Contourlet Transform based Multi-scale Fusion Network for Face Hallucination

W Wei, G Q Feng and D L Cui

Northeastern University, Shenyang 110004, China

Corresponding author: weiwei_nuphd@163.com

Abstract. Recently, Convolutional neural network (CNN) has achieved great success in the field of face hallucination. However, such approaches usually generate blurry and over-smoothed Super Resolution (SR) results, and the performance suffers from degradation when super-resolve a very Low Resolution (LR) face image. To solve these problems, this paper proposes an contourlet transform based accurate CNN architecture, namely Multi-scale Fusion CNN (MSFC), which are able to reconstruct a High Resolution (HR) face image from a very low resolution input. First of all, we present multi-scale fusion CNN (MSFC) to fully detect and exploit features from LR inputs. And then, we formulate the SR problem as the prediction of contourlet transform coefficients, which is able to make MSFC further capture the texture details for super-resolve face images. Extensive qualitative and quantitative experiments show that the proposed method is capable of preserving details and achieves superior SR performance compared to the state-of-the-art methods.

1. Introduction

Face Hallucination (FH) is a domain-specific general image Super-Resolution (SR). The main goal of FH is to reconstruct a High-Resolution (HR) results from a Low-Resolution (LR) face image. Since face SR restores the high-frequency information, it is important for most face-related tasks [1, 2], e.g. face recognition.

Due to the powerful learning ability, many CNN based methods have demonstrated superiority over traditional methods [3] in the field of face image SR. However, they have obvious shortcomings that their SR results are blurry and over-smoothed. Besides, they usually perform well on small scaling factors, but the performance degrades greatly when dealing with very small image (like 16×16 pixels).

Since the first general image SR network (SRCNN) with three convolutional layers was proposed by Dong et al [4], many efforts have been made to construct effective network structures. Many previous works use convolutional layers with plain connections to construct their CNN model. In order to ease the difficulty of training a deep CNN model, residual connections and dense connections are adopted to build basic feature extraction blocks, e.g. residual block [5] and dense block [6]. However, these two kinds of block only adopt convolutional layers with a single size which leading neglect the multi-scale image features in different receptive field. To solve this problem, we propose an efficient multi-scale fusion block (MSFB) in this paper to sufficiently exploit features from multiple kinds of receptive fields.

In the other hand, most of current CNN-based SR methods use pixel-wise loss in the image spatial domain, enforcing the pixel-wise SR output become more and more similar to the ground truth HR images. However, such training process usually generates over-smooth outputs, losing texture details.
In contrast, some efforts are made to solve the problem in the transform domain, which is able to retain the contextual and textural information of an image. For example, wavelet transform has been explored for image SR in CNN models [7]. Wavelet transform effectively deals with the "point singularity" problem of one-dimension signal. However, a common limitation of it is that it cannot well represent the curves and edges of two-dimensional image because of its isotropic property [8]. In order to overcome this disadvantage of wavelet transform, this paper proposes to use the contourlet transform for face image SR. Specially, contourlet transform provide optimal approximation for a piecewise smooth function and could better represent curves of image than spatial domain and wavelet domain.

In this paper, we propose to formulate the face SR problem as the prediction of contourlet transform subbands, which are able to preserve richer texture details than spatial domain.

The contributions of our method can be summarized as follows: (1) a novel feature extraction block, namely Multi-Scale Fusion Block (MSFB), is proposed. In detail, the proposed MSFB is able to extract image features from multiple receptive fields, and use these features to provide more cues for reconstructing high quality SR face images. Based on MSFB, we construct an efficient Multi-scale fusion CNN (MSFC) in a holistic manner. (2) A novel contourlet transform and CNN based approach is proposed, which could further improve SR performance of MSFC. (3) Experimental results show that the proposed method outperforms previous state-of-the-art methods.

2. Proposed method

2.1. Network structure

The network structure of the proposed MSFC is shown in Figure 1. The MSFC consists of three modules, including Shallow Module (SM), the Deep Module (DM), and the Up Sampling Module (USM).

Figure 1. The network structure of the proposed MSFC

Here, the main goal of our network is to optimize the network parameters $\theta$:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L_{SR}^{LR}(F_{\theta}(I_{LR}^i),I_{HR}^i)$$  

(1)

where $N$ indicates the training samples, and $L_{SR}^{LR}$ is the loss function for enforcing the LR $I_{LR}^i$ become more and more similar to ground truth HR $I_{HR}^i$. The SFM is used for extracting shallow features form LR input, which consists of two convolutional layers with 3x3 kernel size. After that, the DFM contains a set of cascaded MSFBs to extract features at different levels. In order to effectively exploit these features for reconstructing HR image, we then adopt global feature fusion to further fuse image features from all preceding MSFBs.

$$F_{GF} = H_{GFF}([F_1,\ldots,F_D])$$  

(2)

where $D$ indicates the number of MSFBs in our CNN model, $[F_1,\ldots,F_D]$ indicates the channel-wise concatenation operator, which combines all of feature maps produced by preceding MSFBs. $H_{GFF}$ indicates the function of global feature fusion, which contains a 1x1 convolutional layer. After that, global residual learning is utilized as follow

$$F_{DF} = F_0 + F_{GF}$$  

(3)
where \( F_0 \) indicates the features captured by SFM. Finally, a 12×12 deconvolutional layer is adopted in USM to generate the SR result with high visual quality.

2.2. Multi-scale fusion block
In this paper, an efficient feature extraction block namely MSFB is utilized to construct the proposed MSFC, which are able to exploit features in multiple receptive fields. In Figure 2, we provide the structure of MSFB. Three-bypass networks are combined with each other in the MSFB, each of which adopts different kernel size of convolutional layers.

Let \( F_{d-1} \) indicate the input of \( d \)-th MSFB, the output \( F_d \) of MSFB can be defined as:

\[
\begin{align*}
C_3 &= \sigma \left( w_{3\times3}^1 \ast F_{d-1} + b^1 \right) \\
C_5 &= \sigma \left( w_{3\times3}^2 \ast F_{d-1} + b^2 \right) \\
C_7 &= \sigma \left( w_{3\times3}^3 \ast F_{d-1} + b^3 \right) \\
H_3 &= \sigma \left( w_{5\times5}^1 \ast \left[ C_3, C_5, C_7 \right] + b^4 \right) \\
H_5 &= \sigma \left( w_{5\times5}^2 \ast \left[ C_3, C_5, C_7 \right] + b^5 \right) \\
H_7 &= \sigma \left( w_{5\times5}^3 \ast \left[ C_3, C_5, C_7 \right] + b^6 \right) \\
F_d &= w_{3\times3}^1 \ast \left[ H_3, H_5, H_7 \right] + b^7 + F_{d-1}
\end{align*}
\]

where \( w \) is the convolutional layer weights, \( b \) indicates the bias terms, and the \( \sigma \) is the function of ReLU.

2.3. Contourlet transform
Wavelet transform has achieved success on representing one-dimensional signals. However, in higher dimensions, other types of singularities are usually present or even dominant, while wavelet is unable to handle them effectively. To solve this problem, Do and Vetterli [8] proposed contourlet transform, which aimed at keeping the advantages and avoid the shortcomings of wavelet transform. Specially, contourlet transform provides optimal approximation for a piecewise smooth function, and could better capture texture feature of image than spatial domain and wavelet domain. The detailed process of contourlet transform implementation can be found in [8].
In Figure 3, the LR face image is firstly decomposed by one level contourlet transform, and then the subbands of contourlet transform are served as the “Input” of CNN model. Finally, the five subbands generated by the CNN are as “Output” to reconstruct the SR results. In principle, the contourlet transform can be used in arbitrary CNN architecture, which efficient way to improve the SR performance. We will evaluate this contribution in Section 3.2.

3. Experiments
In this section, we evaluate our method on the public Celebrity Face Attributes (CelebA) dataset [9]. Here, the widely used PSNR and SSIM [10] are used as the evaluation measurements.

3.1. Implementation
We select 10% of the CelebA dataset, which includes 20K training samples and 260 testing images. And then, these images are aligned and cropped to 128×128 pixels as ground truth. One level contourlet transform is adopted in our method. Our network contains 10 MSFBs. The learning rate of all layers in our model is initialized to 0.0001, and decreases 50% for every 30 epochs. Specially, we train our model on the Tesla P100GPU.

3.2. Effectiveness of contourlet transform
In this paper, we formulate the face image SR task as the prediction of contourlet subbands. In principle, this contribution can be easily adapted to any arbitrary CNN architecture. Next, we will evaluate of this contribution. We use three recently proposed methods (IDN, MNCE and our method) and integrate them with contourlet transform prediction. We provide the comparison results of three approaches in Figure 4, where we can see that all of these methods improve significantly when they are integrated with contourlet transform. Such results demonstrate the effectiveness of contourlet transform prediction for face image SR task.

Figure 4. The effectiveness of contourlet transform
3.3. Qualitative and quantitative comparisons

Here, we compare our method with four recently proposed methods, which include CNN-based general image SR methods IDN [11], and three face specific image super-resolution methods, i.e., LCGE [12], UR-DGN [13], and MNCE [14]. We provide the PSNR and SSIM results in Table 1, where we find that our method performs better than other methods in terms of PSNR and SSIM.

Table 1. Average score in terms of PSNR (dB) and SSIM of different face SR methods.

| Metric | Bicubic | IDN | LCGE | UR-GAN | MNCE | Our method |
|--------|---------|-----|------|--------|------|-------------|
| PSNR   | 22.61   | 24.15| 23.35| 23.55  | 24.36| 24.68       |
| SSIM   | 0.6134  | 0.6854| 0.6673| 0.6696 | 0.6943| 0.7122      |

To better understand the advantages of our method, we also visualize some quantitative results in Figure 5. From these examples, we can see the curvature of face and the detailed texture are better reconstructed by our method. Overall, our method takes more structural correlation information into account by using contourlet transform, which produce more accurate edges textures than other methods.

![Figure 5](image)

Figure 5. Face super resolution results. (a) Bicubic interpolation, (b) IDN [11], (c) LCGE [12], (d) UR-GAN [13], (e) MNCE [14], (f) Our method and (g) Ground-truth.

4. Conclusion

In this paper, we propose a novel method for face image hallucination. To exploit feature of different receptive fields, we propose the MSRB, and based that construct the MSFC based on a set of cascaded MSFBs. In addition, we effectively combine our MSFC with the prediction of contourlet transform, which are able to reconstruct richer texture details. Extensive results demonstrate that the proposed approach outperforms the state-of-the-art methods.
Acknowledgments
This work was supported by the Open Project of National Railway Intelligent Transportation System Engineering Research Center under Grant No. RITS2018KF03.

References
[1] Chen Y, Tai Y, Liu Y M, Shen C H and Yang J 2018 FSRNet: End-to-end learning face super-resolution with facial priors. In Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah, USA, 2018, pp 2492-2501.
[2] Yang J, Luo L, Qian J J, Tai Y, Zhang F L and Xu Y 2017 Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol.39, No.1, pp 156-171.
[3] Schulter S, Leistner C and Bischof H 2015 Fast and accurate image upscaling with super-resolution forests. In CVPR, Boston, USA , pp 3791-3799.
[4] Dong C, Perez-Pellitero E, He K M and Tang X O 2016 Image super-resolution using deep convolutional networks. In PAMI, Vol.38, No.2, pp 295-307.
[5] Zhang Y L, Li K P, Li K, Wang L C, Zhong B N and Fu Y 2018 Image super-resolution using very deep residual channel attention networks. In: Proc. of the European Conference on Computer Vision (ECCV), Munich, Germany, pp 294-310.
[6] Zhang Y L, Tian Y P, Kong Y, Zhong B N and Fun Y 2018 Residual dense network for image super-resolution. In CVPR, Salt Lake City, Utah, USA, pp 2472-2481.
[7] Huang H B, He R, Sun Z N and Tan T N 2017 Wavelet-SRNet: a wavelet-based CNN for multiscale face super-resolution. In ICCV, Venice, Italy, pp 1689-1697.
[8] Do M N and Vetterli M 2005 The Contourlet Transform: An Efficient Directional Multi-resolution Image Representation. In TIP. Vol.14, No.12, pp 2091-2106.
[9] Liu Z W, Luo P, Wang X G and Tang X O 2015 Deep learning face attributes in the wild. In Proc. of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp 3730-3738.
[10] Wang Z, Bovik A C, Sheikh H R, Simoncelli E P 2005 Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (TIP). Vol.13, No.4, pp 600-612.
[11] Hui Z, Wang X M and Gao X B 2018 Fast and accurate single image super-resolution via information distillation network. In CVPR, Salt Lake City, Utah, USA, pp 723-731.
[12] Song Y B, Zhang J W, He S F, Bao L C and Yang Q X 2017. Learning to hallucinate face images via component generation and enhancement. In: Proc. of the International Joint Conference on Artificial Intelligence (IJCAI), Melbourne, Australia, pp 4537-4543.
[13] Yu X and Porikli F M 2016 Ultra-resolving face images by discriminative generative networks. In ECCV, Amsterdam, Netherlands, pp 318-333.
[14] Jiang J J, Yu Y, Hu J H, Tang S H and Ma J Y 2018 Deep CNN denoiser and multi-layer neighbor component embedding for face hallucination. In IJCAI, Stockholm, Sweden, pp 1542-1548.