COUPLED LEARNING BASED ON SINGULAR-VALUES-UNIQUE AND HOG FOR FACE HALLUCINATION

Songze Tang,Liang Xiao,Pengfei Liu,Huicong Wu

Nanjing University of Science and Technology

ABSTRACT

This paper proposed a novel method for face hallucination based on a neighbor embedding technique. Traditional neighbor embedding approaches often offer counterintuitive results because consistency between high resolution images and low resolution images cannot be preserved without taking the intrinsic features of the image patches into account. In order to reinforce the consistency, on the one hand, we exploit the singular-values-unique (SVU) features inspired by singular values decomposition (SVD) successfully applied in image processing. On the other hand, we introduced the Histograms of Oriented Gradients (HOG) features to characterize the local geometric structure of the image patches to alleviate the effects of noise. At last, the learning space is extended to a coupled feature space that combines the SVU and HOG features. Simulation experiments show that this proposed approach could provide competitive results in simulation experiments in subjective and objective quality.

Index Terms— Face hallucination, neighbor embedding, SVU, HOG, coupled feature space

1. INTRODUCTION

Face hallucination, a hot research topic in image processing and computer vision, is aimed at estimating high-resolution (HR) face images from one or many low-resolution (LR) ones. It has broad applications in tasks such as face recognition, long distance video surveillance. Although a lot of related works have been proposed, such as interpolation methods, reconstruction-based methods [1, 2] and neighbor embedding (NE)-based methods [3]. The NE-based methods show the most promising solutions for its impressive performance.

In 2004, inspired by locally linear embedding, Chang et al. [3] firstly introduced an image super-resolution reconstruction method with a manifold assumption. The manifold assumption states that patches in the low- and high-resolution images form manifolds with similar local neighborhood structures. In the wake of Chang's pioneering work, many variant improvements to NE for super-resolution(SR) have been proposed. Chan et al. [4] employed a novel combination of features for better neighborhood preservation, subdivide the training samples with two flexible schedules for guiding the neighbor search and adaptive tuning the neighborhood size, and bootstrap the edge samples for exploiting the capacity of a small training set with less redundancy. Compared with [3], this approach can reconstruct sharper edge details. However, their algorithm produces artifacts in the case of incorrect edge detection. In other words, their approach has not yet settled the problem of ambiguities between LR and HR manifolds. Recently, sparse coding method has achieved a notable success in computer vision. Yang et al. [5] proposed a novel method for adaptively choosing the relevant reconstruction neighbors to represent the relationship between the training data and the input patches based on sparse coding. Dong et al. [6] proposed an adaptive sparse domain selection (ASDS) model for image SR recovery, where self-similarities within the same scale is formulated as a regularization term for more robust reconstruction. Considering both the local sparsity and the nonlocal sparsity constraints, they further proposed a centralized sparse representation model for image restoration, achieving promising performance on image SR reconstruction. Very recently, Li et al. [7] have proposed a method for face hallucination based on learning the sparse local-pixel structures of the target HR facial images. The sparse representation is used to capture the local structures from the HR example faces, and optical flow is applied to make the learning process more accurate.

Obviously, feature representation is key to the neighbor searching and recovery of the high-resolution images in the NE technique. Meanwhile, we note that the previously mentioned methods used intensity information and simple feature, i.e., the value of a pixel and the gradients of each neighborhood, as a way to measure the similarity of patches. However, these features lead to not holding always the abovementioned manifold assumption. Because pixel intensities only exhibit their variance to intensity difference between image patches, whereas gradient features are sensitive to noise. Therefore, the motivation and starting point of this letter is to exploit the intrinsic features of image patches and alleviate the effects caused by noise. In this way, we can get more accurate neighbor searching and better performance. Inspired by singular values decomposition (SVD) successfully applied in image processing [8,9], we proposed a singular-values-unique(SVU)

Thanks to XYZ agency for funding.

feature to characterize the intrinsic structure based on SVD of each image patch. Then, to further reinforce the consistency between the low-resolution manifold (LRM) and the corresponding high-resolution manifold (HRM), we exploited the Histograms of Gradient (HOG) features of image patches to capture the local geometric structure and alleviate the effects from noise [10]. Finally, a coupled feature space is constructed, which combined the exploited singular-values-unique feature together with the HOG features. The similar patches are searched in the coupled space, where the consistency of LRM and the corresponding HRM can preserve effectively.

The rest of this paper is organized as follows: In Section 2, we introduce neighbor embedding method briefly and present the proposed method. Results are provided in Section 3, both simulated and practical results are included. The flowchart of our proposed algorithm is summarized in Fig. 1.

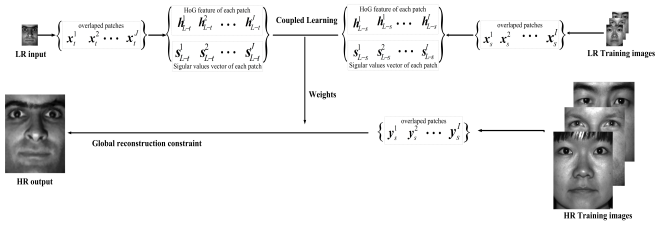

Fig. 1. The flowchart of the proposed algorithm.

2. PROPOSED

2.1. Brief Introduction of NE Method

NE for SR reconstruction assumes that the two manifolds constructed by the LR and HR patches respectively have similar local structures and an HR patch can be reconstructed by a linear combination of its neighbors [3]. Hence, each low- or high-resolution image is represented as a set of small overlapping image patches. Concretely, $X_s = \left\{\boldsymbol{x}_s^i\right\}_{i=1}^{I}$ and $Y_s = \left\{\boldsymbol{y}_s^i\right\}_{i=1}^{I}$ are the training dataset of LR image patches and the corresponding HR image patches, respectively. And the test LR image is denoted by $X_t = \left\{\boldsymbol{x}_t^j\right\}_{j=1}^{J}$, and $Y_t = \left\{\boldsymbol{y}_t^j\right\}_{j=1}^{J}$ is the estimated HR image. Obviously, I and J are the number of image patches in the training dataset and that in the test image, respectively. For each patch $\boldsymbol{x}_t^j$ in image X_t, we find K nearest neighbors in training dataset X_s , i.e.:

$$NN_j = \underset{\boldsymbol{x}_s^i \in X_s}{\arg\min}^{K} \left\|\boldsymbol{x}_t^j - \boldsymbol{x}_s^i\right\|_2^2$$

Then reconstruction weights of the neighbors are computed by minimizing the error of reconstruction $\boldsymbol{x}_t^j$:

$$\boldsymbol{w}_j = \arg\min_{\boldsymbol{w}} \left\|\boldsymbol{x}_t^j - \sum_{k \in NN_j} \boldsymbol{w}_j \boldsymbol{x}_s^k\right\|^2 \qquad s.t. \;\; \boldsymbol{1}^T \boldsymbol{w} = \boldsymbol{1},$$

Then the high-resolution patch $\boldsymbol{y}_t^j$ is computed by using the reconstruction weights: When all the HR patches vectors are computed, we merge all the estimated pixel-based patches to obtain the HR image. Since the LR patches are taken from the input image with some overlap, the final image is then obtained by simply averaging the pixel values in the overlapping regions.

2.2. Coupled Learning Based on Singular-Values-Unique and HOG Features

From the brief introduction of NE method, we note that feature selection and weights computation play an essential role. Although The LR feature vectors look like each other in low-resolution manifold, their corresponding HR feature vectors may have larger variety in appearance due to the one-to-many mappings existing between one LR image and many HR images, as shown in Fig. 2. And this serious ambiguity leads to bad weights computation. Thus, it is not easy to achieve the desired final HR output. With respect to feature selection, inspired by the Singular Values Decomposition (SVD) applied in image processing successfully [8,9], we exploit a simple, yet powerful scheme for describing the underlying features of faces, called singular-values-unique(SVU) features. SVD of an image stems from the fact that it can be factorized into two orthogonal matrices and a diagonal matrix with singular values (SV) on the diagonal. The singular values are unique for the image. Actually SV are likely to reflect the intrinsic structure of an image. Based on the description, we conduct SVD on a local patch P of $p \times p$ pixels as follows: $P = UWV^T$ where $U^TU = V^TV = I$

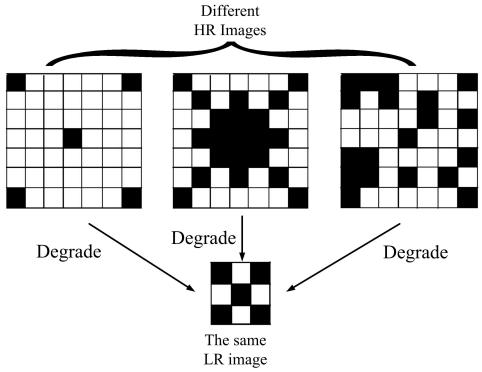

Fig. 2. The one-to-many relationship between one LR image and many HR images.

, and I is the unit matrix. W is a diagonal matrix whose elements λ_i on the diagonal are called singular values, i.e. $W = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$. The singular values vector $\boldsymbol{S}$ of patch P is defined as $\boldsymbol{S}(P) = [\lambda_1, \lambda_2, \ldots, \lambda_p]^T$.

In the following, we analyze robustness of the SVU feature in the face representation. As illustrated in Fig. 3, we obtain an index set of neighborhood patches according to the LR patch based on different features: NE and SVU. Then we present the corresponding HR patches, although they can be degraded to the same LR patch. To illustrate the effectiveness of the proposed feature representation, we calculate the sum of the root mean square error (RMSE) between the original HR patch and the HR ones in the neighborhood. From the values of RMSE, we know that as a feature representation, pixel intensity is inferior to the SVU feature. In other words, the SVU can provide features with more distinctiveness. Thus, we employ the SVU as a novel feature to characterize the image patch. In this way, the neighbor selection will be more accurate. It is conceivable that the local geometry consistency between LR and HR features is guaranteed greater than that of original LR and HR manifolds.

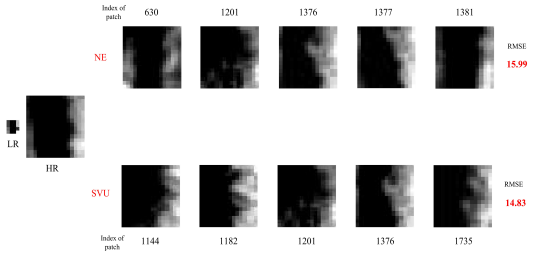

Fig. 3. The Effectiveness of SVU in neighborhood search.

Besides, in order to represent a variety of image patterns and alleviate the effects from noise, as a rather good geometric descriptor, HOG features are used to characterize the local structure of image patches, which makes full use of the distribution of local intensity gradients or edge directions.

To extract the HOG features, 1-D derivation masks $[1, 0, -1]$ and $[1, 0, -1]^T$ are first conducted on the input image I to compute the horizontal gradient G_x and vertical gradient G_y , i.e. $G_x = [1, 0, -1] * I$ and $G_y = [1, 0, -1]^T * I$. Thus the gradients direction $\theta = arctan\left(\frac{G_y}{G_x}\right)$. Next, we can discrete the direction into 9 bins evenly spaced over ("unsigned" gradient). The size of the histogram box is set 20° ($180^\circ/9 = 20^\circ$). We divide the image into some 8×8 spatial regions ("cells"), for each cell accumulating a local 1-D histogram of gradient directions over the pixels of the cell. Finally, the edge orientation histograms of each block are combined and normalized to the unit ℓ_2-norm. The block size is 2×2 cells. Concretely, for an 16×16 image patch, a 36-dimensional HOG feature is constructed.

To improve the accuracy of weights computation, we combined singular values vectors of image patches with the corresponding HOG features as a coupled feature to represent each patch. It means that we extend the learning space from single to a coupled feature space. The advantage of NE-based face hallucination will be reasonably highlighted. Thus, in the NE algorithm, $X_s = \left\{\boldsymbol{x}_s^i\right\}_{i=1}^I$ and $X_t = \left\{\boldsymbol{x}_t^j\right\}_{j=1}^J$ are substitute by $X_s = \left\{\boldsymbol{x}_{s-N}^i\right\}_{i=1}^I$ and $X_t = \left\{\boldsymbol{x}_{t-N}^i\right\}_{j=1}^J$, where $\boldsymbol{x}_{s-N}^i = \left[\boldsymbol{S}(x_s^i); \boldsymbol{h}_s^i\right]$ and $\boldsymbol{x}_{t-N}^j = \left[\boldsymbol{S}(x_t^j); \boldsymbol{h}_t^j\right]$. $\boldsymbol{S}(\cdot)$ denotes the singular values vector. $\boldsymbol{h}$ represents the HOG features. Eq. (1) and (2) can be rewritten as follows:

$$NN_j = \underset{\boldsymbol{x}_{s-N}^i \in X_s}{\arg\min}^K \left\|\boldsymbol{x}_{t-N}^j - \boldsymbol{x}_{s-N}^i\right\|_2^2$$

$$\boldsymbol{w}_j = \underset{\boldsymbol{w}}{\arg\min} \left\|\boldsymbol{x}_{t-N}^j - \sum_{k \in NN_j} \boldsymbol{w}_j \boldsymbol{x}_{s-N}^k\right\|^2 \qquad s.t. \quad \boldsymbol{1}^T \boldsymbol{w} = \boldsymbol{1}$$

2.3. Global Reconstruction Constraint

Face hallucination algorithm based on NE method is based on image patches. Since the patch size is usually relatively small, the representation ability of each patch is local. Usually, the initial estimation produced through the above process does not meet with the global reconstruction constraint perfectly. In this letter, to address this problem, the iterative back-projection (IBP) is employed to guarantee the consistency between the initial HR estimation and the final outcome [11], which should be the best viewed both globally and locally. Let Y_0 denote the initial estimation and Y represent the underlying HR image, which is assumed to get the observed LR observation X after being degraded by the degraded operators D , i.e., $X = DY$. The final reconstructed image is obtained from

$$Y^* = \underset{Y}{\arg\min} \|DY - X\|_2^2 + c\|Y - Y_0\|_2^2$$

where c is a balancing parameter. The solution to this optimization problem can be efficiently computed using gradient descent. The update equation for this iterative method is

$$Y_{n+1} = Y_n + \mu\left[D^T(X - DY_n) + c(Y_n - Y_0)\right]$$

where Y_n is the estimate of the high-resolution image after n th iteration. μ is the step size of the gradient descent. After n steps optimization process of the above, the final result Y^* satisfied the global reconstruction constraint.

3. EXPERIMENTAL RESULTS

The experiment is conducted on extended Yale face database B (B+) [9], which contains 2432 images from 38 subjects and each subject has 64 frontal images but under different illumination conditions. We randomly choose 28 subjects. Each

(a) LR (b) NE (c) ASDS (d) SLPS (e) Proposed (f) GT

Fig. 4. The face hallucination results using different methods. (a) Low-resolution image. (b) Results by NE [10] (28.8456, 0.8297). (c) Results by ASDS [23] (31.5407, 0.8898). (d) Results by SLPS (29.4352, 0.8324). (e)Results by the proposed method (33.2665, 0.9088). (f) Original high-resolution facial image.

subject selects 8 images under the similar illumination condition according to the angle the light source direction makes with the camera axis (about $12°$) as the training set. The rest 10 subjects for testing. In order to simulate the surveillance, the HR facial images are downsampled by a factor of 4 and smoothed by a 7×7 Gaussian kernel of standard deviation 1. Thus the size of LR face images are 48×42 pixels.

After performing our experiences with different patch sizes and overlapping sizes, we find that the PSNR values get larger as the overlap gets larger with the patch size fixed, and the PSNR values decrease as the patch size increases with the overlap fixed. When the patch size is small, artifacts may exist in the results. Thus, we choose to divide the original LR images into 16×16 patches with an overlap of 13 pixels between the adjacent patches. For consistency in the corresponding HR facial images, the patch size is 64×64 with an overlap of 52 pixels for HR images. Besides, the neighborhood size for NE procedure is five.

To assess objectively the performance of the proposed method, peak signal-to-noise ratio (PSNR), and structural similarity (SSIM) are calculated, as shown in Table 1 and Fig. 5, respectively. The PSNR values show that the proposed method is superior to the others. This is partly due to the coupled constraint on the LR and HR image patches reduces the ambiguity between the LR image patches and the HR patches. Moreover, SSIM is based on the HVS. The SSIM scores also suggest the effectiveness of the proposed method. In order to further testify the performance of our method, we perform the proposed method on a set of real LR facial images. These subjects are not present in the database we used for training and testing above. The super-resolved face images are shown in Fig. 6. Note that the quality of the input image is significantly worse. Still our algorithm is able to generate reasonably good results.

Table 2. The average PSNR values of different methods

method	NE	ASDS	SLPS	Proposed
PSNR(dB)	30.1430	31.9976	30.7938	32.8112

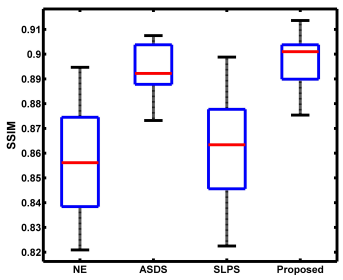

Fig. 5. SSIM comparison with other methods

Fig. 6. Results on a real world image(Top). Some original small images(Middle). The super-resolved faces (Bottom).

4. CONCLUSION

This paper has presented a novel face hallucination scheme. The distinction of the proposed approach is that the neighbor embedding is performed on a coupled space that combine singular-values-unique and HOG features rather than on the original LR space. The approach can effectively enhance the consistency between the LR and HR facial images. Experimental results demonstrate that the proposed method outperforms some state-of-the-art image hallucination algorithms.

5. ACKNOWLEDGMENTS

This work was supported in part by the Natural Science Foundation of China under Grant No. 61171165, National Scientific Equipment Developing Project of China (Grant No. 2012YQ05025004), 333 Project of Jiangsu Province and The Six Top Talents of Jiangsu Province Grant.

6. REFERENCES

[1] J. Sun, J. Sun, Z. Xu, and H. Shum, "Image superresolution using gradient profile prior," in IEEE Conf. CVPR, 2008, pp. 1-8 .

[2] Y. W. Tai, S. Liu, M. S. Brown, and S. Lin, "Super resolution using edge prior and single image detail synthesis," in IEEE Conf. CVPR, 2010, pp. 2400-2407.

[3] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in IEEE Conf. CVPR, 2004, pp. 275-282.

[4] T.M. Chan, J.P. Zhang, J. Pu, and H. Huang, "Neighbor embedding based super-resolution algorithm through edge detection," Pattern Recognition letters., vol. 30, pp. 494-502, 2009.

[5] J. C. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," IEEE Trans. Image Process., vol. 19, no. 11, pp. 2861-2873, 2010.

[6] W. S. Dong, L. Zhang, and G. M. Shi, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," IEEE Trans. Image Process., vol. 20, no. 7, pp. 1838-1857, 2011.

[7] Y. Li, C. Cai, G. Qiu, K. Lam, "Face hallucination based on sparse local-pixel structure," Pattern Recognition, vol. 47 pp. 1261-1270, 2014.

[8] W. Kim, S. Suh, W. Hwang, and J. Han, "SVD Face: Illumination-Invariant Face Representation," IEEE Signal Process. Lett., vol. 21, no. 11, pp. 1336-1340, 2014.

[9] A. Georghiades, P. Belhumeur, and D. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose," IEEE Trans. Patt. Anal. Mach. Intell., vol. 23, no. 6, pp. 643-660, 2001.

[10] Dalal. N and Triggs. B, "Histograms of Oriented Gradients for Human Detection," in IEEE Conf. CVPR, 2005, pp. 1: 886-2893.

[11] M. Irani and S. Peleg, "Motion analysis for image enhancement: Resolution, occlusion and transparency," J. Vis. Commun. Image Represent., vol. 4, no. 4, pp. 324-335, Dec. 1993.