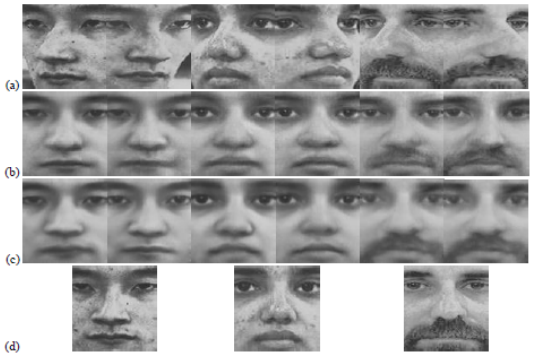

Fig. 3. Results based on different SPC generation schemes with the given viewpoint at $45°$: (a) the input faces, (b) the results generated by taking $\alpha_s = \alpha_f$, (c) the results generated by taking the SPC refinement, and (d) the ground-true frontal-view face images.

5. ACKNOWLEDGEMENT

The project was supported by a Grant from RGC of the HKSAR, China under Project no.PolyU 5187/11E.

6. CONCLUSIONS

In this paper, we have proposed a novel HR virtual-view generation method for face images. Our method uses sparse representation to code the information in a training set. The optical-flow method with a penalty function has been employed to refine the sparse-coding coefficients used to reconstruct the target images. Our method works well even when the change of viewpoint is large, and it can effectively overcome the difficulty in establishing the correspondence across large viewpoint changes, which has been a problem in most of the previous work.

7. REFERENCES

[1] T. Vetter, "Synthesis of novel views from a single face image," *International Journal of Computer Vision*, vol. 28, no. 2, pp. 103–116, 1998. 1

[2] T. Vetter and T. Poggio, "Linear object classes and image synthesis from a single example image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 733–742, 1997. 1

[3] V. Blanz, P. Grother, P.J. Phillips, and T. Vetter, "Face recognition based on frontal views generated from non-frontal images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 2, pp. 454–461. 1

[4] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194. 1

[5] D. Beymer and T. Poggio, "Image representations for visual learning," *SCIENCE-NEW YORK THEN WASHINGTON-*, pp. 1905–1909, 1996. 1

[6] K. Jia and S. Gong, "Generalized face super-resolution," *IEEE Transactions on Image Processing*, vol. 17, no. 6, pp. 873–886, 2008. 1, 4

[7] X. Chai, S. Shan, X. Chen, and W. Gao, "Locally linear regression for pose-invariant face recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 7, pp. 1716–1725, 2007. 1, 2, 4

[8] X. Ma, H. Huang, S. Wang, and C. Qi, "A simple approach to multiview face hallucination," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 579–582, 2010. 1

[9] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006. 1

[10] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: Design of dictionaries for sparse representation," *Proceedings of SPARS*, vol. 5, pp. 9–12, 2005. 1

[11] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 372–386, 2012. 1

[12] K. Jia, T.H. Chan, and Y. Ma, "Robust and practical face recognition via structured sparsity," in *The European Conference on Computer Vision (ECCV)*, 2012. 1

[13] G. Vogiatzis and C. Hernández, "Self-calibrated, multi-spectral photometric stereo for 3d face capture," *International Journal of Computer Vision*, vol. 97, no. 1, pp. 91–103, 2012. 2

[14] C.J. Lin, S.Y. Lin, C.C. Peng, and C.Y. Lee, "A constrained independent component analysis based photometric stereo for 3d human face reconstruction," in *Computer, Consumer and Control (IS3C), 2012 International Symposium on*, 2012, pp. 710–712. 2

[15] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on pure and applied mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004. 2

[16] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 978–994, 2011. 3

[17] Z. Hui and K.M. Lam, "Eigentransformation-based face super-resolution in the wavelet domain," *Pattern Recognition Letters*, 2011. 3

[18] Z. Hui and K.M. Lam, "An efficient local-structure-based face-hallucination method," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 1265–1268. 3

[19] P. Chatterjee and P. Milanfar, "Clustering-based denoising with locally learned dictionaries," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1438–1451, 2009. 3

[20] P. Chatterjee and P. Milanfar, "Patch-based near-optimal image denoising," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1635–1649, 2012. 3

[21] S. Yang, L. Zhao, M. Wang, Y. Zhao, and L. Jiao, "Dictionary learning and similarity regularization based image noise reduction," *Journal of Visual Communication and Image Representation*, 2012. 3

[22] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2002, pp. 46–51. 3

[23] X. Xie and K.M. Lam, "An efficient illumination normalization method for face recognition," *Pattern Recognition Letters*, vol. 27, no. 6, pp. 609–617, 2006. 3

[24] Z. Wang and A.C. Bovik, "Mean squared error: love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009. 4