Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Learning Face Hallucination in the Wild

Erjin Zhou * and **Haoqiang Fan*** **Zhimin Cao** and **Yuning Jiang** and **Qi Yin**
Tsinghua University, Beijing, China
{zej11,fhq13}@mails.tsinghua.edu.cn
Megvii Technology, Beijing, China
{jyn,czm,yq}@megvii.com

Abstract

Face hallucination method is proposed to generate high-resolution images from low-resolution ones for better visualization. However, conventional hallucination methods are often designed for controlled settings and cannot handle varying conditions of pose, resolution degree, and blur. In this paper, we present a new method of face hallucination, which can consistently improve the resolution of face images even with large appearance variations. Our method is based on a novel network architecture called *Bi-channel Convolutional Neural Network* (Bi-channel CNN). It extracts robust face representations from raw input by using deep convolutional network, then adaptively integrates two channels of information (the raw input image and face representations) to predict the high-resolution image. Experimental results show our system outperforms the prior state-of-the-art methods.

1 Introduction

One of the most common challenges to practical face recognition systems is that most face images captured in the wild are of low resolutions. Especially in the standard-definition surveillance videos, a detectable face may be of 20×20 pixels or even smaller. Such low-resolution (LR for short) face images not only bring down the human visual experience but also adversely affect the performance of the followed face recognition and analysis. For this reason, how to infer high-resolution (HR for short) face images from LR ones, namely *face hallucination*, has attracted great research interests in the past years (Baker and Kanade 2000; Liu, Shum, and Zhang 2001; Li and Lin 2004; Wang and Tang 2005; Liu, Shum, and Freeman 2007; Tappen and Liu 2012; Yang, Liu, and Yang 2013; Dong et al. 2014; Cui et al. 2014).

Face hallucination remains an unsolved problem due to the following characteristics of face images in the wild:

1. Large variations. Realistic face images are often with large appearance variations, e.g., the changes in poses, expression and illumination;
2. Uncontrolled blur. Due to the motion or unfocused problem, the face images captured in the wild are often blurred with uncertain kernels.

Conventional face hallucination methods conduct extensive studies on the choice of low-level features, such as global eigen-faces (Wang and Tang 2005), or local texture patches (Baker and Kanade 2000; Liu, Shum, and Zhang 2001; Li and Lin 2004). Unfortunately, since the low-level features are not robust to the appearance variations (e.g., pose and expression), these methods are strictly limited to the frontal face images of constrained conditions. Recently, some algorithms are proposed to handle the variations in poses and expression (Tappen and Liu 2012; Yang, Liu, and Yang 2013). These methods rely on highly similar faces matched in training set (Tappen and Liu 2012) or face structural features extracted by an accurate facial landmark detector (Yang, Liu, and Yang 2013), as shown in Fig. 1. Moreover, these algorithms are unable to well handle variances of blur effects, since either low-level feature or structural feature is not descriptive enough for the blurred patches.

Therefore, it can be summarized that an ideal face hallucination algorithm should satisfy the following requirements: (1) it should take advantages of highly descriptive features, which is able to handle the large appearance variations and provide blur-robust face representations for face images; (2) it should not rely on other prior knowledge such as the location of facial landmarks, which may be unavailable in LR situation.

In this paper, we propose a novel *Bi-channel Convolutional Neural Network* (Bi-channel CNN). Convolutional neural network is proved to be successful in computer vision areas, such as image classification, facial landmark detection, and face recognition (Krizhevsky, Sutskever, and Hinton 2012; Sun, Wang, and Tang 2013; Yaniv et al. 2014; Sun, Wang, and Tang 2014). Benefitting from the multi-layer model, the convolutional architecture can well handle large appearance variations and provide robust representations for the following module. Our model consists of two modules: a feature extractor, and an image generator. The deep convolutional extractor learns from raw LR images and extracts descriptive face representations. The image generator takes two channels of information as inputs: the representations extracted by feature extractor and raw LR image

*This work was done when Erjin Zhou and Haoqiang Fan were visiting students at Megvii Technology.

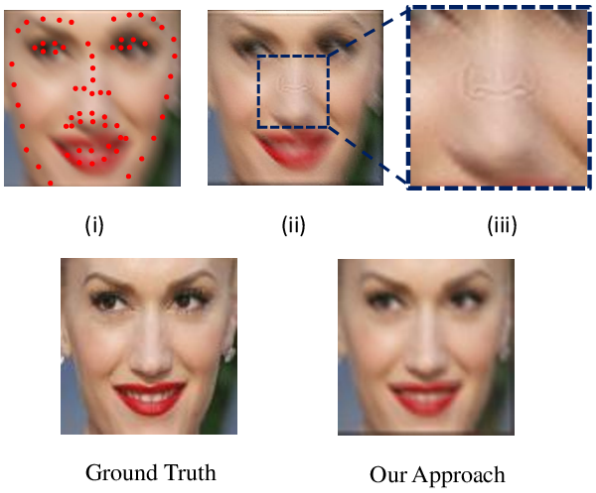


Figure 1: A failed case of structured face hallucination method (Yang, Liu, and Yang 2013). (i) is the low-resolution images with motion blur. It is presented with detected landmarks (denoted in red points); the second (ii) is the reconstructed image; the third (iii) is the close-up looking of the region in the bounding box. The method does try to add high frequency information to the image, but it generates visually implausible artefacts. The last two show our result and ground truth. From the case we can see that our method is more robust. Best viewed on a high-resolution display.

(see Fig. 2). In this paper, we exploit a simple strategy to combine two channels of information by linear combination (see details in Sec. 3). Different from the conventional strict straightforward layered deep architecture, the motivation of this two channel design is to reduce the possible information loss in the feature extractor. The comparison between Bi-channel CNN and the strict straightforward network (i.e., without the raw LR image linking to image generator directly) shows that our design improves the result greatly (see Sec. 4).

Bi-channel CNN has following properties. First, it differs from recent existing methods, in that our model does not rely on highly similar faces in database matched test faces (Tappen and Liu 2012) or an accurate facial landmark detector (Yang, Liu, and Yang 2013). Second, it is an end-to-end model, which takes the raw LR image as input and generates HR face image. We learn the feature extractor and image generator holistically. Third, deep convolutional structure helps learn a robust model to handle images captured in the wild, which contain appearance variations and uncontrolled blur.

The main contribution of this paper is that we introduce a new approach for face hallucination by deep convolutional neural network. Compared to the previous works, the proposed approach is able to handle large appearance variations and blur effects. It is an end-to-end model, which does not rely on any other prior knowledge. Furthermore, we present a novel deep architecture, Bi-channel CNN, which can adaptively integrate two channels of information (the raw input face and face representations extracted by the network). Empirical results demonstrate that this design helps the network to deal with images under different qualities, thus outperforming the state-of-the-art methods significantly on visualization quality.

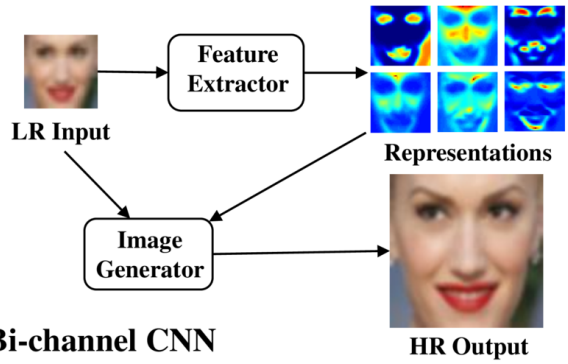


Figure 2: An overview of our system, which consists of two modules: a feature extractor, and an image generator. The deep convolutional extractor extracts descriptive face representations. The image generator takes two channels of information as inputs: the face representations and raw LR image to predict the HR face image. Different from most straightforward deep architecture, our design reduces the possible information loss in the feature extractor, thus improves the hallucination quality greatly (see Sec. 4).

2 Related Work

Face Hallucination

Super resolution techniques (Yang et al. 2008; 2010; Kim and Kwon 2010; Freeman and Liu 2011; Yang et al. 2012) are most straightforward ways to improve the resolution of image. However, all these methods are designed for general images and make weak assumptions on face images.

Baker and Kanade (2000) develop a face hallucination method using Bayesian formulation. It learns the gradient prior from a parent image pyramid for the frontal faces based on training images and incorporates it into the maximum a *posterior* (MAP) model. Liu, Shum, and Zhang (2001) present a two-step approach to hallucinate faces. The method integrates a global parametric model and a local nonparametric model. Wang and Tang (2005) propose a face hallucination method by eigen transformation. However, these methods are limited to the constrained condition or the explicit resolution-reduction function. Those limitations render previous methods impractical to apply to the images captured in the wild.

In recent years, some new face hallucination methods are proposed. Tappen and Liu (2012) introduce a method based on SIFT flow (Liu et al. 2008) and a MAP estimation framework. Their method can handle faces with widely various poses and expressions. But the method performs well only when a highly similar face can be found in training set, otherwise it will introduce artifacts on the patches

failed to match. Yang, Liu, and Yang (2013) present the method which exploits the local structure for face hallucination. It divides the face image into facial components, contours and smooth regions, and then maintains the structure by matching gradients in the reconstructed output. However, this method bases on the accurate facial landmark points that are not always available in the wild. Recently there are some deep learning based methods (Dong et al. 2014; Cui et al. 2014). However, they are patch-based and do not make full use of the faces' global features.

Convolutional Neural Networks

Convolutional neural network is proved to be successful in many computer vision areas (Bengio 2009; Schmidhuber 2014; Krizhevsky, Sutskever, and Hinton 2012; Sun, Wang, and Tang 2013; Zhou et al. 2013; Fan et al. 2014b; Szegedy et al. 2014; Ciresan et al. 2010). Benefitting from the large amount of data and recent high performance training implementation (Jia et al. 2014), the network learns to deal with complex vision problems, such as large-scale image classification (Krizhevsky, Sutskever, and Hinton 2012) and face recognition in the wild (Yaniv et al. 2014; Sun, Wang, and Tang 2014; Fan et al. 2014a). Our model benefits from the convolutional architecture to extract robust face representations from image captured in the wild and provide for face hallucination.

3 Our Approach

Problem Formulation

Let I_L and I_H denote the LR and HR face image. Following in Yang, Liu, and Yang (2013), the process of getting the LR image from HR image can be modeled as

$$I_L = \downarrow (I_H \otimes G). \tag{1}$$

Here G is the blur kernel, $\otimes$ denotes the convolution operation and $\downarrow$ means downsampling.

For a given LR image I_L, the face hallucination system f is expected to predict a hallucinated face as similar as the ground truth I_H by minimizing

$$\|f(I_L, \Phi) - I_H\|_2, \tag{2}$$

where Φ means the parameter of the system.

For a given training set composed of LR and HR image pairs $\mathcal{D} = \{(I_{L_1}, I_{H_1}), (I_{L_2}, I_{H_2}), \ldots, (I_{L_N}, I_{H_N})\}$, the parameter Φ can be determined by minimizing the *objective function* in Eq. 3

$$\arg\min_{\Phi} \quad \frac{1}{N} \sum_{i=1}^{N} \left\| f(I_{L_i}, \Phi) - I_{H_i} \right\|_2. \tag{3}$$

Blur Model

Two types of blur are considered. Gaussian blur is usually caused by out-of-focus. It is defined in Eq. 4

$$(h \otimes g)(u, v) = \frac{1}{S_g} \iint e^{\frac{-x^2}{2\sigma_x^2} + \frac{-y^2}{2\sigma_y^2}} h(u - x, v - y)\, \mathrm{d}x\, \mathrm{d}y, \tag{4}$$

σ_x, σ_y are variance parameters on the horizontal and vertical directions and S_g is a normalization constant.

Motion blur is caused by the relative movement between the image system and subjects. For simplicity, the blur kernel is modeled by two parameters θ, l, which represent the blur direction and moving distance, respectively. S_m is a normalized constant.

$$(h \otimes g)(u, v) = \frac{1}{S_m} \int_0^l h(u + t\cos\theta, v + t\sin\theta)\, \mathrm{d}t. \tag{5}$$

Basic Convolutional Neural Network

First, we present the strict straightforward convolutional model (Basic CNN) for face hallucination. It consists of two modules: a feature extractor, and an image generator. The feature extractor extracts the facial features and outputs the representations of the input image. The image generator recovers the HR image based on the representations.

We denote the network's input as I_{in}. The system can be summarized as

$$f(I_{in}, \Phi) = G(F(I_{in}, \Phi_F), \Phi_G), \tag{6}$$

where F, G represent the feature extractor and image generator and $\Phi = \{\Phi_F, \Phi_g\}$ denotes the parameters to be learned.

The feature extractor contains three convolutional layers and each convolutional layer i contains n_i feature maps denoted as $I_i^j, i = 1, 2, 3,\ j = 1 \ldots n_i$. Each feature map I_i^j is obtained by convolving the previous feature maps I_{i-1}^k with linear filters $h_i^{k,j}$, summing the results with the bias term b_i^j, applying a non-linear function $\tanh(\cdot)$, and then downsampling with a max-pooling layer. The pooling layer chooses the maximum value on every 2×2 non-overlapping sub-region. The operations are formulated in Eq. 7, where P denotes the max-pooling operator.

$$I_i^j = P \tanh\left(\sum_{k=1}^{n_{i-1}} I_{i-1}^k \otimes h_i^{k,j} + b_i^j \right), \quad j = 1 \ldots n_i. \tag{7}$$

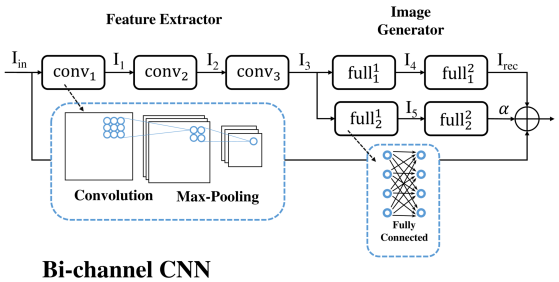
We take feature maps $I_3^j, j = 1 \ldots n_3$ as the face representations of input image I_{in} and flatten them to a vector for convenience.

The following image generator contains two fully connected layers. It takes face representations $F(I_{in}, \Phi_F)$ as input to predict the reconstructed image I_{rec}. Eq. 8 presents details of the generator, where W_j is the weighted matrix in layer j and b_j is bias term.

$$\begin{aligned} I_{rec} &= G(F(I_{in}, \Phi_F), \Phi_G) \\ &= \tanh(W_2 \tanh(W_1 F(I_{in}, \Phi_F) + b_1) + b_2) \end{aligned} \tag{8}$$

Bi-channel Convolutional Neural Network

The feature extractor focuses on the robust global representations and naturally tends to lose information from raw LR input. Thus we provide an extra data path for image generator to obtain the raw LR input directly. Specifically, we want the image generator uses both the raw LR input image



Bi-channel CNN

Figure 3: The network details of our method. The first three convolutional layers extract feature from the LR image I_{in}. Each layer outputs feature maps by convolving the previous feature maps with linear filters, applying a non-linear function $\tanh(\cdot)$, and then downsampling by using max-pooling. The following fully-connected layers are combined into two groups. One group predicts a reconstructed face image I_{rec} and another group estimates a fusion coefficient α. The HR output integrates I_{rec} and I_{in} linearly with α.

and robust representations extracted by the prior extractor to hallucinate the HR output.

$$f(I_{in}, \Phi) = G(F(I_{in}, \Phi_F), I_{in}, \Phi_G) \tag{9}$$

In this paper, we exploit a simple way to integrate both the raw input image and face representations with linear combination. It is controlled by one parameter *fusion coefficient* α as

$$G(F(I_{in}, \Phi_F), I_{in}, \Phi_G) = \alpha \uparrow I_{in} + (1 - \alpha) I_{rec}, \tag{10}$$

where $\uparrow I_{in}$ means upsampled input image by using bicubic interpolation, and α is the fusion coefficient controlled the incorporating behavior. I_{rec} is the intermediate image predicted by a two layers fully-connected network as described in Basic CNN. The fusion coefficient is predicted in the image generator implicitly based on the face representations $F(I_{in}, \Phi_F)$.

Fig. 3 shows network details of our approach. Different from the Basic CNN, the image generator in Bi-channel CNN contains four fully-connected layers. First two layers predict the intermediate image I_{rec} as described in Basic CNN and remains estimate a fusion coefficient α. Eq. 11, 12 present the output of each layer, where W_i^j is the weighted matrix and b_i^j is bias term.

$$\begin{aligned} I_4 &= \tanh(W_1^1 F(I_{in}, \Phi_F) + b_1^1) \\ I_{rec} &= \tanh(W_1^2 I_4 + b_1^2) \end{aligned} \tag{11}$$

$$\begin{aligned} I_5 &= \tanh(W_2^1 F(I_{in}, \Phi_F) + b_2^1) \\ \alpha &= \frac{1}{2}\tanh(W_2^2 I_5 + b_2^2) + \frac{1}{2}{}^{1} \end{aligned} \tag{12}$$

The pipeline of Bi-channel CNN can be summarized as:

1. *Face representations* are obtained from the raw LR input through deep convolutional layers.

[1]We put $\frac{1}{2}$ here to normalize α to $[0, 1]$.

Table 1: The details of the network (see Fig. 3). The number of feature maps in convolutional layer i is denoted as n_i. The input size of the convolutional layer i is $(n_{i-1}, s_{i-1}, s_{i-1})$ and output size is (n_i, s_i, s_i). The filter size in layer i is (n_{i-1}, n_i, w_i, w_i). The size of the fully-connected layer is denoted as $(p_i^j, 1)$ and the corresponding weighted matrix size is (p_i^j, p_i^{j-1}).

Layer	Input Size	Output Size	Filter Size
conv_1	$(3, 48, 48)$	$(32, 22, 22)$	$(3, 32, 5, 5)$
conv_2	$(32, 22, 22)$	$(64, 10, 10)$	$(32, 64, 3, 3)$
conv_3	$(64, 10, 10)$	$(128, 4, 4)$	$(64, 128, 3, 3)$

Layer	Input Size	Output Size	Matrix Size
full_1^1	$(2048, 1)$	$(2000, 1)$	$(2000, 2048)$
full_1^2	$(2000, 1)$	$(30000, 1)$	$(30000, 2000)$
full_2^1	$(2048, 1)$	$(100, 1)$	$(100, 2048)$
full_2^2	$(100, 1)$	$(1, 1)$	$(1, 100)$

2. *Intermediate image* I_{rec} is predicted from the following two fully-connected layers by using face representations as input.

3. *Fusion coefficient* α is estimated from another two fully-connected layers by using face representations as input in parallel.

4. *High-resolution image* integrates the upsampled input image and intermediate image linearly with fusion coefficient α. The upsampled input image is obtained by using bicubic interpolation.

4 Experiments

In this section, we present our experiments. First we describe the dataset used in training and testing. Then we demonstrate our implementation details. Finally we provide the results compared with other methods.

Dataset

Our system is trained from a large collection of photos crawled from the web. As any HR image can be used for our training, the scale of the dataset can be easily made large, which particularly favors the training of neural network. Our dataset contains more than 100,000 faces. The data are divided into three parts. The 60% of faces are used as training set, 20% of faces are used as validation set and the remains are left out for testing. All images are scaled into 100×100-pixel. Both training and validation set are applied Gaussian blur or motion blur randomly, downsampled by a factor from 2 to 5 (i.e., the resolution of LR image lies in the range 20×20 to 50×50 pixels). The variance of Gaussian blur σ_x, σ_y lie in the range 0 to 7. The moving distance of motion blur l lies in the range 0 to 11, and the blur direction θ is uniform selected from $-\pi$ to π. We train our model on training and validation sets and compare with other methods on test set.

Table 2: This table contains the quantitative comparison of our method Bi-channel CNN with super resolution methods SC1 (Yang et al. 2008; 2010), SC2 (Kim and Kwon 2010), recent face hallucination method SFH (Yang, Liu, and Yang 2013), and Basic CNN on images from the test set. We report the average result of PSNR and structural similarity (SSIM) (Wang et al. 2004). The results show our method outperforms others in LR images with blurs.

(a) Quantitative comparison under Gaussian blur

PSNR	Bicubic	SC1	SC2	SFH	Basic CNN	Bi-channel CNN
$\sigma = 1$	32.15	32.89	32.98	32.56	29.78	**33.11**
$\sigma = 3$	30.33	30.30	30.31	30.06	29.78	**30.35**
$\sigma = 5$	29.52	29.48	29.49	29.29	29.61	**29.71**
SSIM	Bicubic	SC1	SC2	SFH	Basic CNN	Bi-channel CNN
$\sigma = 1$	0.84	0.87	0.87	0.88	0.70	**0.89**
$\sigma = 3$	0.68	0.68	0.68	0.67	0.68	**0.71**
$\sigma = 5$	0.58	0.58	0.58	0.58	**0.65**	**0.65**

(b) Quantitative comparison under motion blur

PSNR	Bicubic	SC1	SC2	SFH	Basic CNN	Bi-channel CNN
$l = 2.0$	32.39	34.05	34.23	33.59	29.75	**34.63**
$l = 6.0$	30.11	29.68	29.69	29.53	29.35	**30.23**
$l = 9.0$	28.89	28.76	28.76	28.67	29.07	**29.33**
SSIM	Bicubic	SC1	SC2	SFH	Basic CNN	Bi-channel CNN
$l = 2.0$	0.85	0.91	0.91	0.91	0.69	**0.92**
$l = 6.0$	0.71	0.65	0.65	0.64	0.65	**0.77**
$l = 9.0$	0.52	0.48	0.48	0.48	0.61	**0.68**

Implementation Details

The size of network's input I_{in} is 48×48 pixels with RGB 3-channels. The network's output is HR image with 100×100 pixels and RGB 3-channels. The details of layer structure are summarized in Table 1.

Data Pre-processing We train our model to handle LR input with different resolutions. The resolution of LR input lies in the range 20×20 to 50×50. Since the resolution of network input I_{in} is 48×48, we upsample or downsample the LR image I_L to 48×48 by using bicubic interpolation

$$I_{in} = \begin{cases} \uparrow I_L & \text{resolution of } I_L < 48 \times 48, \\ \downarrow I_L & \text{resolution of } I_L \geq 48 \times 48. \end{cases} \tag{13}$$

All of entries of the input image I_{in} and the groundtruth image I_H are normalized to lie in the range -1 to 1. Specifically, we denote I_{in} and I_H as

$$\begin{aligned} I_{in} &= [I_{in}^R, I_{in}^G, I_{in}^B], \\ I_H &= [I_H^R, I_H^G, I_H^B]. \end{aligned} \tag{14}$$

The input image's mean and standard deviation are computed as

$$M_{in} = \left[M_{in}^R, M_{in}^G, M_{in}^B\right] = \left[\overline{I_{in}^R}, \overline{I_{in}^G}, \overline{I_{in}^B}\right], \tag{15}$$

$$\begin{aligned} S_{in} &= \begin{bmatrix} S_{in}^R & 0 & 0 \\ 0 & S_{in}^G & 0 \\ 0 & 0 & S_{in}^B \end{bmatrix} \\ &= \begin{bmatrix} \sqrt{\text{Var}(I_{in}^R)} & 0 & 0 \\ 0 & \sqrt{\text{Var}(I_{in}^G)} & 0 \\ 0 & 0 & \sqrt{\text{Var}(I_{in}^B)} \end{bmatrix}. \end{aligned} \tag{16}$$

I_{in} and I_H are normalized in Eq. 17.

$$\begin{aligned} \widetilde{I_{in}} &= \tanh\left((I_{in} - M_{in})S_{in}^{-1}\right), \\ \widetilde{I_H} &= \tanh\left((I_H - M_{in})S_{in}^{-1}\right). \end{aligned} \tag{17}$$

We use normalized images $\widetilde{I_{in}}$ and $\widetilde{I_H}$ as input signal and groundtruth for training. Therefore, the network outputs a normalized responding $\widetilde{I_{out}}$. When we test an image, we recover image from its normalized responding $\widetilde{I_{out}}$ in Eq. 18.

$$I_{out} = \text{arctanh}\left(\widetilde{I_{out}}\right) S_{in} + M_{in}. \tag{18}$$

Details of Learning We initialize all of the filters and fully-connected matrices from a zero-mean Gaussian distribution with standard deviation 0.001 and set all biases to 0. We train our model by stochastic gradient descent. All of the parameters are optimized by back-propagation. The data batch size is 200. The initial learning rate is 0.00001 for all layers.

The update rule for parameter $w \in \Phi = \{\Phi_F, \Phi_G\}$ in k-th iteration is

$$v_k = 0.9 * v_{k-1} + \epsilon * \frac{\partial L}{\partial w}\Big|_{w_{k-1}}, \tag{19}$$

$$w_k = w_{k-1} - v_k, \tag{20}$$

where v is the momentum variable, ϵ is the learning rate, and L is the objective function defined in Eq. 3. We adjust learning rate manually during training by dividing it by 10 when the validation error stops decrease. We train our model for about 5000 cycles.

Comparison with Other Methods

We compare our method with generic super resolution methods (Yang et al. 2008; 2010; Kim and Kwon 2010) and the state-of-art face hallucination method (Yang, Liu, and Yang 2013). We generate the test images through Eq. 1 by using Gaussian blur in Eq. 4 or motion blur in Eq. 5, then downsample them into 50×50 pixels to mimic the LR and poor quality images captured in the wild.

Fig. 4 shows qualitative comparisons of our proposed method and other approaches on images with different Gaussian and motion blur. The conventional face hallucination methods fail when LR images are blurred. The patch-based methods (Yang et al. 2008; 2010; Kim and Kwon 2010)

Figure 4: Qualitative comparison under different blurs. We compare the results from our system Bi-channel CNN with the results from generic super resolution method SC1 (Yang et al. 2008; 2010), SC2 (Kim and Kwon 2010), recent face hallucination method SFH (Yang, Liu, and Yang 2013), and Basic CNN. The results show our method outperforms the other methods in LR images with blurs. Best viewed in color.

fail to reconstruct a clear image, because the patches in image are polluted and these methods do not learn the high-level representations from the input. Structured face hallucination method (Yang, Liu, and Yang 2013) reconstructs images based on the accurate facial landmarks. When images are blurred, the landmarks detected are not accurate. On the other hand, both Basic CNN and Bi-channel CNN outperform the previous approaches. Moreover, Bi-channel CNN obtains extra information from raw input directly and combines with reconstructed image adaptively. So compared with Basic CNN, the artifacts in the results produced by Bi-channel CNN are greatly reduced.

We analyze the quality of reconstructed images by comparing PSNR and structural similarity (SSIM) (Wang et al. 2004). Table 2 presents the quantitative comparisons on the test set. The results show our method Bi-channel CNN outperforms other methods on LR images with different blurs. The performance of conventional approaches decreases rapidly when LR images have large blurs. On the other hand, the performance of Basic CNN is stable, which indicates that our face representations extracted from convolutional layers are robust for blurs. Moreover, our method Bi-channel CNN has a big advantage over Basic CNN when LR images have small blurs. The linear combination process improves the performance of system. We also note the performance of Bi-channel CNN approaches Basic CNN when LR images have large blurs. When the image has large blurs, it cannot provide useful information directly, so Bi-channel CNN degenerates to Basic CNN.

5 Conclusion

We propose a new approach for face hallucination using deep neural network and design a new network architecture Bi-channel CNN to integrate the raw input image and the face representations extracted by deep convolutional network. Using robust face representations, our system can handle large variations, like poses, expression, illumination and unknown blur effects, for images with different resolution. The linear integrating process of two channels of information makes our system adaptable for images with different qualities. Experiments show our method produces state-of-art result.

References

Baker, S., and Kanade, T. 2000. Hallucinating faces. In *Automatic Face and Gesture Recognition*, 83–88.

Bengio, Y. 2009. Learning deep architectures for ai. *Foundations and trends in Machine Learning* 2(1):1–127.

Ciresan, D. C.; Meier, U.; Gambardella, L. M.; and Schmidhuber, J. 2010. Deep, big, simple neural nets for handwritten digit recognition. *Neural computation* 22(12):3207–3220.

Cui, Z.; Chang, H.; Shan, S.; Zhong, B.; and Chen, X. 2014. Deep network cascade for image super-resolution. In *European Conference on Computer Vision*. Springer. 49–64.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*. Springer. 184–199.

Fan, H.; Cao, Z.; Jiang, Y.; Yin, Q.; and Doudou, C. 2014a. Learning deep face representation. *arXiv preprint arXiv:1403.2802*.

Fan, H.; Yang, M.; Cao, Z.; Jiang, Y.; and Yin, Q. 2014b. Learning compact face representation: Packing a face into an int32. In *Proceedings of the ACM International Conference on Multimedia*, 933–936. ACM.

Freeman, W., and Liu, C. 2011. Markov random fields for super-resolution and texture synthesis. *Advances in Markov Random Fields for Vision and Image Processing*.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, 675–678. ACM.

Kim, K. I., and Kwon, Y. 2010. Single-image super-resolution using sparse regression and natural image prior. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32(6):1127–1133.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Li, Y., and Lin, X. 2004. An improved two-step approach to hallucinating faces. In *Multi-Agent Security and Survivability*, 298–301.

Liu, C.; Yuen, J.; Torralba, A.; Sivic, J.; and Freeman, W. T. 2008. Sift flow: Dense correspondence across different scenes. In *European Conference on Computer Vision*. Springer. 28–42.

Liu, C.; Shum, H.-Y.; and Freeman, W. T. 2007. Face hallucination: Theory and practice. *International Journal of Computer Vision* 75(1):115–134.

Liu, C.; Shum, H.-Y.; and Zhang, C.-S. 2001. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *Computer Vision and Pattern Recognition*, volume 1, I–192.

Schmidhuber, J. 2014. Deep learning in neural networks: An overview. *arXiv preprint arXiv:1404.7828*.

Sun, Y.; Wang, X.; and Tang, X. 2013. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition*, 3476–3483.

Sun, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation from predicting 10,000 classes. In *Computer Vision and Pattern Recognition*, 1891–1898.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2014. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*.

Tappen, M. F., and Liu, C. 2012. A bayesian approach to alignment-based image hallucination. In *European Conference on Computer Vision*, 236–249.

Wang, X., and Tang, X. 2005. Hallucinating face by eigentransformation. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, , IEEE Transactions on* 35(3):425–434.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *Image Processing, , IEEE Transactions on* 13(4):600–612.

Yang, J.; Wright, J.; Huang, T.; and Ma, Y. 2008. Image super-resolution as sparse representation of raw image patches. In *Computer Vision and Pattern Recognition*, 1–8.

Yang, J.; Wright, J.; Huang, T. S.; and Ma, Y. 2010. Image super-resolution via sparse representation. *Image Processing, IEEE Transactions on* 19(11):2861–2873.

Yang, J.; Wang, Z.; Lin, Z.; Cohen, S.; and Huang, T. 2012. Coupled dictionary training for image super-resolution. *Image Processing, IEEE Transactions on* 21(8):3467–3478.

Yang, C.-Y.; Liu, S.; and Yang, M.-H. 2013. Structured face hallucination. In *Computer Vision and Pattern Recognition*, 1099–1106.

Yaniv, T.; Ming, Y.; MarcAurelio, R.; and Lior, W. 2014. Deepface: Closing the gap to human-level performance in face verication. In *Computer Vision and Pattern Recognition*, 1701–1708.

Zhou, E.; Fan, H.; Cao, Z.; Jiang, Y.; and Yin, Q. 2013. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *International Conference on Computer Vision Workshops*, 386–391. IEEE.