FACE HALLUCINATION USING REGION-BASED DEEP CONVOLUTIONAL NETWORKS

Tao Lu[1], Hao Wang[1], Zixiang Xiong[2], Junjun Jiang[3], Yanduo Zhang[1], Huabing Zhou[1], Zhongyuan Wang[4],

[1] School of Computer Science and Engineering, Wuhan institute of Technology, Wuhan, China, 430073
[2]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843
[3]School of Computer Science, China University of Geosciences, Wuhan, China, 430074
[4]School of Computer Science, Wuhan University, Wuhan, China, 430072

ABSTRACT

Most deep learning based face hallucinations exploit random patch prior from training samples, then to learn the mapping functions between low-resolution (LR) and high-resolution (HR) images, and achieve satisfactory reconstruction performance. However, most of them do not take into account the *prior information on facial structure*, which is pivotal for face hallucination. Different from random patch prior based deep learning approaches, in this paper, we utilize *facial structural prior* and develop a simple yet powerful face hallucination, named region-based deep convolutional networks (RDCN). Firstly, we divide facial image into several regions of interest, then to train multiple parallel subnetworks of these regions for exacting better structure priors, finally HR output is reconstructed by stitching facial parts. Experiments on the FEI database demonstrate that the proposed region-based convolution networks outperform other state-of-the-art, including recently proposed deep learning based approaches, both in subjective and objective reconstruction qualities.

Index Terms— face hallucination, structural prior, deep convolutional networks, region-based deep learning

1. INTRODUCTION

In real surveillance scenario, interested of object such as face images are often far away the cameras, leading very LR and poor quality observed images. In order to enhance the quality and resolution of input facial images, learning-based face hallucination algorithms which infer the HR output images from LR inputs by the prior provided from training samples are more popular due to their superb reconstruction performance [1].

Super-resolution is an ill-posed problem, that uses different types of image prior to guide the reconstruction of an HR output. Thus, the performance of learning-based super-resolution algorithms relies on the prior information provided from training database. It is an impossible mission to learn one general dictionary (represent image prior) for all different size of images. Aviable solution is self-similarity prior at patch level is qualified to provide image content specific prior, which contains accurate, flexible and adaptive patch prior. Generally speaking, there are two popular ways to utilize a image self-similar prior from patch levels, the first one is based on position-patch (use facial structured patch), the other one is random-patch (sample random patch from image) based.

Face images are highly structured with stable spatial configuration. morphable model [2], facial component mask [3], hierarchical compositional model [4] are used to exploit facial structural regions. Considering face position content, Ma *et al.* [5] proposed a position-patch based least squares representation approach for accurate patch prior. After that, various regularization terms *e.g.,* sparse representation [6], locality-constrained representation [7], weighted sparse regularization [8], are introduced on position representation weights. Furthermore, some iterative optimization based approaches [9, 10] alternately updated dictionary and representation weights to boost performance. Although above position-patch based methods achieve superb performance, they only considerate local position prior which limits their reconstruction ability.

The other kind of approaches randomly select patches from images then to cluster predefined dictionaries for input patches. Baker *et al.* [11] first proposed multi-resolution patches matching based super-resolution algorithm to settle HR facial image reconstruction. Chang *et al.* [12] proposed a local linear embedding (LLE) approach by the k-nearest neighbor patches. Yang *et al.* [13] introduce spares representation regularization by designing raw patch dictionary in super-resolution algorithms. Recently, deep learning offered a highly effective paradigm for various vision-based applications. Dong *et al.* [14] proposed a deep convolutional network for image super-resolution algorithm which has an end-to-end mapping between LR and HR random patches without pre- or post- processing. Zhou *et al.* [15] used Bi-channel convolutional neural network to exploit patch prior.

This work is supported by the National Natural Science Foundation of China (61502354, 61501413, 61671332, 41501505), the Natural Science Foundation of Hubei Province of China (2015CFB451, 2014CFA130, 2012FFA099, 2012FFA134, 2013CF125), Hubei Chenguang Talented Youth Development Foundation, Scientific Research Foundation of Wuhan Institute of Technology.

Zhu *et al.* [16] cascaded bi-network to estimate dense corresponding filed for better texture details. Above random-patch based approaches treat every patch with same manner, however they ignore the region prior from facial structures which degrades performance of face hallucination.

Inspired by region-based convolutional network [17], we propose a novel region-based deep convolutional network for face hallucination to exploit highly structured facial prior rather than random patch and position-patch. We first exact facial regions to cluster structure priors into regions, then we build multiple sub-networks for all regions, finally, we stitch all reconstructed HR regions to render the final outputs. The main contributions are summarized as following: (1) We first introduce facial region structural prior into deep learning framework to further exploit accurate prior from training samples. (2) The multiple convolutional sub-networks can be parallel optimized, it not only obtains flexible prior but also is time efficient.

2. RELATED WORK

2.1. Face hallucination

Suppose the observed LR facial image $X \in \Re^{m \times n}$, then the purpose of face hallucination is to infer the HR latent image $Y \in \Re^{mt \times nt}$ with training samples database $\{H_i\}_{i=1}^{M}$, here, m, n represent size of image, t is the amplification scale factor, M is the number of HR training samples. As usual, we adopt image degradation model to describe the relationship between vector versions of Y and X as $\boldsymbol{y}$ and $\boldsymbol{x}$ by following expression:

$$\boldsymbol{x} = DB\boldsymbol{y} + \boldsymbol{n}, \tag{1}$$

where B is a blurring function of imaging system, D is a the decimation operator matrix, and $\boldsymbol{n}$ represents additive Gaussian white noise. In this paper, we consider a special case of this model, in which system blurring operator and the noise term are removed, then above image degradation model becomes: $\boldsymbol{x} = D\boldsymbol{y}$.

As we know, super-resolution is an inverse process of image degradation. So the aims of face hallucination is to learn a mapping function as

$$Y = F(X), \tag{2}$$

where mapping function F can be treated as HR restoration of image X. Therefore, all of face hallucinations attempt to find the best mapping function for the best performance.

2.2. Convolutional neural network

Recently, convolutional neural networks have been successfully applied in image classification, image super-resolution and other computer vision fields. Learning the end-to-end mapping function F requires the estimation of network parameters $\Theta = \{W_1, B_1, ...W_i, B_i, ...W_q, B_q\}$, W_i and B_i denote the ith layer filter and biases respectively, q is the depth

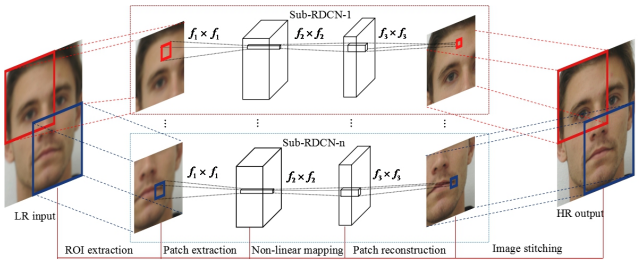

Fig. 1. Outline of our proposed region-based deep convolutional networks.

of this network, then the ith layer mapping function is expressed as an operation F_i:

$$F_i(X) = W_i * h(F_{i-1}(X)) + B_i, \tag{3}$$

where $*$ denotes the convolution operation, function $h(x)$ is the activation function on filter responses. F_{i-1} is the pre-layer feature maps, thus all the network parameter set Θ can be optimized layer-by-layer [14].

3. HALLUCINATION VIA REGION-BASED DEEP CONVOLUTIONAL NETWORKS

Outline of RDCN is shown in Fig.1, the first layer is region of interest extraction layer for clustering S regions. the second layer is convolutional layer of sub-networks to extract a set of feature maps, the third layer is nonlinearity layer to obtain the nonlinear relationship mapping, the fourth layer is reconduction layer for transferring feature maps into pixel level to render image patch, the last layer is image stitching layer combines the regional predictions to produce the final HR output.

3.1. Network structure

For training data $\{H_i\}_{i=1}^{M}$, we use formulate (1) to downsample HR images into LR version, then we upsacle them to HR size as $\{L_i\}_{i=1}^{M}$. Although L_i has same resolution with H_i, we still call them as "low-resolution" for the ease of expression.

3.1.1. ROI extraction layer

In this paper, we consider a simple special case: position region as an example. position region contains real meaning that same position area always has similar facial component. These position regions naturally contain more facial structure prior than random patches. Suppose r is the size of region, thus, we divide the $mt \times nt$ image into $S = \lceil mt/r \rceil \times \lceil nt/r \rceil$ regions, here $\lceil \cdot \rceil$ denotes round up operator, furthermore regions are allowed overlapped for smooth edge of regions. Thus, the LR and HR training sample database are transferred into region segmentations as:

$$\{L_i^s, H_i^s \mid 1 \le s \le S, 1 \le i \le M\} \in \Re^{(r \times r) \times M}, \tag{4}$$

where L_i^s and H_i^s represent the s-th region with i-th coupled LR and HR training samples with same indices.

3.1.2. Patch extracion layer

For the s-th region, we further divide training region into pathes.As shown in Fig.2, yellow square denotes region size, red square represents patch size, when learning LR and HR relationship, the marginal of HR patch is $d = f_1 + f_2 + f_3 - t$, where f_1, f_2, and f_3 are spatial filter sizes of patch extraction layer, nonlinear mapping layer and reconstruction layer, t is scale factor. Thus,the HR patch is smaller than LR patch, this means more local pixels in LR domain (more prior information) participate reconstructing of HR patches. Let u as the patch stride, then in this region, we have $[(r-u)/(p-u)]^2$ patches for training. Thus, mapping function of patch extraction layer is obtained by

$$F_1^s(L^s) = h(W_1^s * L^s + B_1^s), \tag{5}$$

where superscript s indicates index of region, and W_1^s and B_1^s reprsent the s-th sub-network filter and biases respectively. Here, W_1^s has n_1 filter of size $c \times f_1 \times f_1$, c is the number of channels of input images. B_1^s is a n_1-dimensional vector. we use the rectified linear unit (ReLU) which makes convergence much faster as activation functions:

$$h(x) = max(0, x). \tag{6}$$

3.1.3. Nonlinear-Mapping

After pre-layer extracts n_1 feature maps, in order to exploit the nonlinear mapping between LR and HR features, we project these features into another ones. We apply spatial support 1×1 for filter to finish nonlinear mapping. The operation is :

$$F_2^s(L^s) = max(0, (W_2^s * F_1^s + B_2^s), \tag{7}$$

here, W_2^s has n_2 filter of size $n_1 \times f_2 \times f_2$, B_2^s is an n_2-dimensional vector. These n_2-dimensional features are used in next step for reconstruction. It is worthy to notice that adding more nonlinear mapping layer can increase the representation ability.

3.1.4. Reconstruction layer

In order to reconstruct HR patch from features, we define a convolutional layer to render final HR patch:

$$F^s(L^s) = W_3^s * F_2^s(L^s) + B_3^s, \tag{8}$$

here, W_3^s has n_3 filter of size $n_2 \times f_2 \times f_2$, B_3^s is an c-dimensional vector, and $n_3 = 1$. Same with traditional methods, the overlapped patch are averaged to produce final region parts.

3.1.5. Image stitching

In order to render the whole image, once all S regions are finished for training, we combine all S regions together for final HR image by $Y = \sum_{s=1}^{S} F^s(\bar{X})$. The overlapped pixels between regions are averaged for smooth image content. We plug the final image Y in image degradation model to keep data fidelity:

$$Y^* = \arg\min_Y ||X - DY||_2^2. \tag{9}$$

This iterative back projection refinement generally reduce the gap between region parts and global image.

3.2. Training network

Back propagation and stochastic gradient descent algorithm is used to optimize RDCN, we define Mean Squared Error (MSE) as loss function which favors a high PSNR. It is worthy noticing that this convolution network is qualified by other kinds of loss functions, e.g., SSIM, MSSIM. Let network parameters $\Theta = \{W_i^s, B_i^s | 1 \le s \le S, 1 \le i \le M\}$, the s-th sub-network reconstructed HR image $F^s(L_i^s; \Theta)$, then MSE loss function is defined as

$$E(\Theta) = \frac{1}{M} \sum_{i=1}^{M} \left\| F^s(L_i^s; \Theta) - H_i^s \right\|^2, \tag{10}$$

where M represents the number of training samples. We use stochastic gradient descent wiht standard backpropagation [18] to minimize the above loss function. In each layer, weight matrix W_{i+1} is updated as

$$W_{i+1} = W_i + \rho * \Delta_i + \alpha \frac{\partial E}{\partial W_i}, \tag{11}$$

where ρ represent momentum as a constant, α represent the learning rate, $\frac{\partial E}{\partial W_i}$ is the derivative. We set learning rate at 10^{-4} for fast converge. Here, we can use parallel computing respectively optimize each sub-network, which is efficient in training time.

3.3. Testing

Given an input LR face image X, we interpolate it to HR size $\bar{X}$, then divide it into S regions for testing with corresponding sub-networks. The s-th region of desired HR image $Y^s = F^s(F_2^s(F_1^s(\bar{X}^s)))$. Then, we use (9) to further optimize the HR image as final output.

4. EXPERIMENTAL RESULTS

4.1. Experimental data and parameter settings

We conduct experiments on FEI face database [19], and use PSNR and SSIM as evaluation metrics. This database has 400 frontal facial images, here we random select 320 images as training set, 40 images as validation set, and the remaining 40 images are employed as testing set. The HR image is in size of 120×100 pixels, scale factor $t \in \{4, 8\}$, we use formulate (1) to generate LR images, then interpolate the LR version images into original size as LR samples. For RDCN, we set $S = 4, n_1 = 64, f_1 = 11, n_2 = 32, f_2 = 1, n_3 = 1, f_3 = 3$, where ns represent the number of feature maps, fs represent the size of convolution kernel, stride = 1, padding = 0, $\rho =$

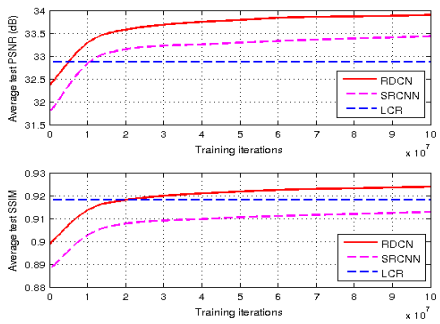

Fig. 2. PSNR and SSIM increase along with more training iterations.

0.9, $\alpha = 10^{-4}$. The filter weights of each layer are initialized by Gaussian distribution with zero mean and variance 0.001, and 0 for the bias.

4.2. Effects on number of training iterations

If $S = 4$, then $r = 64$, for $t = 4$, we define patch size $p = 32$, patch sliding stride $u = 14$, when $t = 8$, patch size $p = 40$, patch sliding stride $u = 16$. We extract 5120 patches for training, and 640 patches for validation. We use "11-1-3" network structure to test its performance. The training iteration plays important role in RDCN. However, RDCN surpasses SRCNN just few interations, and after a moderate tarning (after 2×10^7 iterations) it surpasses LCR which has the best performance in position-based face hallucination. As shown in Fig.3, the performance may further boost with more training iterations.

4.3. Network structure and parameters setting

In order to test network structure, we conduct experiments on different network structure, we just apply a basic there-layer network architecture. With 10^7 training iterations, we modify different filter sizes and fix other configurations. Table.1, gives average test PSNR in different settings, "11-1-3" structure achieves best performance in case of down-sampling by 4 and 8 times.

Table 1. PSNR with different network structures, $n_1 = 64, n_2 = 32, n_3 = 1$.

f_1	f_2	f_3	PSNR	f_1	f_2	f_3	PSNR
7	1	5	32.4933	9	1	7	32.6847
7	1	7	32.4959	11	1	3	32.8133
9	1	3	32.6573	11	1	5	32.6255

4.4. Effects on number of facial regions

We test the number of regions in this subsection for varying S at different level. As we know, regions of facial image contain local prior more than patch, especially are useful to exploit structure information for accurate prior. As shown in Table.2, when $S = 4$ the performances achieve the best level. In the same time, S should not too big and too small for better structure prior information.

4.5. Comparison to state-of-the-art

We select some state-of-the-art including sparse representation (SR) [13], locality linear embedding (LLE) [12], Least squares representation (LSR) [5], locality-constrained representation (LCR) [7], SRCNN [14] and recently Cascaded Bi-Network (CBN) [16] as benchmarks. Parameters in these papers are fine tuned for their best performance. Table.3. gives the average PSNR, SSIM and running time in case of scale factor as 4 and 8. Our method yields best performance both on PSNR and SSIM on matter with scale factors. Comparing with second best algorithm SRCNN, RCDN improves 0.46 db and 0.011 respectively in case of upscale 4 times, 1.04 db and 0.032 in case of upscale 8 times. CBN method do not train with FEI database, so the performances is lower than other methods which are trained on this database. This phenomenon further proves that specific prior is important for face hallucination. In Fig.4. and Fig.5, we list three representative images to show the subjective quality of different face hallucinations. From left to right, (a) Nearest, (b) Bicubic , (c) SR, (d) LLE, (e) LSR, (f) LCR, (g) SRCNN, (h) RDCN, (i) HR ground truth. It is easy to observe that our method yields better texture details (the red arrow marked regions).

Table 2. PSNR and SSIM with different face hallucinations

	Scale	Bicubic	SR	LLE	LSR	LCR	SRCNN	CBN	RDCN
PSNR	4	30.11	32.70	32.81	32.26	32.86	33.43	28.5445	33.89
	8	23.67	27.32	27.91	27.35	28.24	28.31	28.0809	29.35
SSIM	4	0.8750	0.9184	0.9164	0.9131	0.9182	0.9128	0.8452	0.9238
	8	0.6797	0.7828	0.8062	0.7851	0.8108	0.7836	0.8147	0.8156
time	4	0.0167	2.0900	0.9970	2.1915	2.4646	1.2683	0.1730	1.0866
	8	0.0152	0.7828	0.8062	0.7851	0.8108	1.3354	0.1389	1.1164

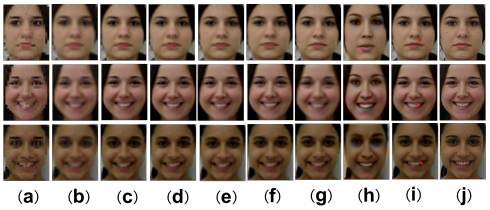

Fig. 3. Subjective results of different approaches at upscale 4 times.

Fig. 4. Subjective results of different approaches at upscale 8 times.

Therefore, comparing with above state-of-the-art, the proposed approach achieves best performance both on subjective and objective image qualities and with less time cost.

5. CONLUSION

In this paper, we present a novel face hallucination method based on region guided deep convolutional neural network. It is simple yet powerful as it outperforms state-of-the-art including basic convolutional networks, and newest cascaded Bi-Network, both on subjective and objective qualities. More complex, deeper and efficient networks will be further investigated in future.

6. REFERENCES

[1] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine Vision and Applications*, vol. 25, no. 6, pp. 1423–1468, 2014.

[2] D. Zhang, J. He, and M. Du, "Morphable model space based face super-resolution reconstruction and recognition," *Image and Vision Computing*, vol. 30, no. 2, pp. 100 – 108, 2012.

[3] C. Y. Yang, S. Liu, and M. H. Yang, "Structured face hallucination," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 1099–1106.

[4] Z. Xu, H. Chen, S. C. Zhu, and J. Luo, "A hierarchical compositional model for face representation and sketching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 955–969, June 2008.

[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, June 2011.

[7] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Noise Robust Face Hallucination via Locality-Constrained Representation," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1268–1281, 2014.

[8] 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.

[9] J. Jiang, R. Hu, Z. Wang, Z. Han, and J. Ma, "Facial image hallucination through coupled-layer neighbor embedding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 9, pp. 1674–1684, Sept 2016.

[10] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4220–4231, Oct 2014.

[11] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1167–1183, Sep 2002.

[12] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 1, no. 3, pp. I–I, 2004.

[13] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *2008 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, June 2008, pp. 1–8.

[14] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb 2016.

[15] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Learning face hallucination in the wild," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence(AAAI2015)*, Jan 2015, pp. 3871–3877.

[16] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *2016 14th European Conference Computer Vision (ECCV2016)*, October 2016, pp. 83–88.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, Jan 2016.

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[19] C. E. Thomaz and G. A. Giraldi, "A new ranking method for principal components analysis and its application to face image analysis," *Image and Vision Computing*, vol. 28, no. 6, pp. 902 – 913, 2010.