COMPRESSED FACE HALLUCINATION

Sifei Liu *Ming-Hsuan Yang*

Electrical Engineering and Computer Science
University of California, Merced, CA 95344, USA

ABSTRACT

In this paper, we propose an algorithm to hallucinate faces in the JPEG compressed domain, which has not been well addressed in the literature. The proposed approach hallucinates compressed face images through an exemplar-based framework and solves two main problems. First, image noise introduced by JPEG compression is exacerbated through the super-resolution process. We present a novel formulation for face hallucination that uses the JPEG quantization intervals as constraints to recover the feasible intensity values from each image patch of a low-resolution input. Second, existing face hallucination methods are sensitive to noise contained in the compressed images. We regularize the compression noise caused by block discrete cosine transform coding, and reconstruct high-resolution images with the proposed gradient-guided total variation. Numerous experimental results show that the proposed algorithm generates favorable results than the combination of state-of-the-art face hallucination and denoising algorithms.

Index Terms— Face Hallucination, Compressed Domain

1. INTRODUCTION

Face hallucination is a domain-specific super-resolution problem, which aims to generate high-quality high-resolution (HR) face images from low-resolution (LR) inputs. It is a well-known ill-posed problem as pixel values in the HR space need to be recovered based on a LR image with limited intensity or color information. Numerous methods have been proposed to address the face hallucination problem [1, 2, 3, 4, 5, 6, 7, 8, 9]. In [1], exemplar HR patches are exploited to overcome the limitation of insufficient high-frequency details in images generated by super-resolution algorithms with linear constraints. However, due to ambiguities between LR and HR patches (i.e., many HR patches can be mapped to the same LR patch), the retrieved HR patches may not be effective for reconstructing HR images without artifacts. To reduce ambiguities of HR face images, linear subspace model learned from HR exemplars are proposed [4, 6], the high-frequency details are learned from exemplar patches to the reconstructed HR images via a Markov Random Field model [4] or sparse dictionaries [6]. While the main parts of faces (e.g., eyes and noses) are reconstructed well under linear subspace constraints, contour regions contain significant ghost effects since large shape variations are modelled less effectively by a global linear subspace. Although local linear subspaces are exploited [7] to reduce ghost effects, high-frequency details of reconstructed HR images are missing (as subspace methods can be considered as low-pass filters for denoising). The exemplar-based method [9] decomposes face into several parts according to facial structure, while it generates pleasant visual quality for non-compressed images, it doesn't perform well for compressed inputs.

To simplify this ill-posed problem, most algorithms assume that the input images do not contain significant amount of noise. However, in practice a large amount of images are compressed for storage or transfer. The block discrete cosine transform (BDCT) coding scheme (e.g., JPEG) is one of the most commonly adopted methods due to the high efficiency of compression ratio versus visual quality. It encodes images using non-overlapping blocks (typically with 8×8 pixels) independently with different compression qualities, which can be obtained from the header of a JPEG image. When a LR input is compressed, both blocking and ringing noisy effects are noticeable compared to a non-compressed image. In addition, details of face contours, eyes and mouths are missing, and colors of different regions are smeared. Although one possible remedy for dealing with compressed images is to first denoise LR images and then generate the corresponding HR results, we show that a straightforward combination of restoration schemes [10, 11] and face hallucination methods does not lead to high-quality HR images. Recently, a two-step approach is proposed [12] where compressed frames are first preprocessed by a deblocking algorithm to reduce artifacts and then upsampled with an edge-enhancing prior to generate HR results. Although sharp edges are restored, texture details are missing as the deblocking process eliminates high-frequency components, and the edge-based priors are not effective for reconstructing texture regions. To the best of our knowledge, no significant attempts have been made to address the face hallucination problem in the compressed domain. As facial components contain unique high-frequency details, it is of great importance to exploit specific textures rather than only edge-based prior from natural images.

In this paper, we address both denoising and super-resolution problems in one unified framework, and show that high-quality HR images can be generated from compressed

LR face images. The main contributions are summarized as follows. First, we propose a novel unified approach for practical face hallucination applications. We remove the compression noise in the hallucination process by relaxing the back-projection constraint to a quantization interval, which is defined in a compressed image. Second, we develop a gradient-guided total variation method to preserve gradients from the LR input image in the reconstructed facial textures. In addition, the color channels are restored via matching exemplars for better HR results.

2. GRADIENT-GUIDED OPTIMIZATION

We propose a gradient-guided optimization algorithm to generate high-quality hallucinated face images while minimizing compression artifacts. Given a LR test image I_l, a HR image gradient map U is generated through matching image from a LR exemplar set to its corresponding non-compressed HR set, and U is a guided gradient map. We estimate the hallucinated face image I_h by minimizing the difference between the gradient map of the HR result ∇I and the gradient map U

$$I_h = \arg\min_I \|\nabla I - U\|^2, \tag{1}$$

where ∇ is a gradient operator.

To reconstruct the hallucinated image, a regularization term is required. Suppose I_l is a non-compressed LR image and I_h is its corresponding HR image, I_l can be modeled by downsampling from a smoothed I_h:

$$(I_h \otimes G) \downarrow = I_l, \tag{2}$$

where $\otimes$ is the convolution operator, G is a Gaussian kernel with kernel width σ, and $\downarrow$ is a downsampling operator. For the compression noise, a novel regularization term is proposed to reduce the amount of undesired JPEG artifacts of the back-projected result and to resemble the LR input. Each non-overlapping 8×8 block is first converted through BDCT, and then further quantized by a matrix determined by the compression quality, thereby introducing significant noise.

To regularize compression noise, we enforce the BDCT coefficients of the downsampled I_h within the quantization limits defined by the JPEG quantization matrix T. Given the quantized integer coefficient $q_{i,j}$ of a JPEG LR image, for each BDCT block with size of 8 pixels, the BDCT coefficients $d_{i,j}$ of the downsampled I_h in (2) should be regularized in the quantization interval

$$d_{i,j} \in [T_{i,j} \cdot (q_{i,j} - 0.5), T_{i,j} \cdot (q_{i,j} + 0.5)], i, j \in [1, 8]. \tag{3}$$

In the spatial domain, the compressed JPEG image intensities of a block are equivalent to its associated JPEG image intensities of the same block. For the whole image,

$$J_Q [(I_h \otimes G) \downarrow] = I_l, \tag{4}$$

where J_Q is a compression operator with quality Q. Here Q is set to be equal to the compression quality of I_l. The spatial formulation in (4) is equivalent to the frequency formulation

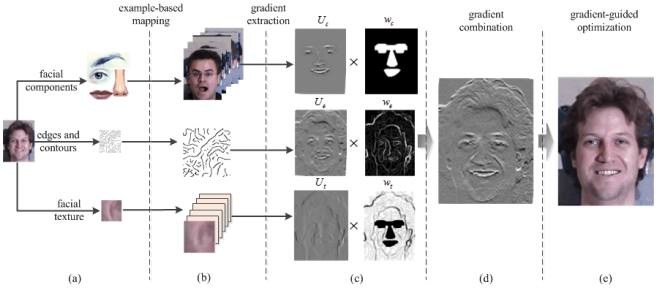

Fig. 1. (a) a LR face image is decomposed into facial components, edges and textures. (b) each part is restored through exemplar-based matching to obtain the corresponding HR counterpart. (c) gradient maps are extracted from the restored HR images and combined through weight maps to generate a gradient map. (d) resulting gradient-guided map U. (e) HR result is obtained through the proposed gradient-guided optimization. (d) and (e) are of the same image size as (c), and are magnified for illustration purpose.

of (3). The task for face hallucination in the compressed domain is formulated by

$$I_h = \arg\min_I \|\nabla I - U\|^2 \text{ s.t. } J_Q [(I \otimes G) \downarrow] = I_l. \tag{5}$$

3. GENERATING GRADIENT MAP

Three gradient maps from facial components U_c, edges U_e and textures U_t, are constructed based on the exemplar-based method. These maps are then integrated to generate the guided gradient map U. Figure 1 shows the main steps to generate the gradient map for a HR image from a LR input. The gradient maps U_c and U_e are generated in a way similar to [9]. Due to space constraints, we only describe how the maps are generated in the JPEG compressed domain in this section. More details can be found in [9].

3.1. Facial Components and Edges

The gradient map of facial components is generated by searching for the corresponding components with the maximal similarity. Several facial components, including left and right eyebrows, eyes, nose and mouth, are detected via a landmark detection algorithm [13]. These landmarks are used to align all the exemplar face images to the test image, and search for each component such that gradients are only generated in facial component regions. Once the best matched facial components are determined, a gradient map U_c is generated (See Figure 1(b)-(c)).

We model the local edge properties based on a statistical edge prior between LR and HR to produce sharp HR contours. Facial edges are first detected by the Canny edge detector on the exemplar images. For each edge pixel c with gradient magnitude m_c, we extract the gradient m_p for any of its nearby pixel p, as well as its closest distance d_p to p. The exemplar-based mapping function is built by setting a lookup-table with corresponding features of (m_c, m_p, d_p) between the LR and HR datasets. While extracting the edge features (m'_c, m'_p, d'_p) that detected by the Canny for a test image, the corresponding HR gradients information can be

restored through the lookup-table. An illustration of the gradient map U_e is shown in Figure 1(c).

3.2. Facial Textures

We use the exemplar patch match algorithm [14] to reconstruct the regions (mostly facial textures and background contents) other than facial components and edges. For a patch that densely sampled with every pixel, the most similar patch in the LR exemplar set is determined and the corresponding HR patch is retrieved. Each pixel in a LR image results in a $z \times z$ reconstructed square box, where z is the scaling factor (z=4 in our experiments). However, the patch match algorithm does not perform well in the compressed domain for two reasons. First, the reconstructed square boxes are generated independently without considering the neighbors. As a result, the reconstructed HR images are not smooth at the box boundaries. Second, the reconstructed HR image contains significant compression noise from the LR image. We propose a gradient-guided total variation method to address these issues.

3.3. Gradient-Guided Total Variation

We improve estimated HR facial textures in two aspects. First, the facial texture to be estimated, denoted by I_t, should be similar to the patch match result, denoted by I_p, to ensure the optimal HR mapping. Second, the pixel values of overlapping blocks should correlate with each other such that the image contents are continuous to reduce blocky artifacts. Instead of using a Markov random field to select the best candidate for each patch in the intensity domain [4], we enforce the pixel correlation by exploiting the image structure of the input LR image via its gradient map V_L to estimate I_t. We use three regularization terms based on total variation (TV) for image restoration:

- A L2-norm similarity term that enforces the resulting image to be similar to the HR patch match result.
- A L2-norm gradient-guided term that enforces the gradients of the estimated HR to be similar to the gradients of the LR input.
- A L1-norm TV term that eliminates the image noise introduced by the reconstructed image of a compressed input based on patch match.

The regularized face image based on texture is generated by

$$I_t^* = \arg\min_{I_t} \|I_p - I_t\|_2^2 + \beta \|V - D(I_t)\|_2^2 + \lambda \|D(I_t)\|_1 , \quad (6)$$

where D is a matrix such that $D(x)$ is the vector of first order differential of x, non-negative β controls the degree of gradient guiding, and non-negative λ controls the degree of smoothing. The vectorized gradient map V is generated by bilinear interpolation of the input image gradient map V_L to have the same size as $D(I_t)$. Both V and $D(I_t)$ are raster-scanned vectors.

For a JPEG image, the BDCT coefficients are computed based on non-overlapping blocks. Thus, gradient values change significantly along block boundaries. As the size of JPEG blocks are typically fixed as 8×8 pixels, the block positions are easy to locate. An improved gradient map V_L is computed through estimating those pixel gradient values on the block boundaries by the central difference equation,

$$V_L(i) = [V_L(i+1) + V_L(i-1)]/2, \quad i \bmod 8 = 0, \quad (7)$$

where i is a multiple of 8. To solve (6), we iteratively update I_t by using the majorization-minimization method [11]. Figure 2 shows that the reconstructed image based on the gradient-guided patch match method contains fewer artifacts, and thus better gradient map U_t from I_t^* (See Figure 1(b)-(c)).

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

Fig. 2. Reconstructed images based on facial textures. (a) HR result generated by the patch match method with blocky artifacts. (c) result generated by the proposed gradient-guided patch match method. (b)(d) zoom-in views of (a)(c).

3.4. Integrating Gradient Maps

The gradient maps of facial components U_c, facial edges U_e and facial textures U_t are integrated into U of (5). The gradient maps are combined based on specific regions using soft masks as weight maps as shown in Figure 1(c)-(d),

$$U = w_c U_c + w_e U_e + w_t U_t, \; w_t = 1 - w_c - w_e, \quad (8)$$

where w_c, w_e and w_t are the weight maps for facial components, edges and textures, respectively. With the gradient map U, a hallucinated face image can thus be computed by (5).

3.5. Restoring Color Channel

Since human eyes are more sensitive to the brightness than to the color components, the Cb and Cr color channels are downsampled (typically reduced by a factor of 2) before computing BDCT coefficients by the compression standard. As a result, the color information needs to be better restored rather than bilinear interpolation [9]. In this work, we reconstruct HR outputs in color channels by patch match [14] to reconstruct HR outputs. Thus the color details are better reconstructed without blocky artifacts.

4. EXPERIMENTAL RESULTS

We first present experimental results using the Multi-PIE [15] dataset containing face images with pose labels and landmarks. Two sets are selected for training: one with 2,184 frontal face images of 320×240 pixels, and the other 283 images of the same size with pose of 30 degrees in yaw. Each HR image in the training set is downsampled by (2) to generate its corresponding exemplar LR image of 80×60

Table 1. Quantitative evaluation of Multi-PIE frontal face images with JPEG compression Q=75. set is listed below.

Q=25	Liu [4]	Yang [6]	Ma [7]	Yang [9]	[10] + [9]	[11] + [9]	Proposed
PSNR	18.537	27.789	22.079	30.279	29.580	**30.339**	29.821
SSIM [19]	0.596	0.698	0.752	0.793	0.801	0.808	**0.817**
DIIVINE idx. [14]	33.214	**27.498**	51.622	38.439	38.835	38.643	33.434

Fig. 3. Hallucinated images using the Multi-PIE (top 2 rows) and PubFig datasets. LR input with JPEG compression Q=75. (Results are best viewed on a high-resolution display.)

pixels. The remaining 342 frontal facial images and 9 images with pose variations are used to form the test set with no overlap of identity to the training set. The LR test images are generated by (4) (smoothing, downsampling and compression) with Gaussian kernel width of 1.6 and downsampling factor of 4. Facial landmarks are identified by [13] to align the facial components from exemplars to the test images. All the face hallucination results are evaluated quantitatively using PSNR, structural similarity (SSIM) [16] and DIIVINE index [17]. Implemented in MATLAB, it takes 1 minute to upsample a LR test image with a scaling factor of 4 on a 3.4G Hz Quad Core CPU.

We evaluate the proposed algorithm with state-of-the-art algorithms by using the released code of [9] and implementing the methods of [4, 7, 6], in which the same settings for training and test datasets are used. Figure 3 shows the hallucinated face images (with Q=75) with a frontal face image with a frontal smiling face image (first row), and a face at 30 degrees yaw pose (second row). We also use face images from the PubFig dataset [18] for evaluation where photos are taken in unconstrained environments and further compressed with Q=75 (third row). We note that some hallucination methods of [4, 6, 7] do not regularize the hallucinated results to be close to input images via the back-projection [19], the HR images do not contain magnified JPEG compression noise, as [9] does. Although the methods based on high-frequency texture reconstruction [4] and sparse coding [6] generate high-frequency details, their results do not contain clear facial components details or textures. The HR images generated by [7] are over-smoothed. In contrast, the proposed algorithm performs well in the regions of facial components, and texture regions without much JPEG compression noise.

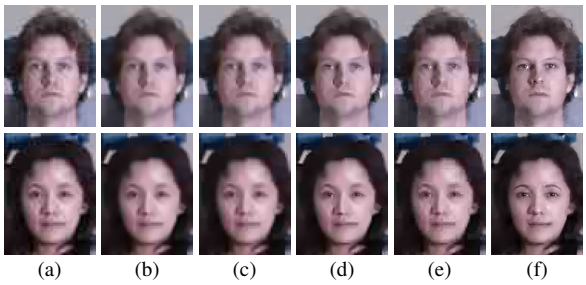

Fig. 4. From left: LR input with Q=25, LR with NLM denoising [10], TV denoising [11], Yang [9] preprocessed with [10], Yang [9] preprocessed with [11], and the proposed method. (a)-(c) upsampled by nearest neighbor interpolation for better observation. (Results are best viewed on a high-resolution display.)

We also compare the proposed algorithm with the two-step approach that carries out denoising and face hallucination sequentially. The method of [9] which performs perfect on Multi-PIE with non-compressed image are used here for comparison. In Figure 4(a), we show a face image that highly compressed with Q=25 and use respectively the non-local means (NLM) [10] and TV [11] algorithms to denoise the compressed LR input image. The denoised images, as shown in Figure 4(b)(c), show that both methods can remove JPEG blocking and ringing noise significantly. However, these denoising algorithms smooth out noise as well as facial textures of LR images, and thus the generated HR results do not contain details. Although the blocky and ringing effects are reduced, the hallucinated images are over-smoothed. All quantitative evaluations on the frontal face test set are listed in Table 1. Overall, the proposed algorithm performs well with compressed or non-compressed images. More results and details can be found at `http://graduatestudents.ucmerced.edu/sliu32/home`.

5. CONCLUSION

In this paper, a novel approach to generate HR images by estimating the structure from exemplars and the input prior is proposed. To estimate the HR details and remove JPEG artifacts, a structural total variation regularization method is proposed. Experimental results show that the proposed method generates high-quality images from highly compressed inputs with favourable performance than both state-of-the-arts and alternative methods based on straightforward combination of denoising and super-resolution techniques.

6. REFERENCES

[1] Simon Baker and Takeo Kanade, "Limits on super-resolution and how to break them," *IEEE PAMI*, vol. 24, no. 9, 2002.

[2] X. Wang and X. Tang, "Hallucinating face by eigen-transformation," *IEEE SMC*, vol. 35, no. 3, pp. 425 – 434, 2005.

[3] K. Jia and S. Gong, "Multi-modal tensor face for simultaneous super-resolution and recognition," in *ICCV*, 2005.

[4] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *IJCV*, vol. 75, no. 1, pp. 115–134, 2007.

[5] J.-S. Park and S.-W. Lee, "An example-based face hallucination method for single-frame, low-resolution facial images," *IEEE TIP*, vol. 17, no. 10, pp. 1806–1816, 2008.

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

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

[8] M. F. Tappen and C. Liu, "A Bayesian approach to alignment-based image hallucination," in *ECCV*, 2012.

[9] C.-Y. Yang, S. Liu, and M.-H. Yang, "Structured face hallucination," in *CVPR*, 2013.

[10] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," *ICCV*, 2009.

[11] M. A T Figueiredo, J. B. Dias, J. P. Oliveira, and R.D. Nowak, "On total variation denoising: A new majorization-minimization algorithm and an experimental comparisonwith wavalet denoising," in *ICIP*, 2006.

[12] Z. Xiong, X. Sun, and F. Wu, "Robust web image/video super-resolution," *IEEE TIP*, vol. 19, no. 8, pp. 2017–2028, 2010.

[13] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012.

[14] C. Barnes, E. Shechtman, D. Goldman, and A. Finkelstein, "The generalized patchmatch correspondence algorithm," in *ECCV*, 2010.

[15] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *FG*, 2008.

[16] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600 –612, 2004.

[17] A.K. Moorthy and A.C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE TIP*, vol. 20, no. 12, pp. 3350 –3364, 2011.

[18] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar, "Attribute and simile classifiers for face verification," in *ICCV*, 2009.

[19] M. Irani and S. Peleg, "Improving resolution by image registration," *Graphical Models and Image Processing*, vol. 53, no. 3, pp. 231–239, 1991.