MULTI-POSE FACE HALLUCINATION VIA NEIGHBOR EMBEDDING FOR FACIAL COMPONENTS

Yanghao Li, Jiaying Liu, Wenhan Yang, Zongming Guo*

Institute of Computer Science and Technology, Peking University, Beijing, P.R.China, 100871

ABSTRACT

In this paper, we propose a novel multi-pose face hallucination method based on Neighbor Embedding for Facial Components (NEFC) to magnify face images with various poses and expressions. To represent the structure of a face, a facial component decomposition is employed on each face image. Then, a neighbor embedding reconstruction method with locality-constraint is performed for each facial component. For the video scenario, we utilize optical flow to locate the position of each patch among the neighboring frames and make use of the Intra and Inter Nonlocal Means method to preserve consistency between neighboring frames. Experimental results evaluate the effectiveness and adaptability of our algorithm. It shows that our method achieves better performance than the state-of-the-art methods, especially on the face images with various poses and expressions.

Index Terms— Face hallucination, super resolution, neighbor embedding, nonlocal means

1. INTRODUCTION

Face hallucination is a domain-specific super resolution (SR) problem, whose goal is to estimate a high resolution (HR) face image from its low resolution (LR) counterpart. The general super resolution methods do not operate effectively on the face images due to the lack of employing structural priors of faces. Thus, many researchers devoted to face hallucination in recent years as it is so fundamental to many applications, such as enhancement and recognition of face images from surveillance videos. In order to generate high resolution images effectively and provide more facial details, numerous methods have been proposed in the last decade.

In [1], Baker and Kanada proposed a probabilistic framework to model the relationship between LR and HR image patches. However, some artifacts can also be introduced as the transformation does not consider the structure of face images in this domain-specific SR problem. To model the structure features of face images, Wang and Tang [2] proposed a global face hallucination method using Principal Component Analysis (PCA). This method reconstructs HR images by the PCA coefficients in the LR subspace. Because of the limitation of linear subspace representations, this global method requires that all the training and testing images are precisely aligned at the fixed pose and expression. Otherwise, the results usually contain ghosty effects.

To improve the result of global methods, some two-step approaches [3, 4] were proposed to combine both global and local methods. These methods first hallucinate a smooth global face image by global methods, such as PCA, and Non-negative Matrix Factorization (NMF). Then they compensate the residue by local patch-based methods, like a patch-based Markov Random Field model in [3] and a sparse representation based method in [4]. However, these methods still have the same problem as the global methods do because of the holistic constraint model.

In order to avoid the drawbacks in global or two-step methods, [5] adopted a patch-based local method by employing position prior. This method reconstructs HR face image patches by linearly combining image patches at the same position of each training image. Following this work, many position-based methods are proposed. Jung *et al.* [6] improved this method using convex optimization instead of the least square estimation. In addition, Wang *et al.* [7] employed the weighted adaptive sparse regularization. However, when the faces are not precisely aligned, the reconstruction results are often not very good as a result of the position prior.

Recently, an algorithm was introduced [8] to handle faces with various poses and expressions by employing the most similar facial components in the training set directly. This method performs well when the training faces are highly similar to the test face in terms of the pose and expression. Thus, it needs not only a large training set but also the precise landmark localization. If no similar facial component is found in the training set or the landmark localization has deviation, there are significant distortions in the estimated HR images.

The critical problem in the above face hallucination methods is that they require training and testing face images are well-aligned or highly similar. When the poses and expressions of the testing images are different from training images, their performance will decrease significantly.

In our work, we develop a multi-pose face hallucination method regardless of variations of poses and expressions.

∗Corresponding author

This work was supported by National High-tech Technology R&D Program (863 Program) of China under Grant 2014AA015205, National Natural Science Foundation of China under contract No. 61472011 and Beijing Natural Science Foundation under contract No.4142021.

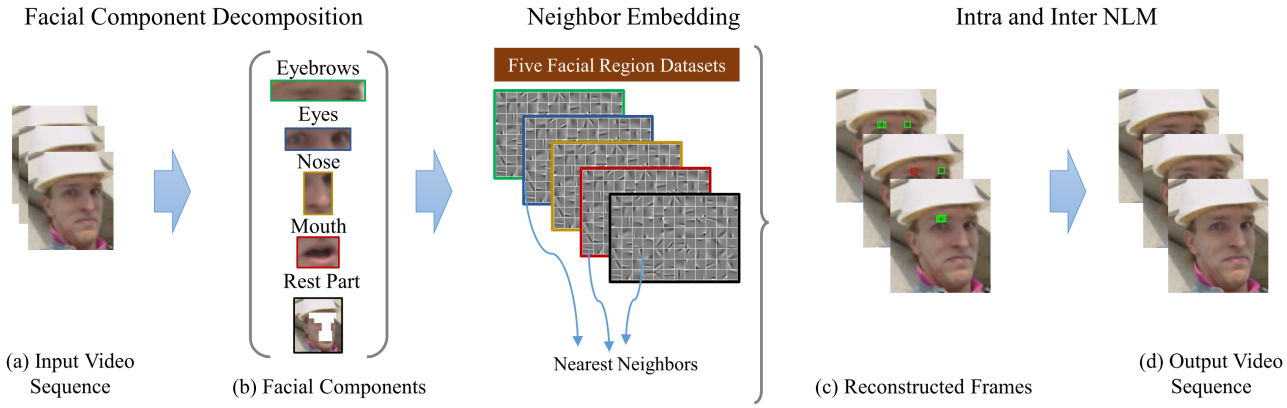


Fig. 1. An overview of the proposed method. (a) A LR input video sequence. (b) Facial components of each input frame are generated by facial component decomposition. (c) HR frames are reconstructed by neighbor embedding method with locality-constraint, then each frame is processed by Intra and Inter Nonlocal Means method. (The green blocks show the similar patches corresponding to the patch in the red block). (d) The output HR video sequence.

First, we consider the structure of a face image in terms of facial components. Then we propose a patch-based neighbor embedding method with locality-constraint to reconstruct each facial component respectively. Thus, there is no requirement for our algorithm to align the images. Finally, since our method is capable to handle face images with different poses and expressions, we also apply our method to reconstructing face images in videos. In order to preserve consistency between neighboring frames, we use optical flow to link corresponding patches in the neighboring frames, then develop an Intra and Inter Nonlocal Means method to refine each reconstructed frame. Fig. 1 illustrates the framework of our approach.

The remainder of the paper is organized as follows: we explain our proposed Neighbor Embedding for Facial Components (NEFC) method in Section 2, and show the experimental results in Section 3. The conclusion is drawn in Section 4.

2. NEIGHBOR EMBEDDING FOR FACIAL COMPONENTS

In this section, we first introduce our NEFC method for the multi-pose face hallucination, then illustrate the extended version of our algorithm that applied in the video scenario.

2.1. Face Image Structure with Facial Components

To model the structure of face images, we divide a face image into five different regions. Using a face detection and landmark localization method [9], each face is annotated by several landmark points such that all the facial components are recognized (Fig. 2). We use $c \in \{1 \sim 5\}$ to denote each facial component, including eyebrows, eyes, nose, mouth, and the rest part. For all training images and testing images, we extract these facial regions by this process at first. Then, each region of the testing image is reconstructed by the following neighbor embedding method with corresponding regions of images in the training set.

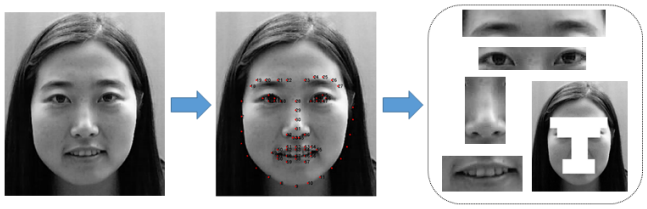


Fig. 2. Extracting facial components of face images. (a) A face image. (b) Detected landmark points (c) Facial components of the face image.

The reasons why we deal with these facial regions separately are threefold. (i) Since we use [9] to extract facial components and then utilized a patch-based reconstruction method, it does not need to align each face image any more like in [2–7]. (ii) The ambiguity between LR and HR images is a critical problem in the SR reconstruction. Different HR image patches may correspond to the same LR patch. Since the most appropriate similar patches should locate in the same region of face images, our facial component decomposition can reduce these ambiguities significantly. (iii) The four facial components contains the most important information for a face image. Via reconstructing them specifically, our method can improve the final reconstructed facial details.

2.2. Neighbor Embedding with Locality-Constraint

In this subsection, we explain how to reconstruct each facial region of a testing image by a neighbor embedding method with locality-constraint.

Let $X^c = \{x_i^c\}_{i=1}^N$ and $Y^c = \{y_i^c\}_{i=1}^N$ be the LR and HR patch dictionaries of facial component c, where N is the training set size, x_i^c and y_i^c are the corresponding LR and HR patch features. Given an input LR image X_t, we first divide them into different facial regions according to Sec 2.1, then separate them into patches.

Neighbor Embedding. Neighbor embedding approaches [10, 11] assume that small image patches in LR and HR

images form manifolds with similar local geometry. Under this assumption, target HR patch can be reconstructed as a weighted average of neighbors using the same weights in the LR feature domain. Specifically, in our algorithm, we formulate this problem as follows.

First, for each testing LR patch x_t^c , we use nearest neighbor searching (*K-NN*) to obtain the K neighbors $N_t^x = [x_{i_1}^c, x_{i_2}^c, ..., x_{i_K}^c]$ for x_t^c in the training dataset X^c. Then, the optimal reconstruction coefficients can be solved by:

$$\min_{\alpha} \|x_t^c - N_t^x \alpha\|_2^2 \qquad s.t.\, \mathbf{1}^\mathsf{T}\alpha = 1. \tag{1}$$

Locality-constraint. To improve the representation performance in Eq. (1), some regularization terms are often used to constrain the solution space, such as sparsity regularization term (l_1) and ridge regression term (l_2). In [12–14], Locality-constraint Linear Coding (LLC) is proved to have better performance than sparsity in the least square problem. In our method, we also incorporate the locality-constraint instead of the sparsity-constraint in Eq. (1):

$$\min_{\alpha} \|x_t^c - N_t^x \alpha\|_2^2 + \lambda \cdot \|d \odot \alpha\|_2^2 \quad s.t.\, \mathbf{1}^\mathsf{T}\alpha = 1, \tag{2}$$

where $\odot$ denotes the element-wise multiplication, λ is a parameter used to balance the contribution of the reconstruction error term and the regularization term, and $d = [d_1, ..., d_K]^\mathsf{T}$is a K dimensional locality adaptor that gives different freedom for each nearest patch proportional to its similarity to the input LR patch x_t^c. Specifically,

$$d_j = exp\left(\frac{\|x_t^c - x_{i_j}^c\|_2^2}{\sigma}\right), \tag{3}$$

where σ controls how diverse the nearest patches are.

Analytical solution. Following [13], the solution of Eq. (2) can be derived analytically by:

$$\hat{\alpha} = (C + \lambda \cdot diag(d))\backslash \mathbf{1}, \tag{4}$$

where $C = (N_t^x - x_i^c\mathbf{1}^\mathsf{T})(N_t^x - x_i^c\mathbf{1}^\mathsf{T})^\mathsf{T}$ denotes the data covariance matrix. The final coefficient is obtained by rescaling $\hat{\alpha}$:

$$\alpha = \hat{\alpha}/\mathbf{1}^\mathsf{T}\hat{\alpha}. \tag{5}$$

Finally, the output HR patch can be calculated by the same coefficients α:

$$y_t^c = N_t^y \alpha, \tag{6}$$

where $N_t^y = [y_{i_1}^c, y_{i_2}^c, ..., y_{i_K}^c]$ is the neighborhood in the HR space corresponding to N_t^x.

2.3. Consistency Constraint in Video Face Hallucination

When applying in the video scenario, we employ temporal redundancies to preserve consistency between neighboring frames.

First, for each patch x_t^c in the current frame t, we use Optical Flow [15] to find corresponding patches (x_{t-1}^c, x_{t+1}^c) in the previous and next frames. In our neighbor embedding method, we need to search K nearest neighbors for x_t^c in the training set . In order to preserve consistency among consecutive frames, we develop the distance function $dis(x_t^c, x^c)$ to a new intra and inter form in *K-NN* searching as follows:

$$dis'(x_t^c, x^c) = \eta \cdot dis(x_t^c, x^c) + dis(x_{t-1}^c, x^c) + dis(x_{t+1}^c, x^c), \tag{7}$$

where $dis(x_t^c, x^c)$ is the Euclidean distance between two patches, and η is used for adjusting the weight for current frame t. According to Eq. (7), K new neighbors $N_t^{x'}$ can be obtained for x_t^c.

Second, the nonlocal redundancies existing in natural images are very useful for image super resolution [11, 16]. We further introduce the Intra and Inter Nonlocal Means into our method. Besides searching for similar patches in the current image, we extend the searching range to the neighboring frames.

We formulate this problem as follows. For each local patch $y_t(i)$ in the current reconstructed frame y_t, we search for similar patches to it in the current and neighboring frames $\{y_k\}$. A patch $y_k^l(i)$ in the frame k is selected as a similar patch to $y_t(i)$ if satisfying $\|y_k^l(i) - y_t(i)\|_2^2 <= T$, where T is a preset threshold. After choosing the first L closest patches to $y_t(i)$, the nonlocal estimation can be solved by:

$$\min_{y_t} \sum_{y_t(i) \in y_t} \|y_t(i) - B(i)\beta(i)\|_2^2, \tag{8}$$

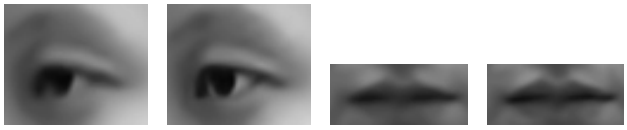
where $B(i) = [y_{k_1}^1(i), ..., y_{k_L}^L(i)]$ contains all similar patches, and $\beta(i) = [\beta^1(i), ..., \beta^L(i)]$ corresponds to the nonlocal weights:

$$\beta^l(i) = exp\left(\frac{-\|y_t(i) - y_{k_l}^l(i)\|_2^2}{h}\right)/Z(i), \tag{9}$$

where h is the controlling factor of the weight, and $Z(i) = \sum_{l=1}^{L} exp(\frac{-\|y_t(i) - y_{k_l}^l(i)\|_2^2}{h})$ is the normalization factor.

3. EXPERIMENTAL RESULTS

In this section, we evaluate our method with the criterions of the reconstruction precision and the visual quality. We use the CAS-PEAL-R1 database [17], which consists of human faces with various poses and expressions, for image face hallucination, and some standard video sequences for video face hallucination. All images taken from CAS-PEAL-R1 are under the same illumination condition. One set with 300 images at upright frontal pose and another set with 300 images at different poses, such as 30, 45, 67 degrees, are utilized as two training datasets. The test set also consists of two parts, including 50 images at frontal pose and another 50 images at profile poses. In our experiment, the scaling factor is 4, and the number of neighbors K is set to 9. The parameters λ and h are set to 0.15 and 65.



(a) Without FCD (b) With FCD (c) Without FCD (d) With FCD

Fig. 4. Effectiveness of the facial component decomposition (FCD). (a)(c) Results of our method without FCD. (b)(d) Results of our method with FCD.

Table 1. Average PSNR(dB) results by 4× on two testing sets

Methods	frontal pose	various poses
Bicubic	31.62	30.17
NRSR [11]	33.08	31.37
EFH [2]	24.37	20.92
PFH [5]	27.95	22.10
SFH [1] [8]	26.91	-
Proposed	**34.00**	**32.50**

Effectiveness of Facial Component Decomposition. As explained in the Sec 2.1, we divide each image into five different facial regions and reconstruct them respectively. Fig. 4 shows the effectiveness of this facial component decomposition process. It is clear that with the help of the facial component decomposition, our method can generate better facial details, such as the clear eyeball and lip in Fig. 4.

Objective Quality Comparison. We compare our hallucination results on face images with different methods, including Bicubic, a generic super resolution method [11] (NRSR), the global face hallucination method [2] (EFH), the position based method [5] (PFH), and the structured face hallucination method [8] (SFH). We implement algorithms in EFH [2] and PFH [5], while others are compared with the original authors' codes.

Table. 1 shows the final results on the two different testing sets, while some examples are shown in Fig. 3. One set only contains frontal images, and another contains images with various poses. Fig. 3 shows that NRSR generates more blurred result than our proposed method, since it does not employ any structure information of face images. In addition, due to the limitation of holistic constraint or position-prior, the results of EFH and PFH have some artifacts. SFH also has some unnatural and distorted effects because of the failure of finding highly similar component. As a result of our multi-pose patch-based framework which needs no alignment, our proposed method is illustrated to outperform the state-of-arts, especially on the images with different poses.

(a) Bicubic (b) Proposed

Fig. 5. Results by 4× of 23th and 50th frames in the *Foreman-CIF* sequence.

Video Face Hallucination. We also test our method on some video sequences which contain people faces. Fig. 5 is the result of two frames in the *Foreman* sequence. It shows that our method can handle various poses and expressions in the video scenario. More results can be found at [2].

4. CONCLUSION

In this paper, a novel multi-pose face hallucination approach of Neighbor Embedding for Facial Components is proposed. The structure of face images is represented by different facial components, then they are reconstructed respectively. Since our method is capable of handling different poses and expressions, we also extend our algorithm to the video scenario by employing some temporal redundancies. Experimental result shows that the proposed algorithm is more adaptive and achieves better results than state-of-the-art algorithms.

[1] The original authors' code does not support reconstructing face images with poses of more than 30 degrees.

[2] http://www.icst.pku.edu.cn/course/icb/Projects/NEFC.html

(a) Bicubic (b) NRSR [11] (c) EFH [2] (d) PFH [5] (e) SFH [8] (f) Proposed (g) Ground Truth

Fig. 3. The visual results by 4× on the two face images with frontal and 30 degree poses. Our method generates clearer and more natural facial details. The effect is better viewed in zoomed PDF.

5. REFERENCES

[1] S. Baker and T. Kanade, "Hallucinating faces," in *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 2000.

[2] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, vol. 35, no. 3, pp. 425–434, 2005.

[3] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *Int'l Journal of Computer Vision*, vol. 75, no. 1, pp. 115–134, 2007.

[4] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[5] X. Ma, J. Zhang, and C. Qi, "Hallucinating face by position-patch," *Pattern Recognition*, vol. 43, no. 6, pp. 2224–2236, 2010.

[6] C. Jung, L. Jiao, B. Liu, and M. Gong, "Position-patch based face hallucination using convex optimization," *IEEE Signal Processing Letters*, vol. 18, no. 6, pp. 367–370, 2011.

[7] Z. Wang, R. Hu, S. Wang, and J. Jiang, "Face hallucination via weighted adaptive sparse regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 802–813, 2014.

[8] C.-Y. Yang, S. Liu, and M.-H. Yang, "Structured face hallucination," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2013.

[9] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2012.

[10] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2004.

[11] Y. Li, J. Liu, W. Yang, and Z. Guo, "Neighborhood regression for edge-preserving image super-resolution," in *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, 2015.

[12] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. Advances in Neural Information Processing Systems*, 2009.

[13] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2010.

[14] J. Jiang, R. Hu, Z. Han, T. Lu, and K. Huang, "Position-patch based face hallucination via locality-constrained representation," in *Proc. IEEE Int'l Conf. Multimedia and Expo*, 2012.

[15] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," *Doctoral Thesis, Massachusetts Institute of Technology*, 2009.

[16] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 1838–1857, 2011.

[17] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The cas-peal large-scale chinese face database and baseline evaluations," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 38, no. 1, pp. 149–161, 2008.