A JOINT LEARNING BASED FACE SUPER RESOLUTION APPROACH VIA CONTEXTUAL TOPOLOGICAL STRUCTURE

Liang Chen, Ruimin Hu, Zhen Han, Zhongyuan Wang, Qing Li

State Key Laboratory of Software Engineering, Wuhan University
Collaborative Innovation Center of Geospatial Technology, Wuhan
NERCMS, School of Computer, Wuhan University
MERC, Department of Computer Science, City University of Hong Kong, Hong Kong

ABSTRACT

Face Super Resolution(FSR) is to infer High Resolution(HR) facial images from given Low Resolution(LR) ones with the assistance of LR and HR training pairs. Among existing methods, local patch based methods are superior in visual and objective quality than global based methods. These local patch based methods are based on the consistency assumption that the neighbors in HR/LR space form similar local geometry. But when LR images are with low quality, the LR space is seriously contaminated that even two distinct patches look similar, which means that the consistency assumption is not well held anymore. To this end, in this paper we introduce the contextual topological structure of target patch to improve the consistency. The contextual topological structure consists of the target patch as well as its adjacent patches, we explore the relationship between them based on statistical probability and apply the relationship for joint learning progress of mapping from LR to HR. By incorporating the contextual topological structure, the robustness to noise of approach is increased as well as the LR/HR consistency. The effectiveness of proposed method is verified both quantitatively and qualitatively.

Index Terms— face super resolution, contextual topological structure, consistency enhancement, low level vision

1. INTRODUCTION

Face hallucination, also known as Face Super Resolution(FSR), is to infer High Resolution(HR) facial images from given Low Resolution(LR) facial images based on HR/LR training sample pairs. Considering that face is a highly structured object and position prior plays a crucial role in SR process, FSR is a domain-specific SR technique focus on face images specifically. FSR is also extensively used for pre- and/or post-processing in video applications, such as video surveillance of smart city.

Since it was firstly proposed by Baker et al. [1], a large number of FSR approaches have been developed in the past decades. Because face is a highly structured object, the position prior plays vital role in FSR. Based on different position priors applied in the methods, current FSR methods can be divided into two categories: global based methods [2] [3] and local patch based methods [4] [5] [6] [7].

The global based methods apply global facial structure as prior, i.e., they process facial images as a whole [2] [3] [8]. The typical global based methods usually view whole facial image as a data and decompose LR/HR training images to obtain LR/HR basis for representation. It usually involves several steps in its solution pipeline: give an input LR facial image, firstly it is decomposed in LR space and represented with the LR basis, then the representation coefficients are mapping into HR space. The output HR estimation is obtained through the integration of HR basis and coefficients.

Compare to global based methods, local patch based methods change the scope of attention from the whole face to the local facial patch [9][10] [11] [12], i.e., the whole learning process is composed of multiple sub-learning processes that are just the same with the global learning process only focus on the local facial patches. Because the effort in detail refinement, these methods can provide better visual details and higher magnification factors. In this paper, we focus on this category in this paper.

Previous researchers have proposed a lot of work to promote the local framework. One category focuses on learning the relevant subspace from training data for specific input [9][10] [11] [12]. Another category focuses on refinement of coefficients [13] [6] [7]. In addition, a lot of deep learning based methods have been proposed to simplify the LR to HR learning process as an fine tuned end to end black box [14][15][16]. In [16], Dong et al. have shown that the aforementioned pipeline is equivalent to a deep convolutional neural network, they are considered as the sub-class of the local based method.

Methods in local patch category are usually based on the consistency assumption: HR/LR spaces form similar local geometry structures. The HR neighbors of HR image patch and the LR neighbors of its LR version are HR/LR pairs. But the degradation decreases the patch distinctiveness, and deteriorates the one-to-many problem, which leads to the fact that the consistency assumption does not well hold anymore.

Previous FSR works usually apply new computing tools(sparse [13], deep neural network[15],.etc) to solve the FSR problem based on the traditional pipeline. But there is still much room for FSR improvement in perspective of facial pattern and prior exploration, which will also change the current FSR framework enormously.

In this paper, we propose a patch based FSR method via local facial topological structure to enhance consistency. The local facial topological structure applies the adjacent patches of target patch and the relation between them. The adjacent patches are used to provide extra structure for target patch to improve the distinctiveness of target patch, the consistency increases when distinctiveness increases. The relation between the target patch and the adjacent patch

is to determine the contribution of adjacent patch in distinctiveness improvement of target patch, after all, not every adjacent patch is helpful for target patch. In this way, the consistency assumption can well hold under degradations.

2. PROPOSED METHOD

2.1. Motivation

The adjacent patches can increase the patch distinctiveness even under degradation. That is the reason we apply adjacent patch for topological composition. See Fig.1, the patch in blue box denotes the target patch, the patch in black box denotes the combination of target patch and adjacent patches of target patch. Then we give the similarity comparison of with/without contextual information for HR/LR patch pair of same person and LR/LR patches of different persons. From the comparison we can see, after applying the contextual information of adjacent patch, the consistency between same person increases, and the distinctiveness between different persons increases even in LR degradations.

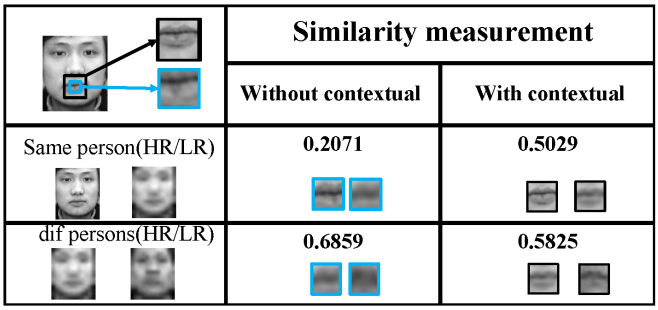

Fig. 1. The distinctiveness advantage of contextual model.

2.2. Overview

There are mainly two stages in proposed method: the training stage and the testing stage. In the training stage, the local topological structures of every facial position are adaptively built based on HR training samples. In the testing stage, every patch of input LR image finds its the corresponding local topological structure, and apply each node of the topological structure in joint learning of HR patch, the contribution of every node is determined by the weight of the node in the local undirected graph. The overview of proposed method is in Fig.2.

2.3. Notions

Suppose we have a training set contains n HR/LR image pairs denoted as $Y = \{y_j\}_{j=1}^n$ and $X = \{x_j\}_{j=1}^n$ respectively. Every image in the training set is divided into p small patches. All the i^{th} $(i = 1, 2, ..., p)$ patch of HR training data is denoted as $\{y_j^i\}_{j=1}^n$, and the corresponding LR patch set is $\{x_j^i\}_{j=1}^n$. The input LR face image is denoted as x_{in}. The i^{th} patch of x_{in} is denoted as x_{in}^i.

2.4. Training Stage

In the training stage, only the HR training samples are used. For a specific patch setting, we get a graph set $G = \{g^i\}_{i=1}^p$ from HR training samples. Every graph g^i corresponds to a patch position i. A graph, a.k.a., a local facial topological structure g^i consists of nodes and the relations between nodes.

The nodes come from 9 adjacent patches of the target patch v^i, denoted as $\{v_1^i, ..., v_{num}^i\}$, where num is the number of nodes, because the maximum number of adjacent patch is 9, $num = 1, 2, ..., 9$. Because not all the adjacent patches are helpful in joint learning, only some of them are selected and applied in learning process. Therefore, for different i, different number of nodes are selected, thus different topological structures of graphs are built.

The relations of a graph are the weights between every two patches(target and adjacent), denoted as $\{\omega_1^i, ..., \omega_{num}^i\}$. The target patch as well as its adjacent patches are used to compose a undirected weighted graph, every weight of the undirected graph is obtained through the texture similarity between corresponding two patches. In this paper, we adopt a widely used feature: HOG (Histogram of Oriented Gradient [17]) as the texture feature for its comprehensive performance in description of texture. Therefore, the obtainment of weight can be formulized as:

$$\begin{aligned} &\omega_1^i = p\left(v_1^i|v^i\right) \\ &\omega_2^i = p\left(v_2^i|v^i\right) \\ &.... \\ &\omega_num^i = p\left(v_num^i|v^i\right) \end{aligned} \tag{1}$$

And we estimate the $p\left(v_1^i|v^i\right)$ in this form:

$$\begin{aligned} &p\left(v_1^i|v^i\right) \propto \exp(-\left\|HOG(v^i) - HOG(v_1^i)\right\|_2^2/h^2) \\ &p\left(v_2^i|v^i\right) \propto \exp(-\left\|HOG(v^i) - HOG(v_2^i)\right\|_2^2/h^2) \\ &.... \\ &p\left(v_{num}^i|v^i\right) \propto \exp(-\left\|HOG(v^i) - HOG(v_{num}^i)\right\|_2^2/h^2) \end{aligned} \tag{2}$$

where h is the length of patch side, HOG is to distinguish hog features from pixel features. Then we use all patches in HR training set to estimate the p:

$$\begin{aligned} &p\left(v_1^i|v^i\right) \propto \exp(-\left\|HOG(v^i) - HOG(v_1^i)\right\|_2^2/h^2) \\ &= \exp(-\sum_{j=1}^n \left\|HOG(y_j^i) - HOG(y_{1,j}^i)\right\|_2^2/h^2) \\ &p\left(v_2^i|v^i\right) \propto \exp(-\left\|HOG(v^i) - HOG(v_2^i)\right\|_2^2/h^2) \\ &= \exp(-\sum_{j=1}^n \left\|HOG(y_j^i) - HOG(y_{2,j}^i)\right\|_2^2/h^2) \\ &.... \\ &p\left(v_{num}^i|v^i\right) \propto \exp(-\left\|HOG(v^i) - HOG(v_{num}^i)\right\|_2^2/h^2) \\ &= \exp(-\sum_{j=1}^n \left\|HOG(y_j^i) - HOG(y_{num,j}^i)\right\|_2^2/h^2) \end{aligned} \tag{3}$$

where $y_{num,j}^i$ denotes the num^{th} adjacent patch of y_j^i.

Then we get graph g^i: node set $\{v_1^i, ..., v_{num}^i\}$ and weight set $\{\omega_1^i, ..., \omega_{num}^i\}$. We put the graph g^i into G, and use G as look up table in test stage. After go through all the position i, the G is obtained, and the training stage is over.

2.5. Testing Stage

In the testing stage, break the given input LR image x_{in} into patches firstly. For its patch in position i, i.e.,x_{in}^i, go to graph set G and find the g^i, obtain the node set $\{v_1^i, ..., v_{num}^i\}$ and weight set $\{\omega_1^i, ..., \omega_{num}^i\}$. Use the nodes to joint represent the x_{in} in turn, we can obtain combination set $\{C_1^i, ..., C_{num}^i\}$. Combine the $\{x_j^i\}_{j=1}^n$

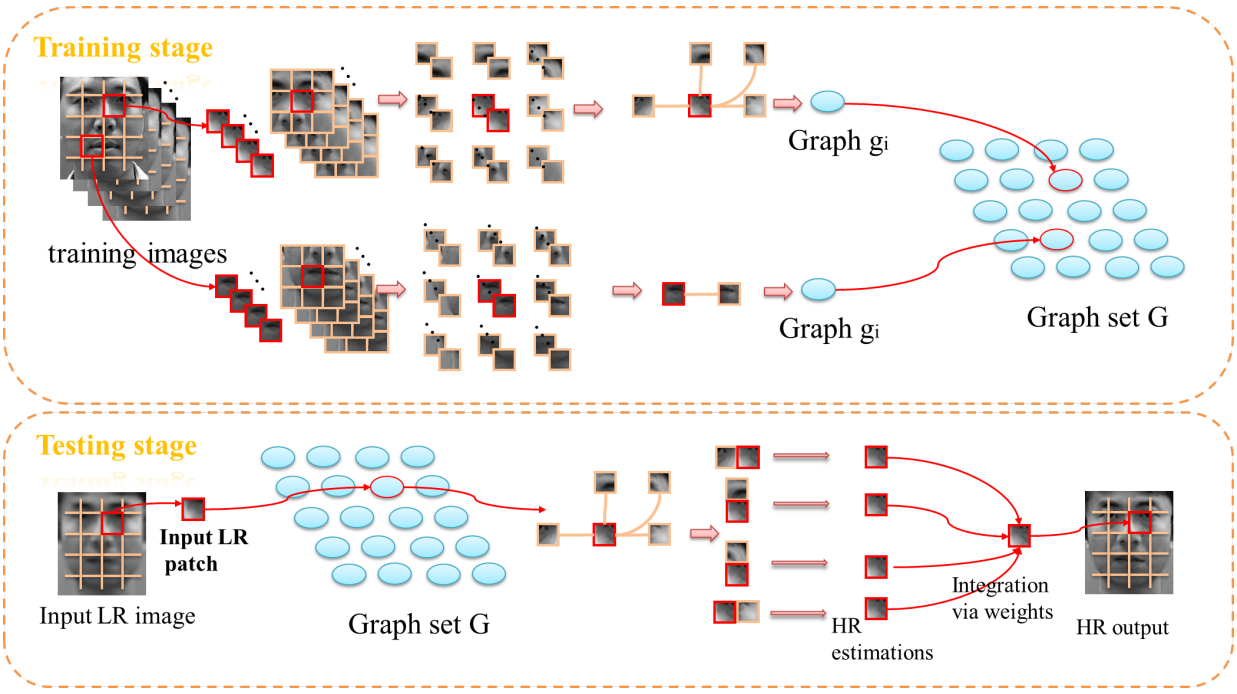

Fig. 2. The overview of first stage.

with its num^{th} adjacent patch set in the same way, we can get the combination of training samples $\{x^i_{j,1}\}^n_{j=1},...,\{x^i_{j,num}\}^n_{j=1}$. The C^i_1 can be coded as weighted combination of $\{x^i_{j,1}\}^n_{j=1}$ in form:

$$\begin{aligned} C^i_1 &\approx x^i_{1,1}\gamma^i_{1,1} + x^i_{2,1}\gamma^i_{2,1} + ... + x^i_{n,1}\gamma^i_{n,1} \\ C^i_2 &\approx x^i_{1,2}\gamma^i_{1,2} + x^i_{2,1}\gamma^i_{2,2} + ... + x^i_{n,2}\gamma^i_{n,2} \\ &.... \\ C^i_{num} &\approx x^i_{1,num}\gamma^i_{1,num} + x^i_{2,num}\gamma^i_{2,num} + ... + x^i_{n,num}\gamma^i_{n,num} \end{aligned} \tag{4}$$

where $\gamma^i_{j,num}$ is the j^{th} coefficient of C_num^i, $j = 1, 2, ..., n$. Then we can get the num HR reconstructions: $y^{i,1}_{out}, ..., y^{i,num}_{out}$, where:

$$\begin{aligned} y^{i,1}_{out} &= y^i_1\gamma^i_{1,1} + y^i_2\gamma^i_{2,1} + ... + y^i_n\gamma^i_{n,1} \\ y^{i,2}_{out} &= y^i_1\gamma^i_{1,2} + y^i_2\gamma^i_{2,2} + ... + y^i_n\gamma^i_{n,2} \\ &.... \\ y^{i,num}_{out} &= y^i_1\gamma^i_{1,num} + y^i_2\gamma^i_{2,num} + ... + y^i_n\gamma^i_{n,num} \end{aligned} \tag{5}$$

The final HR estimation is obtained through weights $\{\omega^i_1, ..., \omega^i_{num}\}$ in following form:

$$y^{i,num}_{out} = \sum_{j=1}^{num} y^{i,j}_{out}\omega^i_j \Big/ \sum_{j=1}^{num} \omega^i_j \tag{6}$$

2.5.1. Relation To Prior Work

In order to verify the advantage of proposed linear contextual graph in degradation resistance, we compare our method with the prior approach [18]. The local structure explored in [18] is in pixel level, which is easily distorted by degradations and relies heavily on accurate facial alignments. We all know that the accurate facial alignments is generally not possible in typical very low quality scenarios such as surveillance. In addition to the patch-level contextual information, we further tailor the contextual model by applying the possibility of distinctiveness texture, provide different positions different topological structures adaptively, which improve the efficiency.

3. EXPERIMENTS

In this section, we perform experiments on the CAS-PEAL-R1 database [19] to verify the proposed method. There are 30867 images of 1040 subjects in this database. We select 1040 images with neutral expression and frontal pose of 1040 subjects, and divide them into a training set (1000 images) and a test set (40 images) randomly. Therefore the test images are not in the training set anymore. The images are aligned by five manually selected feature points and cropped to 112 $\times$ 96 pixels. In this paper, the original patch size of test image is 12 $\times$ 12 pixels with 8 overlapped pixels. The size of adjacent patch is 12 $\times$ 12 pixels.

3.1. Simulation Experiment on Low Quality Images

In order to simulate the images with low quality degradation, we give two degradations: noise and blurry. For noise, the test images are down-sampled by a factor of 4 to LR 28 $\times$ 24 pixels images and a white Gaussian noise is added with σ 0.015. For blurry, images are down-sampled by a factor of 8 to LR 14 $\times$ 12 pixels images.

The baseline [3] is a classical global based method and the rest [12] [7] [4], and [18] are classical local based methods. We can conclude that the proposed method outperforms existing methods both qualitatively (Fig.3) and quantitatively (Peak Signal-to-Noise Ratio(PSNR) and Structural Similarity(SSIM)[20], Table 1). The same conclusion can be made from the results under blurry.

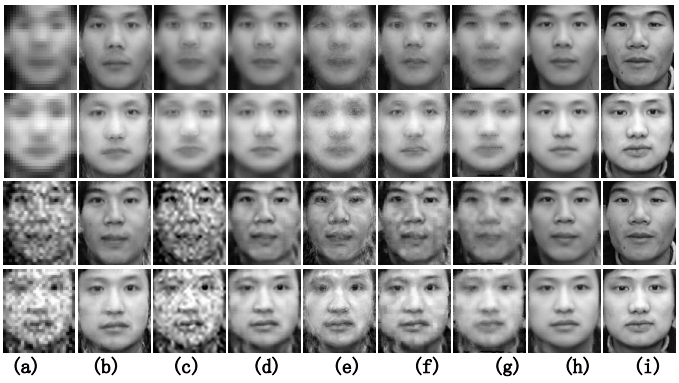

Fig. 3. Results of LR images with noise (The upper two rows) and blur (the bottom two rows). Column:(a) nearest interpolation; (b) NE(24×24); (c) [12]; (d) [7]; (e) [3]; (f) [4]; (g) [18]; (h)ours; (i) HR groundtruth. (The added noise after down-sampling process can be enlarged and strengthened to a great extent, therefore the input images can be in low quality eventhough sigma is only 0.015)

Table 1. Objective comparison.

	noise		blurry	
algorithm	PSNR	SSIM	PSNR	SSIM
[12]	19.3123	0.4948	21.5349	0.6214
[7]($\tau = 5$)	21.7215	0.6616	21.3953	0.6224
[3]	21.6226	0.6569	21.0023	0.5815
[4]	21.8865	0.6711	21.5947	0.6859
[18]	22.9102	0.7391	22.0624	0.7022
NE	20.6077	0.6006	21.0290	0.6076
proposed	**23.3262**	**0.7667**	**22.2280**	**0.7023**

3.2. Consistency performance

We evaluate the consistency performance of the proposed method under serious blur and noise degradation, measure by the Neighbor Preservation Ratio(NPR)[21]. As shown in Fig.4. Two kinds of degradations are given—serious blur and serious noise (same with section3.1). From these figures in Fig.4 we can see, the consistency can be better preserved with contextual information. This should be attributed to the effective structural compensation provided by the contextual topology, which corrects the wrong neighbor relation of local patches caused by serious degradations.

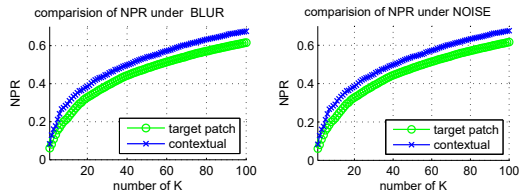

Fig. 4. NPR comparison with degradations. Left: blurry; right: noise. The x-axis indicates numbers of Ks, the y-axis is the NPR performances. Each performance with fixed K is the average of 506 patches in test images. The blue line and the green line denote the NPR with and without contextual information respectively.

3.3. Experiment on Real World Low Quality Images

We also evaluate proposed method on real world images: facial images captured in surveillance scenario with underexposure, and the results are shown in Fig.5. We can see that proposed method outperforms other baselines in details as well.

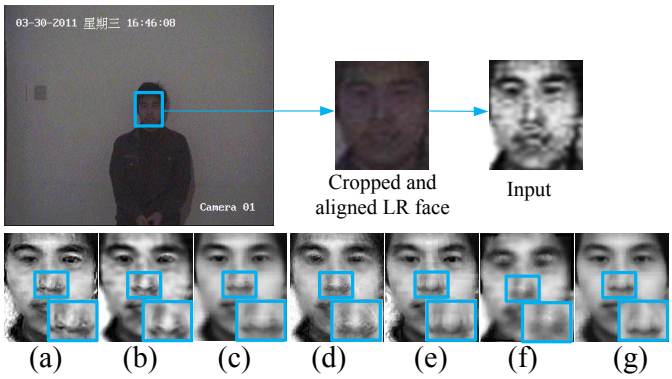

Fig. 5. Results of surveillance images. (a)NE; (b) [12]; (c) [7]; (d) [3]; (e) [4]; (f) [18]; (g)proposed.

4. CONCLUSION

In this paper, we present a joint learning based face super resolution approach via contextual topological structure of undirected graph. The wrong neighbor relation can be reduced by introducing the contextual topological structure. The effectiveness of proposed method is verified both quantitatively and qualitatively. Our future work will focus on the multi-view face hallucination and facial sketch synthesis.

5. ACKNOWLEDGEMENT

The research is supported by the National Nature Science Foundation of China (No. 61671332,61231015, 61172173, 61303114, U1404618, 61501413, 61502354), the National High Technology Research and Development Program of China (863 Program) (No. 2015AA016306, 2013AA014602); the Natural Science Fund of Hubei Province (2016CFB573); the Internet of Things Development Funding Project of Ministry of industry in 2013(No. 25); Applied Basic Research Program of Wuhan City (2016010101010025); the Technology Research Program of Ministry of Public Security (No. 2014JSYJA016); the China Postdoctoral Science Foundation funded project (2013M530350,2014M562058); the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130141120024); the Fundamental Research Funds for the Central Universities(2042014kf0025). This work was partly supported by the EU FP7 QUICK project under Grant Agreement No. PIRSESGA-2013-612652.

6. REFERENCES

[1] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *Pattern Analysis and Machine Intelligence,*

IEEE Transactions on, vol. 24, no. 9, pp. 1167–1183, Sep 2002.

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

[3] Chengdong Lan, Ruimin Hu, Zhen Han, and Zhongyuan Wang, "A face super-resolution approach using shape semantic mode regularization," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, Sept 2010, pp. 2021–2024.

[4] Xiaohui Dong, Ruimin Hu, Junjun Jiang, Zhen Han, Liang Chen, and Ge Gao, "Noise face image hallucination via data-driven local eigentransformation," in *Pacific Rim Conference on Multimedia*. Springer, 2014, pp. 183–192.

[5] Hua Huang, Huiting He, Xin Fan, and Junping Zhang, "Super-resolution of human face image using canonical correlation analysis," *Pattern Recognition*, vol. 43, no. 7, pp. 2532 – 2543, 2010.

[6] Xiang Ma, Junping Zhang, and Chun Qi, "Position-based face hallucination method," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, June 2009, pp. 290–293.

[7] Junjun Jiang, Ruimin Hu, Zhongyuan Wang, and Zhen Han, "Noise robust face hallucination via locality-constrained representation," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1268–1281, 2014.

[8] Soheil Kolouri and Gustavo K Rohde, "Transport-based single frame super resolution of very low resolution face images," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 4876–4884.

[9] Hong Chang, Dit-Yan Yeung, and Yimin Xiong, "Super-resolution through neighbor embedding," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, June 2004, vol. 1, pp. I–I.

[10] Liang Chen, Ruimin Hu, Chao Liang, Qing Li, and Zhen Han, "A novel face super resolution approach for noisy images using contour feature and standard deviation prior," *Multimedia Tools and Applications*, pp. 1–27, 2016.

[11] R. Timofte, V. De, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, Dec 2013, pp. 1920–1927.

[12] Hua Huang and Ning Wu, "Fast facial image super-resolution via local linear transformations for resource-limited applications," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 10, pp. 1363–1377, Oct 2011.

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

[14] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang, "Deep cascaded bi-network for face hallucination," *arXiv preprint arXiv:1607.05046*, 2016.

[15] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin, "Learning face hallucination in the wild.," in *AAAI*, 2015, pp. 3871–3877.

[16] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[17] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[18] Zhuo Hui and Kin-Man Lam, "An efficient local-structure-based face-hallucination method," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1265–1268.

[19] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao, "The cas-peal large-scale chinese face database and baseline evaluations," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 38, no. 1, pp. 149–161, Jan 2008.

[20] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[21] Kevin Su, Qi Tian, Qing Xue, Nicu Sebe, and Jingsheng Ma, "Neighborhood issue in single-frame image super-resolution," in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 4–pp.