Pattern Recognition 47 (2014) 1261–1270

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Face hallucination based on sparse local-pixel structure

Yongchao Li [a,1], Cheng Cai [a,*], Guoping Qiu [b,c], Kin-Man Lam [d]

[a] Department of Computer Science, College of Information Engineering, Northwest A&F University, Xi'an, China
[b] School of Computer Science, University of Nottingham, United Kingdom
[c] International Doctoral Innovation Centre, The University of Nottingham, Ningbo, China
[d] Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Article history:
Received 21 January 2013
Received in revised form
28 June 2013
Accepted 16 September 2013
Available online 25 September 2013

Keywords:
Face hallucination
Sparse local-pixel structure
Super-resolution
Sparse representation

ABSTRACT

In this paper, we propose a face-hallucination method, namely face hallucination based on sparse local-pixel structure. In our framework, a high resolution (HR) face is estimated from a single frame low resolution (LR) face with the help of the facial dataset. Unlike many existing face-hallucination methods such as the from local-pixel structure to global image super-resolution method (LPS-GIS) and the super-resolution through neighbor embedding, where the prior models are learned by employing the least-square methods, our framework aims to shape the prior model using sparse representation. Then this learned prior model is employed to guide the reconstruction process. Experiments show that our framework is very flexible, and achieves a competitive or even superior performance in terms of both reconstruction error and visual quality. Our method still exhibits an impressive ability to generate plausible HR facial images based on their sparse local structures.

1. Introduction

The idea of super-resolution (SR) was first presented by Tsai and Huang [1], and significant progress has been made with it over the last few decades. Since SR is an ill-posed problem, prior constraints are necessary to attain a good performance. Based on the different approaches to attaining these prior constraints, SR methods can be broadly classified into two categories: one is the conventional approach, which is also widely known as multi-image SR [2–5] or regularization-based SR, and which reconstructs a HR image from a sequence of LR images of the same scene. These algorithms mainly employ regularization models to solve the ill-posed image SR, and use smooth constraints as the prior constraints, which are defined artificially. The other approach is single-frame SR [6–11], which is also called learning-based SR or example-based SR. These methods generate a HR image from a single LR image with the information learned from a set of LR–HR training image pairs. These algorithms attain the prior constraints between the HR images and the corresponding LR images through a learning process. Many example-based or learning-based algorithms [6–16] have been proposed in the field of image processing. Also in SR, Qiu [13] and Baker and Kanade [17] have demonstrated that the smooth prior constraints used in many regularization-based methods will become less effective at solving the SR problem as the zooming factor increases, while example-based approaches have the potential to overcome this problem using advances in machine learning and computer vision. In this paper, we focus on the single-image SR problem.

Fig. 1 shows a general framework of example-based SR: the input LR image is first interpolated, using the conventional methods, to the size of the target HR image, and the input interpolated LR image – a blurry image lack of high-frequency information – is then used as the initial estimation of the target HR. The input LR image is also divided into either overlapping or non-overlapping image patch, and the example-based framework will use the image patches to find out the most matched examples by searching a training dataset of LR–HR image pairs. The selected HR examples are then employed to learn the HR information as the prior constraints. Finally, the learned HR information and the input interpolated image are combined to evaluate the target HR image.

The idea of face hallucination was first proposed by Baker and Kanade [18], and it was then used for SR problems in [8,11,16,19]. Example-based face hallucination is a subcategory of the

Abbreviations: SR, super resolution; HR, high resolution; LR, low resolution; PSNR, peak signal to noise ratio; SSIM, Structural Similarity Index; SRM, sparse representation models; NARM, nonlocal autoregressive model

* Corresponding author. Tel.: +86 1899 1291 338; fax: +86 029 87092353.
E-mail addresses: ychao_lee@nwsuaf.edu.cn (Y. Li), cheney.chengcai@gmail.com (C. Cai), qiu@cs.nott.ac.uk (G. Qiu), enkmlan@polyu.edu.hk (K.-M. Lam).
[1] Postal address: Room 322, College of Information Engineering, Northwest A&F University, Yangling, Xi'an, Shaanxi 712100, China.

0031-3203
http://dx.doi.org/10.1016/j.patcog.2013.09.012

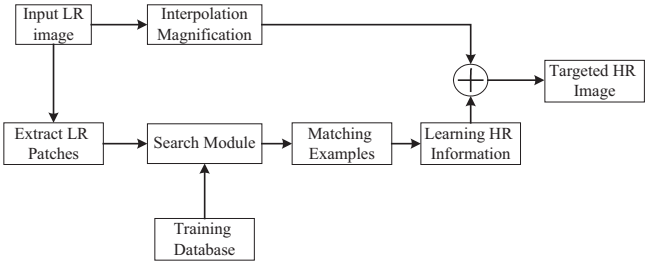

Fig. 1. A general framework of example-based SR methods.

framework shown in Fig. 1, and it is also a specific and important category of image SR. In [20], Liang conducted a good survey of face hallucination, and summarized the existing face-hallucination methods into two approaches: namely global similarity and local similarity. The theoretical backgrounds and practical results of the existing face-hallucination techniques and algorithms are compared. Based on the comparison results, the strengths and weaknesses of each algorithm are summarized, which forms a base for proposing an effective method to hallucinate mis-aligned face images.

In [16], Liu et al. argued that a successful face-hallucination algorithm should meet the following three constraints:

1. *Sanity constraint*: the target HR image should be very close to the input LR image when smoothed and down-sampled.
2. *Global constraint*: the target HR image should have the common characteristics of human faces, e.g., possessing a mouth and a nose, being symmetrical, etc.
3. *Local constraint*: the target HR image should have the specific characteristics of the original LR face image, with photorealistic local features.

Furthermore, a two-step approach was developed for face hallucination, in which a Bayesian formulation and a nonparametric Markov network are employed to deal with face hallucination. Of all the various methods for face hallucination, Hu et al. [11] first learned local-pixel structures from the most matched example images explicitly, which are then used directly as reconstruction priors. A three-stage face-hallucination framework was proposed in [11], which is called Local-Pixel Structure to Global Image Super-Resolution (LPS-GIS). In Stage 1, k pairs of example faces which have a similar pixel structure to the input LR face are selected from a training dataset using k-Nearest Neighbors (KNN). They are then subjected to warping using optical flow, so that the corresponding target HR image can be reconstructed more accurately. In Stage 2, the LPS-GIS method learns the face structures, which are represented as coefficients using a standard Gaussian function; the learned coefficients are updated according to the warped errors. In Stage 3, LPS-GIS constrains the revised face structures, namely the revised coefficients to the input LR face, and then reconstructs the target HR image using an iterative method.

Unlike the abovementioned methods, we propose a face-hallucination framework which utilizes the sparse local-pixel structure as the prior model in the reconstruction of HR faces. The use of sparse local-pixel structure allows our method to reconstruct the details in HR faces flexibly. Furthermore, the global structure of faces is also considered, which enables the proposed method to produce plausible facial components.

As for the organization of this paper, Section 2 gives a brief introduction to the theory of sparse representation and its recent applications to super-resolution. Section 3 provides a detailed introduction to a concept called 'local-pixel structure with sparsity'. The details of our proposed framework are presented in Section 4. Section 5 presents the experiments and an evaluation of the proposed framework. Finally, the concluding remarks are given in Section 6.

2. Related works on super-resolution with sparse representation

Single-image super-resolution produces a reconstructed HR image from an input LR image using the prior knowledge learned from a set of LR–HR training image pairs, and the reconstructed HR image should be consistent with the LR input. An observed model between a HR image and its corresponding LR counterpart is given as follows:

$$\mathbf{I_l} = \mathbf{I_h}\mathbf{H}S(r) + \mathbf{N}, \tag{1}$$

where $\mathbf{I_l}$ and $\mathbf{I_h}$ denote the LR and HR images, respectively; $\mathbf{H}$ represents a blurring filter; $S(r)$ is a down-sampling operator with a scaling factor of r in the horizontal and vertical dimensions; and $\mathbf{N}$ is a noise vector, such as the Gaussian white noise. Here, we will focus on the situation whereby the blur kernel is the Dirac delta function as [11,44], i.e. $\mathbf{H}$ is the identity matrix. Thus, Eq. (1) can be rewritten as follows:

$$\mathbf{I_l} = \mathbf{I_h}\mathbf{H}S(r) + \mathbf{N}. \tag{2}$$

Therefore, the purpose of SR is to recover as much of the information lost in the down-sampling process as possible. Since the reconstruction process still remains ill-posed, different priors can be used to guide and constrain the reconstruction results. In recent years, the sparse representation model (SRM) has been used as the prior model, and has shown promising results in image super-resolution.

Sparse representation of a signal is based on the assumption that most or all signals can be represented as a linear combination of a small number of elementary signals only, called atoms, from an overcomplete dictionary. Compared with other conventional methods, sparse representation can usually offer a better performance, with its capacity for efficient signal modeling [21]. The sparse representation of signals has already been applied in many fields, such as object recognition [22,23], text categorization [24], signal classification [21], etc.

In the sparse representation, a common formulation of the problem of finding the sparse representation of a signal using an overcomplete dictionary is described as follows:

$$\hat{\omega}_0 = \min \|\omega\|_0, \quad \text{s.t.} \quad \psi = \mathbf{A}\omega, \tag{3}$$

where $\mathbf{A}$ is an $M \times N$ matrix whose columns are the elements of the overcomplete dictionary, with $M < N$, and $\psi \in R^{M\times 1}$ is an observational signal. The purpose of sparse representation is to find an $N \times 1$ coefficient vector ω, which is considered to be a sparse vector, i.e. most of its entries are zeros, except for those elements in the overcomplete dictionary $\mathbf{A}$ which are associated with the observational signal ψ. Solving the sparsest solution for (3) has been found to be *NP*-hard, and it is even difficult to approximate [25]. However, some recent results [26,27] indicate that if the vector ω in (3) is sparse enough, then the problem can be solved efficiently by minimizing the ℓ_1-norm instead, as follows:

$$\hat{\omega}_1 = \min \|\omega\|_1, \quad \text{s.t.} \quad \psi = \mathbf{A}\omega. \tag{4}$$

In fact, as long as the number of nonzero components in ω_0 is a small fraction of the dimension M, the ℓ_1-norm can replace and recover the ℓ_0-norm efficiently [22]. In addition, the optimization problem of the ℓ_1-norm can be solved in polynomial time [28,29]. However, in real applications, the data in the dictionary $\mathbf{A}$ are, in general, noisy. This will lead to the result whereby the sparse representation of an observational signal, in terms of the training data in $\mathbf{A}$, may not be accurate. In order to deal with the problem, (4) can be relaxed to a modified form as follows:

$$\hat{\omega}_1 = \min \|\omega\|_1, \quad \text{s.t.} \quad \|\psi - \mathbf{A}\omega\|_2 \le \varepsilon. \tag{5}$$

Lagrange multipliers offer an equivalent formulation, as shown in the following equation:

$$\arg\min \tfrac{1}{2}\|\mathbf{A}\omega-\psi\|_2^2+\lambda\|\omega\|_1, \tag{6}$$

where $\lambda \in R^+$ is a regularization parameter which balances the sparsity of the solution and the fidelity of the approximation to ψ. This is actually a typical convex-optimization problem, and it can be efficiently solved using the method of Large-Scale L_1 Regularized Least Squares (L1LS) [30].

In [41–43], Yang et al. used sparse representation for face hallucination. The proposed method, denoted as ScSR, is based on the idea of sparse signal representation whereby the linear relationships among HR training signals can be accurately recovered from their low-dimensional projections. The structures of LR images are used to form a sparse prior model, which is then employed to reconstruct the HR images or HR patches. The differences between ScSR and our method are that ScSR represents image patches as a sparse linear combination of elements from an appropriately chosen overcomplete dictionary, while in our method, a pixel is represented as a sparse linear combination of elements from its neighboring pixels. Our method seeks a sparse representation for each patch of the LR input image from an LR overcomplete dictionary; the coefficients of this representation are then used to generate the HR target image using the HR overcomplete dictionary. One important process in ScSR is the training of two dictionaries for the LR and HR image patches. In our method, central pixels replace patches, and only the HR dictionary is needed; it is constructed directly from the HR example faces.

In [44], Dong et al. also proposed an image interpolation method based sparse representation, which is abbreviated as NARM-SRM-NL. In the method, a nonlocal autoregressive model (NARM) was proposed and taken as the data-fidelity term in the sparse representation model (SRM). The patches in the estimated HR image are reconstructed using the nonlocal neighboring patches. The method assumes that the nonlocal similar patches in an image have similar coding coefficients with the same overcomplete dictionary; the coefficients are then embedded into SRM and NARM to reconstruct the HR images.

Although a lot of the literature, including this paper, has employed sparse representation to deal with the SR problems, differences are still very obvious. For example, sparse models are diverse, which will lead to different ways of constructing the overcomplete dictionary. In ScSR, the method assumes that image patches can be well represented as a sparse linear combination of elements from a specific dictionary, and a pair of HR–LR dictionaries is constructed to force LR–HR patches to have the same sparse coefficients. In NARM-SRM-NL, the sparse model employed assumes that an image patch can have many similar patches among its nonlocal neighboring patches, and the local PCA dictionary is used to span adaptively the sparse domain for signal representation. In our method, the sparse local-pixel structure is proposed and the dictionary is constructed using the neighboring pixels of those missing pixels.

3. Local-pixel structure with sparsity

As we know, a HR face and its corresponding LR face have a common global face structure. Therefore, we can assume that they also have similar local-pixel structures, and that the local image information about the input LR alone should be sufficient to predict the missing HR details. In our algorithm, we use the neighboring pixels of a missing pixel to estimate the target HR face. This idea is similar to neighbor embedding in [6,10]. The following formulation describes the model used in our method:

$$I(x,y)=\sum_{\mu,\nu\in C}\alpha_{\mu,\nu}(x,y)\times I(x+\mu,y+\nu), \tag{7}$$

where $I(x, y)$ is a pixel at location (x, y), $\alpha_{\mu,\nu}(x,y)$ denotes the weight of the neighboring pixel $I(x+\mu,y+\nu)$ contributed to the pixel $I(x, y)$ with a relative displacement of (μ, ν), μ and ν cannot be zero at the same time, and C denotes a local window centered at (x, y).

Based on the above assumption of similar pixel structures between the HR–LR face pairs, the weights are almost the same at the same position in a HR face and its corresponding LR face image. Our algorithm searches k similar LR example faces to the input LR face from a dataset of LR–HR face pairs. Then, the neighboring weights of the pixel structures in the k-HR example faces are utilized to estimate the information lost in the input LR face. In order to learn the embedded weights for the central pixels or patches, [6,10,11] used the ℓ_2-norm methods, such as Gaussian functions and least-square methods. However, a high visual quality image is very sharp because it contains sharp edges, high-frequency information, and discontinuities. A sharp image means that the local-pixel structures have some sparse properties, and this can be interpreted as indicating that the pixel $I(x, y)$ in (7) can be better reconstructed using only a fraction of the neighboring pixels. Thus, our algorithm will use sparse representation as the prior model to learn the embedded weights.

In the previous section, we stated that many face-hallucination methods use the ℓ_2-norm model to learn the prior knowledge, and we believe that the local-pixel structures in a high visual quality image exhibit the sparse property. In other words, most of the neighboring pixels of the pixel $I(x, y)$ in (7) can be regarded as outliers. Fig. 2 shows a central pixel with its neighboring pixels, and it fully represents the sparse property of an image patch. It is obviously observed that only a small number of neighbors are closely related to the central pixel, while others can be regarded as outliers. It has been demonstrated in [22,31–34] that, compared with the ℓ_1-norm, the ℓ_2-norm is less robust and more sensitive to outliers, which will usually decrease the accuracy of image super-resolution. Furthermore, in [22,34], the use of the ℓ_1-norm produces more robust and better results for face image super-resolution and face recognition, respectively, especially when the signal is sparse and discontinuous.

A toy example is provided in Fig. 3 to illustrate that the ℓ_1-norm is more robust to outliers. Suppose 11 points are given, and a line model $y = kx+b$ is used to fit these points. The ℓ_1-norm model and ℓ_2-norm model are utilized to solve this problem. Fig. 3(a) shows that both the ℓ_1-norm and the ℓ_2-norm produce a similar estimation when there are no outliers in the input data. However, when there are outliers, the results are quite different. In Fig. 3(b), with the two outliers, the ℓ_1-norm still produces similar results to Fig. 3(a), while it can be seen that the ℓ_2-norm result is seriously affected by the outliers. This demonstrates that the ℓ_1-norm is more robust to outliers, and this toy example makes us believe that, with sparse local-pixel structures, sparse representation can provide a better performance in face hallucination. Furthermore, we believe that using sparse

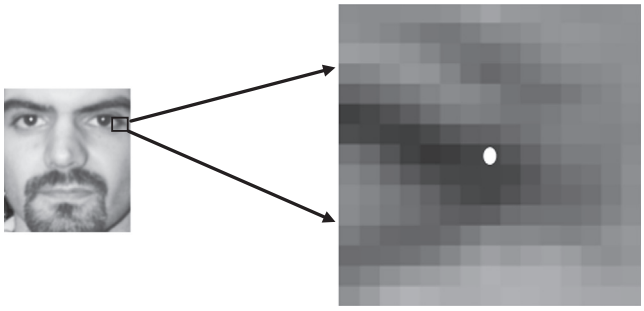

Fig. 2. A central pixel, marked in white, and its neighbors in a 15×15 pixel window.

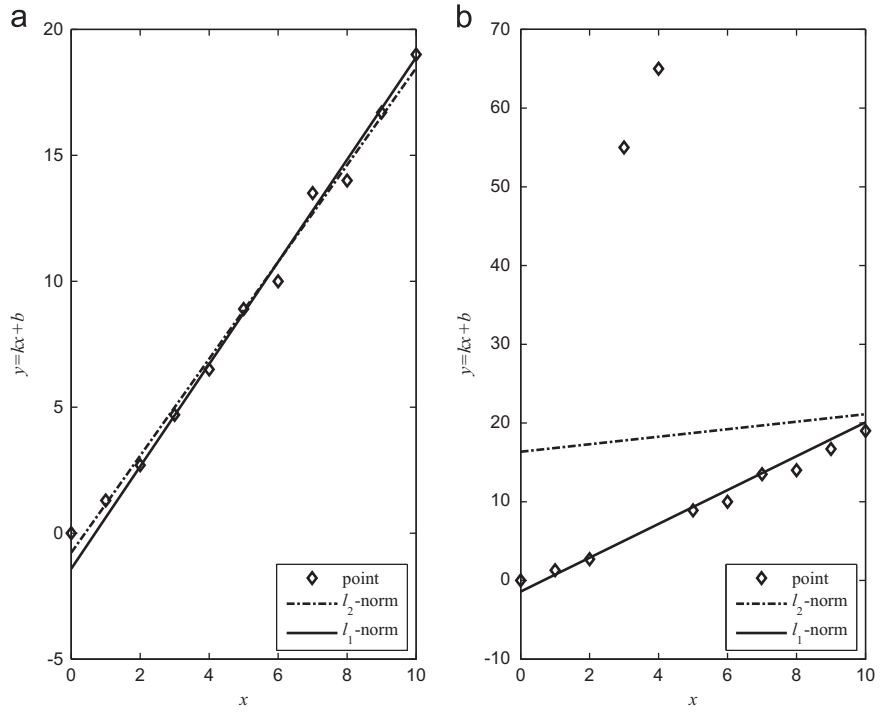

Fig. 3. Fitting a line to 11 given points using the ℓ_1-norm and the ℓ_2-norm: (a) without outliers and (b) with two outliers.

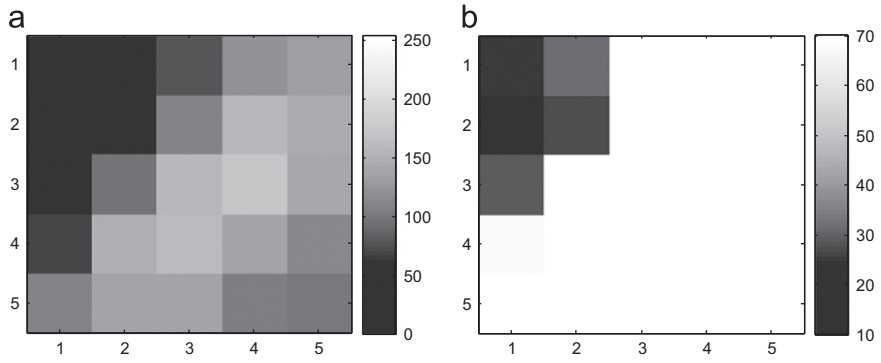

Fig. 4. A central pixel with its 5×5 neighboring-pixels window in a facial image, and the corresponding counterpart whose pixel value is between the range of [10,70].

representation to learn and characterize the prior model can improve the performance of our proposed method.

We have one more example to show that sparse representation can better represent sparse local-pixel structures. Fig. 4(a) shows a randomly selected pixel and its neighbors ($p=5$) in the ground-truth HR image of a LR face. It is obvious that the 5 pixels at the upper-left corner have the biggest difference to the central pixel. By simply changing the range of the pixel values to [10,70], we can observe that most of the pixels have a similar value, except the five pixels at the top-left corner (Fig. 4(b)). Thus, we can consider these pixels as belonging to another class, and we call them outlier pixels. Table 1(a) and (b) shows the learned local structures using the Gaussian function method in [11] and sparse representation, respectively (if a value is less than 0.001, it is set at zero). It can be seen that only two pixels in the top-left corner in Table 1(a) have their weights equal to zero, while four pixels in Table 1(b) have their weights equal to zero. In order to compare Table 1(a) and (b) at the same level, we define the following measure to calculate the percentage of outlier pixels' weights:

$$W_{out} = \sum \alpha_{u,v}^{out}(x,y) / \sum \alpha_{u,v}(x,y) \times 100\% \quad (8)$$

where $\alpha_{u,v}^{out}(x,y)$ is the weight of an outlier pixel. By applying (8) to Table 1, the results are $W_{out}^{a} = 6.4\%$ and $W_{out}^{b} = 4.5\%$. Thus, with the use of sparse representation, the percentage of outlier pixels' weights decreases. This means that the outlier pixels will contribute less in the iterative reconstruction process, which can help to produce a better SR result.

Table 1
The learned coefficient matrices corresponding to the pixel patch in Fig. 4(a) ($\times 10^{-4}$).

a					b				
0	**0**	0	1004	4647	**0**	**0**	0	10	53
36	**369**	2125	4812	3516	**0**	**0**	119	1391	306
2064	6093	Center	2994	908	**45**	5404	Center	157	40
2224	3518	1874	622	540	642	86	48	21	39
29	139	179	274	631	11	12	13	13	1577

4. Detailed procedure of the framework

Fig. 5 illustrates the three major steps in our proposed framework. In Step I, the input LR face is used to search a face dataset and identify the k pairs of LR–HR example faces having the most

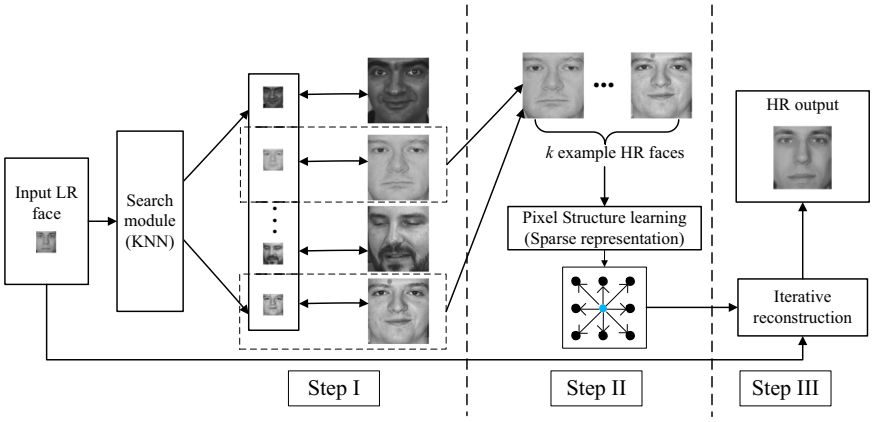

Fig. 5. The implementation procedure for our proposed face-hallucination framework.

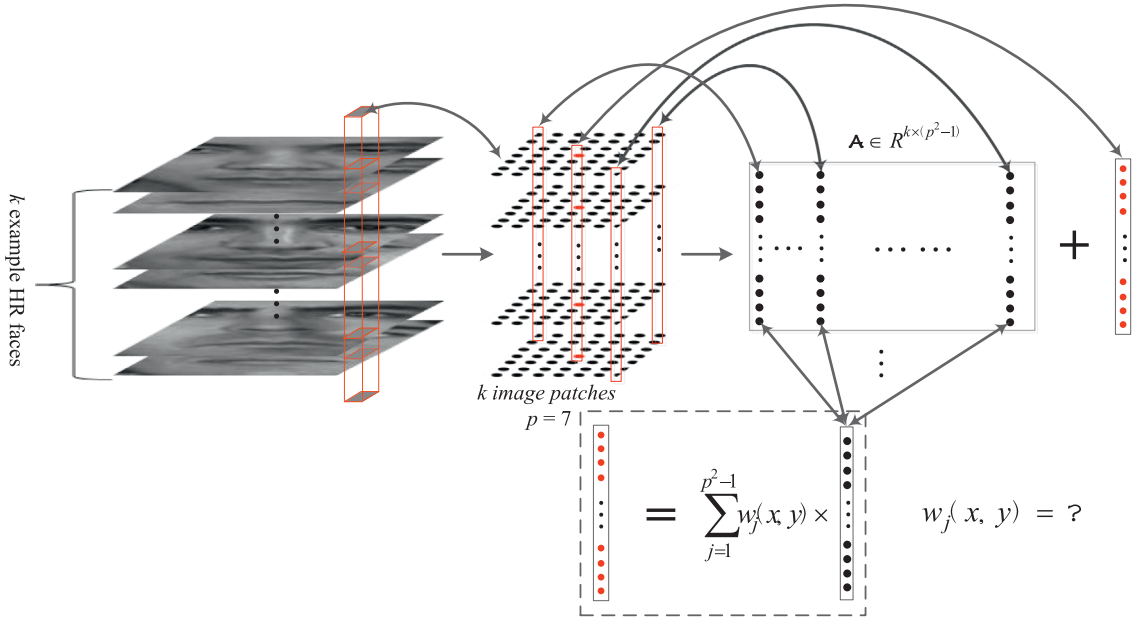

Fig. 6. Illustration of learning the sparse local-pixel structure of a face patch from example HR faces.

similar local-pixel structures to the input LR face, using Principal Component Analysis (PCA) and KNN. The k pairs of example faces are composed of k LR faces and their corresponding k-HR faces. Then, the k-HR example faces are employed and are warped to the input LR face using optical flow, so as to make the estimation of the target HR face more accurate. This process will produce k-HR warped example faces. Warped errors will be used in the next step. In Step II, the local-pixel structures, which are represented by the weights of the neighboring pixels, are learned from the k-HR warped example faces using sparse representation. Then, the accuracy of the weights is improved by using the warped errors produced in Step I. Finally, in Step III, the weights of the neighbors are employed to estimate the target HR face using an iterative method. Following are the details of these three steps.

4.1. Step I: searching and warping

In our face-hallucination framework, finding the example faces of the input LR face in a face dataset composed of GT [38] and FERET [39] is the first step. Here, the example faces used for reconstructing the HR face are composed of k LR faces and their corresponding HR counterparts. In the k pairs of example faces, the k LR example faces have the most remarkably similar pixel structures to the input LR face. In our experiments, the dataset contains 1552 LR–HR face pairs; they have all been aligned using the method in [35], and normalized using the illumination–normalization technique employed in [11]. The LR faces are also magnified to the size of the HR faces using bicubic interpolation. Principal Component Analysis (PCA) is used to represent the interpolated LR face images, and the KNN method is adopted to search for the k LR example faces that are the most similar to the LR input face. Then, the selected k corresponding HR faces are used as example faces. A warping operation is employed to make the estimation more accurate. We use optical flow, which has been used in SR in [11,34,36], to warp the example faces. In our paper, firstly, the flow field between the input LR face and each of the k LR example faces is derived, and then the corresponding HR example faces are warped accordingly based on the k flow fields. The purpose of the warping operation is to force the k-HR example faces to have more similar local-pixel structures to the input LR face. By using the warped errors, we can determine the importance of the neighboring pixels in estimating the HR pixels.

4.2. Step II: learning the neighboring weights via sparse representation

In Fig. 2, the sparse local-pixel structure of a face patch has been described. If we can use all the neighboring pixels of a central pixel to reconstruct the central pixel, then it can be represented effectively as a linear combination of its neighboring pixels using sparse representation. Fig. 6 illustrates the learning of the sparse local-pixel structure of a face patch from the example HR faces in our method. K image patches with the size of $p \times p$ from the warped example HR faces provide the data to construct the overcomplete dictionary matrix $\mathbf{A}$ and the central pixel vector (marked in red in the dashed box in Fig. 6). Combining the model shown in the dashed box, we propose using the following formulation to model the learned sparse local-pixel structure:

$$\psi(x,y) = \sum_{j=1}^{p^2-1} \omega_j(x,y) \times \mathbf{A}(:,j), \tag{9}$$

where $\psi(x,y) = [I^1_{ex}(x,y)\cdots I^k_{ex}(x,y)]^T$ and k is the number of HR example faces, and $\omega_j(x,y)$ denotes the weight between the pixel $I(x, y)$ and its neighboring pixel $I(x+\mu, y+\nu)$ with a relative displacement of (μ, ν). The variables μ and ν cannot be zero at the same time. The ith row of the matrix $\mathbf{A}$ in (6) and (9) are the neighboring pixels of the central pixel $\psi(x, y)$ of the ith example face, and the size of the neighborhood is $p \times p$ ($p=7$ in Fig. 6). The definition and illustration of $\mathbf{A}$ are given in (10) and Fig. 6, respectively:

$$\mathbf{A} = [I^1_{\mu,\nu}(x+\mu, y+\nu)\cdots I^k_{\mu,\nu}(x+\mu, y+\nu)] \in R^{k\times(p^2-1)}. \tag{10}$$

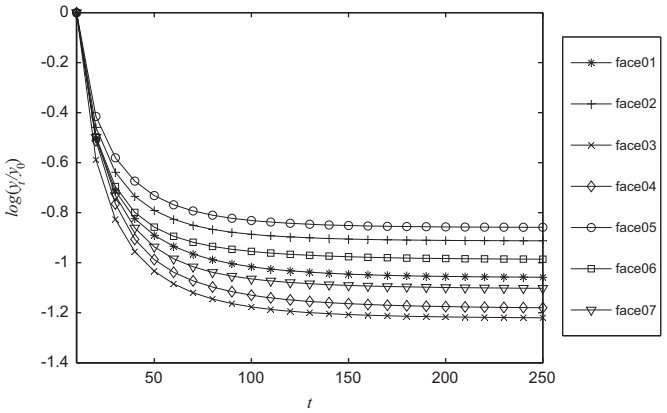

Fig. 7. The variations of y for different numbers of iterations with seven different face images.

Here, we assume that $\omega = [0\cdots\omega_{\mu,\nu}(x,y)\cdots 0]^T$ is the coefficient vector used in (6), whose entries are all zeros, except those associated with the central pixel vector ψ. We use L1LS [30] to solve (6) using the matrix $\mathbf{A}$ and the central pixel vector ψ.

After computing the coefficient vector ω using L1LS, a refinement procedure is performed according to the warped errors produced in the first step. ω can be rewritten as $\omega' = c_{x,y} \times \omega$, and $c_{x,y}$ represents the refinement procedure. Thus, (9) can be rewritten as follows:

$$\psi' = c_{x,y} \times A\omega = c_{x,y} \times \psi, \tag{11}$$

where ω has been calculated using L1LS. The definition of $c_{x,y}$ is the same as in [11]:

$$c_{x,y} = \arg\min \|\psi' - c_{x,y}\psi\|^2_{\mathbf{B}_{x,y}}, \tag{12}$$

where $\|\cdot\|^2_{\mathbf{B}_{x,y}}$ is the operator of the weighted ℓ_2-norm, and $\mathbf{B}_{x,y}$ is a diagonal weight matrix with the form $\mathbf{B}_{x,y} = diag(b_1(x,y), \ldots, b_k(x,y))$, where $b_k(x, y)$ is calculated using the weight assigned to the kth example face at location (x, y). The warping operation from example faces to the target input face may not be sufficiently accurate, so the weights of the warped HR examples should be dependent on the corresponding warping errors. Here the weights are calculated as follows:

$$b_k(x,y) = (\sum Er_k(x+\rho, y+\sigma)+\varepsilon)^{-\beta} / \sum_k (\sum Er_k(x+\rho, y+\sigma)+\varepsilon)^{-\beta}, \tag{13}$$

where Er_k represents the warped errors between the input interpolated LR face and the kth-HR warped face at location (x, y), and $(\rho, \sigma) \in \Omega$, where Ω is a patch centered at (x, y). In our experiment, the size of Ω is 9×9 , and β is a controlling parameter which can balance the effect of (13). ε is a small positive value used to prevent the denominator from being zero. Thus, the contribution of each example face in computing the weight is dependent on its warping errors. For those examples with large warping errors at location (x, y), the corresponding $b_k(x, y)$ will be reduced. Then, $c_{x,y}$ can be calculated as follows:

$$c_{x,y} = \psi'^T \mathbf{B}_{x,y}\psi / (\psi^T \mathbf{B}_{x,y}\psi + \varepsilon), \tag{14}$$

where ε has the same effect as it does in (13). Then, the final neighboring weights for the target HR face at location (x, y) can be obtained.

4.3. Step III: reconstructing the target HR face

By now, the local structures have been learned. The major task of this step is to apply the sparse local structures from the HR example faces to the input interpolated LR face. An iterative method [37] is employed in this step. The pixels of the input LR image are used as the anchor points in the iterations. The target

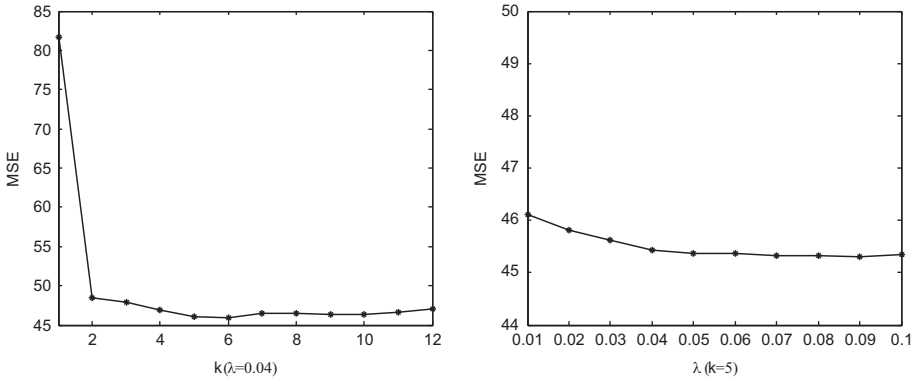

Fig. 8. The average MSE results using different k and λ with a magnification factor (mag) of 4.

HR pixel values are confined within the range of [0, 255], i.e. whenever the pixel value is smaller than 0 or higher than 255, it will be set at 0 or 255, respectively. The target HR face pixels are reconstructed as follows:

$$\Delta_t(x,y) = \hat{\mathbf{I}}_{\mathbf{h}}^{\mathbf{t}}(x,y) - \sum_{\mu,\nu \in C} w_{\mu,\nu}(x,y)\hat{\mathbf{I}}_{\mathbf{h}}^{\mathbf{t}}(x+\mu, y+\nu), \quad (15)$$

$$\hat{\mathbf{I}}_{\mathbf{h}}^{\mathbf{t+1}}(x,y) = \hat{\mathbf{I}}_{\mathbf{h}}^{\mathbf{t}}(x,y) - g\Delta_t(x,y), \quad (16)$$

where $\Delta_t(x,y)$ is the regularization parameter between the iterations, and g is a scale factor. In each iteration, we have $\mathbf{I}_{\mathbf{h}}(x,y) = \mathbf{I}_{\mathbf{l}}(x/r, y/r)$, where r is the down-sampling factor in both dimensions in (2). The input interpolated LR face is selected as the initial estimate $\hat{\mathbf{I}}_{\mathbf{h}}^{\mathbf{0}}$. Here, we propose a method to observe the relationship between the number of iterations and the variable $\Delta_t(x,y)$. We define $y_t' = \sum |\Delta_t(x,y)|$, where $0 \le x < m$ and $0 \le y < n$ (i.e. $m \times n$ is the size of the target HR face). Then, we set $y = \log(y_t'/y_0)$ to make the curves more intuitive. Fig. 7 shows the variations of y for 7 randomly selected face samples with the scale factor $g=0.05$. We can see that, after 150 iterations, the changes become stable. Thus, in the following experiments, we set the number of iterations to be performed at 150.

5. Experimental results and discussions

5.1. Experimental settings

In our experiments, the experimental subset contains 1552 images which are selected from the GT database [38] and the FERET databases [39]. With a down-sampling factor of 4 for each dimension, the resolutions of the HR faces and the corresponding LR faces are 124×108 and 31×27, respectively. Then, the parameters k in (10), i.e. the number of example faces, and λ in (6) are determined empirically. Fig. 8 shows the performances of our proposed framework with different values of k and λ. When $k=5$ and $\lambda=0.04$, the average mean squared error (MSE) of 100 samples in our algorithm becomes stable. The average SSIM (Structural Similarity Index) [40] and PSNR with different neighborhood sizes are tabulated in Table 2. When $p=3$, both SSIM and PSNR are at their largest. Therefore, in the following experiments, we set $k=5$, $\lambda=0.04$, and $p=3$.

Table 2
The average SSIM and PSNR with different neighborhood sizes (p).

p	**3**	5	7	9	11	13	15
SSIM	**0.8639**	0.8585	0.8538	0.8504	0.8480	0.8455	0.8430
PSNR	**28.9054**	28.7435	28.5636	28.4443	28.3593	28.2840	28.2163

5.2. Face reconstruction

To measure the performance of our proposed face-hallucination method, we first reconstruct HR face images with a magnification factor (*mag*) of 4. The resolution of the HR faces and the corresponding LR faces are 124×108 and 31×27, respectively. A testing dataset is formed by choosing 150 HR–LR faces from the 1552 selected images, and the performance is evaluated using the "leave-one-out" method. Our proposed algorithm is compared with Hu's method [11]

Table 3
The average PSNR and SSIM of the testing subset with *mag*=4.

Different methods	PSNR (dB)	SSIM
Bicubic interpolation	25.589	0.7659
Hu's method	27.850	0.8448
Chang's method	27.317	0.8203
Our proposed method	**28.901**	**0.8638**

Table 4
The average PSNR and SSIM of the testing subset with *mag*=6.

Different methods	PSNR (dB)	SSIM
Bicubic interpolation	23.372	0.7278
Hu's method	26.790	0.8086
Chang's method	23.342	0.7176
Our proposed method	**27.734**	**0.8156**

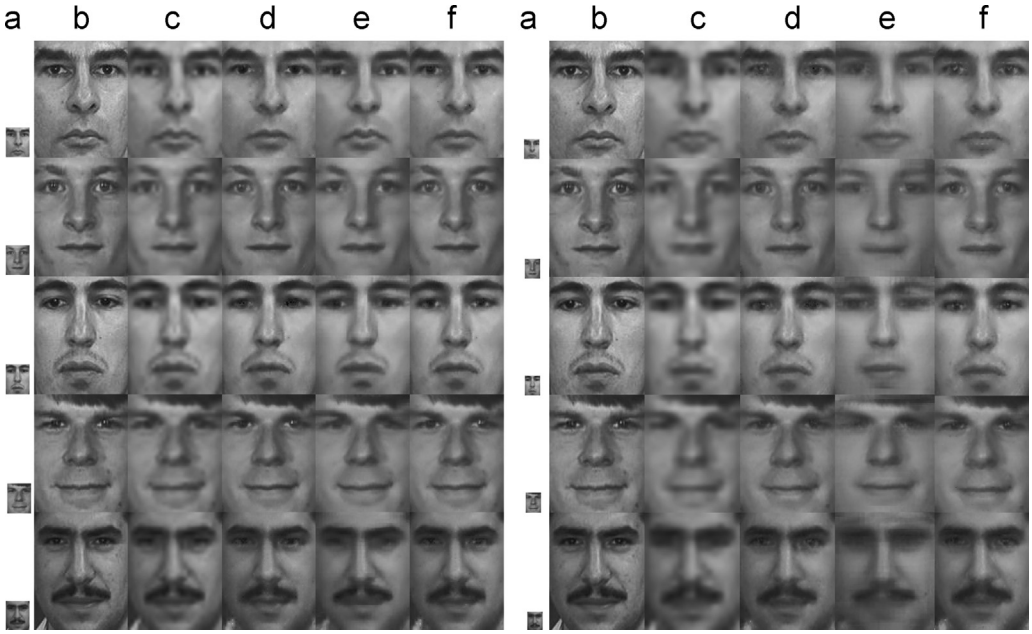

Fig. 9. HR faces reconstructed using different methods with *mag*=4(left) and *mag*=6(right): (a) the input LR faces, (b) the original HR faces, (c) bicubic interpolation, (d) Hu's method, (e) Chang's method, and (f) our proposed framework.

and Chang's method [6], as well as with bicubic interpolation. The reconstructed HR faces of some testing samples with two different magnification factors using the different methods are shown in Fig. 9. It is obvious that the results using bicubic interpolation are the blurriest, while the other methods can provide results of much better visual quality, especially Hu's method and our proposed method. To be specific, Hu's method and our proposed method achieve a better performance on the eyes, mouth and eyebrows. Our method can achieve even better results in edge regions than Chang's and Hu's methods can. This is mainly due to the fact that the ℓ_1-norm is more robust than the ℓ_2-norm. Next, we measure the performances of the different methods in terms of PSNR and SSIM with the testing set, and the average results are tabulated in Table 3. Our method outperforms all the other methods in terms of PSNR and SSIM. As mentioned above, Hu's method is superior to Chang's method, and bicubic interpolation produces the worst results. The results shown in Fig. 9 and Table 3 demonstrate that our method can achieve the best performance in terms of both visual quality and reconstruction error.

We also measure the performance of the different methods when the magnification factor is 6. The sizes of the HR and LR images in the testing set are reshaped to 126×108 and 21×18, respectively. The right-hand side of Fig. 9 shows the reconstructed HR faces using the different methods when the magnification factor is 4. The corresponding average PSNR and SSIM of the testing images are tabulated in Table 4. We can see that the performance of Chang's method degrades significantly here, and is even worse than bicubic interpolation. However, our proposed method can still retain a steady performance in terms of both visual quality and reconstruction error. These experiments demonstrate that our proposed framework is robust to larger magnification factors; this is due to the fact that, even with larger magnification factors, the sparse local-pixel structures learned from the k-HR example faces remain unchanged, and can still be captured accurately. The gradual decrease in the performance is mainly due to the reduced amount of initial information on the aligned input LR face when the magnification factor increases.

Because of the use of sparse representation in our method, we also compare our proposed method with the sparse representation-based image-interpolation method, NARM-SRM-NL, proposed in [44]. All the experiments in [44] were conducted with magnification factors of 2 and 3, so we compare NARM-SRM-NL to our method with the magnification factors being 2, 3, and 4. Another testing set was constructed with the same number of images as the previous one. Five of the reconstructed HR faces selected randomly from the testing set and with the different magnification factors are shown in Fig. 10, and the corresponding average PSNR and SSIM with the different magnification factors are tabulated in Table 5.

In the provided source code of NARM-SRM-NL, a number of parameters are to be determined; only the parameter values for the magnification factors 2 and 3 are given. In the experiment, we set the parameters for the magnification factor of 4 the same as those for the magnification factor of 3. As shown in Fig. 10 and Table 5, when the magnification factor is 2, NARM-SRM-NR outperforms our method, but when the magnification factor is 3 or 4, our method outperforms it significantly in terms of both image visual quality and reconstruction error. NARM-SRM-NL reconstructs an image patch using non-local neighboring patches, and the NARM matrix is used to further improve the incoherence between the sampling matrix and the adaptive local PCA dictionary. The experimental results show that NARM-SRM-NL is more appropriate for reconstructing images which are very smooth or which contain fewer local features. When an input LR image drops plenty of local structures of its HR counterpart, NARM-SRM-NL will produce a weak result. For face hallucination, the facial images are highly structured. When the magnification factor is small, compared with HR faces, the input LR faces still include most of the local structures. Thus, the reconstructed results of NARM-SRM-NL are excellent, as shown in the second row of Fig. 10. However, when the magnification factor increases, more local information will be lost in the LR faces. Therefore, NARM-SRM-NL cannot reconstruct the fine individual facial details or the photorealistic local features, and consequently, the reconstructed faces become much smoother, as shown in the third and the last rows of Fig. 10, and exhibit larger reconstruction errors. Nevertheless, our method still produces plausible faces with smaller reconstruction errors due to the use of local-pixel structures.

Table 5
The average PSNR and SSIM with different magnification factors.

mag	Methods	PSNR (dB)	SSIM
2	NARM-SRM-NL	**36.948**	**0.9568**
	Our method	35.247	0.9435
3	NARM-SRM-NL	29.896	0.8532
	Our method	**32.921**	**0.9074**
4	NARM-SRM-NL	26.180	0.7517
	Our method	**30.889**	**0.8748**

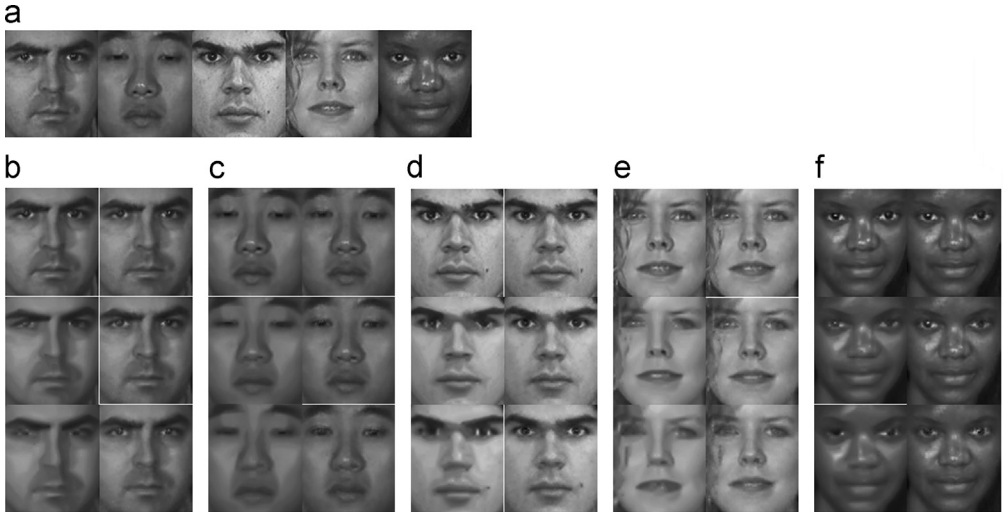

Fig. 10. Five HR faces reconstructed using our method and NARM-SRM-NL with *mag*=2, 3, and 4: (a) the ground-true HR face images; (b)–(f) the reconstructed HR faces of the five faces – the first, second and third rows show the reconstructed HR faces with the magnification factor = 2, 3, and 4, respectively. Those in the left and right columns are the results based on NARM-SRM-NL and our proposed method, respectively.

The proposed algorithm in this paper was simulated on a computer of 2.7 GHz CPU with 8GByte SDRAM, and was implemented using MATLAB. Our code is available online: http://as.nwsuaf.edu.cn/fhsr.html.

6. Conclusions

In this paper, we have proposed a method for face hallucination based on learning the sparse local-pixel structures of the target HR facial images. The sparse representation is used to capture the local structures from the HR example faces, and optical flow is applied to make the learning process more accurate. The experimental results have demonstrated that our proposed framework is competitive and can achieve superior performance compared to other state-of-the-art face-hallucination methods. The superior performance of our algorithm is mainly due to the fact that the example faces can provide both the holistic and the pixel-wise information for reconstructing the target HR facial images, and it can estimate the sparse local-pixel structures of the target HR faces more accurately from the example faces using sparse representation. Our proposed method can maintain the impressive capability of inferring fine facial details and generating plausible HR facial images when the input face images are of very low resolution.

Conflict of interest

None declared.

Acknowledgments

Project 61202188, 61303125 supported by National Natural Science Foundation of China. Grant EP/J020257/1 supported by the UK EPSRC, research project by the International Doctoral Innovation Centre (IDIC), the University of Nottingham Ningbo China (UNNC), and Project 2012B10055 by Ningbo Science & Technology Bureau Science and Technology.

References

[1] R. Tsai, T.S. Huang, Multiframe image restoration and registration, Advances in Computer Vision and Image Processing 1 (1984) 317–339.
[2] R.R Schultz, R.L Stevenson, Extracion of high resolution frames from video sequence enhancement, IEEE Transactions on Image Processing 5 (1998) 996–1011.
[3] M. Elad, A. Feuer, Restoration of a single super-resolution image from several blurred, noisy and undersampled measured images, IEEE Transactions on Image Processing 6 (1997) 1646–1658.
[4] S. Borman, R.L. Stevenson, Super-resolution from image sequences—a review, in: Proceedings of the 1998 Midwest Symposium on Circuits and Systems, 1998, pp. 374–378.
[5] S. Farsiu, M.D. Robinson, M. Elad, P. Milanfar, Fast and robust multiframe super resolution, IEEE Transactions on Image Processing 13 (2004) 1327–1344.
[6] H. Chang, D.-Y. Yeung, Y. Xiong, Super-resolution through neighbor embedding, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, pp. 275–282.
[7] D. Glasner, S. Bagon, M. Irani, Super-resolution from a single image, in: IEEE International Conference on Computer Vision, 2009, pp. 349–356.
[8] J.-S. Park, S.-W. Lee, An example-based face hallucination method for single-frame, low-resolution facial images, IEEE Transactions on Image Processing 17 (2008) 1806–1816.
[9] K.I. Kim, Y. Kwon, Example-based learning for single-image super-resolution, *computer* science, Pattern Recognition 5096 (2008) 456–465.
[10] M. Gong, K. He, J. Zhou, J. Zhang, Single color image super-resolution through neighbor embedding, Journal of Computational Information Systems 7 (2011) 49–56.
[11] Y. Hu, K.M. Lam, G. Qiu, T. Shen, From local pixel structure to global image super-resolution: a new face hallucination framework, IEEE Transactions on Image Processing 20 (2011) 433–445.
[12] G. Qiu, A progressively predictive image pyramid for efficient lossless coding, IEEE Transactions on Image Processing 8 (1999) 109–115.
[13] G. Qiu, Inter-resolution look-up table for improved spatial magnification of image, Journal of Visual Communications and Image Representation 11 (2000) 360–373.
[14] W.T. Freeman, T.R. Jones, E.C. Pasztor, Example-based super-resolution, IEEE Computer Graphics and Applications 22 (2002) 56–65.
[15] X. Li, K.M. Lam, G. Qiu, L. Shen, S. Wang, Example-based image super-resolution with class-specific predictors, Journal of Visual Communications and Image Representation 20 (2009) 312–322.
[16] C. Liu, H.Y. Shum, C.S. Zhang, A Two-Step Approach to hallucinating faces: global parametric model and local nonparametric model, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001, pp. 192–198.
[17] S. Baker, T. Kanade, Limits on super-resolution and how to break them, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 1167–1183.
[18] S. Baker, T. Kanade, Hallucinating faces, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 83–88.
[19] C. Liu, H.-Y. Shun, W.T. Freeman, Face hallucination: theory and practice, International Journal of Computer Vision 75 (2007) 115–134.
[20] Y. Liang, J. Lai, W. Zheng, Z. Cai, A survey of face hallucination, in: Proceedings, CCBR'12 of the 7th Chinese Conference on Biometric Recognition, 2012, 7701, pp. 83–93.
[21] K. Huang, S. Aviyente, Sparse Representation for Signal Classification, Presented at the Neural Information Processing Systems, 2006.
[22] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 210–227.
[23] G.S.V.S. Sivaram, S. Ganapathy, H. Hermansky, Sparse auto-associative neural networks: theory and application to speech recognition, in: Proceedings of the INTERSPEECH'2010, September 2010, pp. 2270–2273.
[24] T.N. Sainath, S. Maskey, D. Kanevsky, B. Ramabhadran, D. Nahamoo, J. Hirschberg, Sparse representations for text categorization, in: Proceedings of the INTERSPEECH'2010, September 2010, pp. 2266–2269.
[25] E. Amaldi, V. Kann, On the approximability of minimization nonzero variables or unsatisfied relations in linear systems, Theoretical Computer Science 209 (1998) 237–260.
[26] D.L. Donolo, For most large underdetermined systems of linear equations, the minimal l^1-norm solution is also the sparsest solution, Communication on Pure and Applied Mathematics 59 (2006) 797–829.
[27] E. Candes, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Communications on Pure and Applied Mathematics 59 (2006) 1207–1223.
[28] D.L. Donoho, Y. Tsaig, I. Drori, J.-L. Starck, Sparse Solution of Underdetermined Linear Equations by Stagewise Orthogonal Matching Pursuit, Preprint, December 2007.
[29] S.J. Wright, R.D. Nowak, M.a.A.T. Figueiredo, Sparse reconstruction by separable approximation, in: Proceedings of the ICASSP'2008, May 2008, pp. 3373–3376.
[30] X. Mei, H. Ling, D.W. Jacobs, Sparse representation of cast shadows via L1-regularized least squares, in: Proceedings of the International Conference on Computer Vision (ICCV), 2009, pp. 583–590.
[31] Q. Ke, T. Kanade, Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 739–746.
[32] A Guitton, D.J. Verschuur, Adaptive subtraction of multiples using the L1-norm, European Association of Geosientists and Engineers (2004) 27–38.
[33] N. Kwak, Principal component analysis based on L1-norm maximization, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 1672–1680.
[34] D. Mitzel, T. Pock, T. Schoenemann, D. Cremers, Video super resolution using duality based TV-L1 optical flow, Computer Science Pattern Recognition 5748 (2009) 432–441.
[35] K.-W. Wong, K.-M. Lam, W.-C. Siu, An efficient algorithm for human face detection and facial feature extraction under different conditions, Pattern Recognition 34 (2001) 1993–2004.
[36] W. Zhao, H.S. Sawhney, Is super-resolution with optical flow feasible, in: Seventh European Conference on Computer Vision, 2002, pp. 599–613.
[37] P. Wesseling, An Introduction to Multigrid Methods, Wiley, Chichester, England, 1992.
[38] Georgia Tech Face Database, ⟨http://www.anefian.com/research/face_reco.htm⟩.
[39] P.J. Phillips, H. Wechsler, J. Huang, P.J. Rauss, The FERET database and evaluation procedure for face-recognition algorithms, Image and Vision Computing 16 (1998) 295–306.
[40] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612.
[41] J. Yang, J. Wright, H. Tang, Y. Ma, Image super-resolution as sparse representation of raw image patches, IEEE Computer Vision and Pattern Recognition (CVPR) (2006) 1–8.
[42] J. Yang, H. Tang, Y. Ma, T. Huang, Face hallucination via sparse coding, in: International Conference on Image Processing (ICIP), 2008, pp. 1264–1267.
[43] J. Yang, H. Tang, Y. Ma, T. Huang, Image super-resolution via sparse representation, IEEE Transactions on Image Processing 19 (2010) 2861–2873.
[44] W. Dong, L. Zhang, R. Lukac, G. Shi, Sparse representaion based image interpolation with nonlocal autoregressive modeling, IEEE Transactions on Image Processing 22 (2013) 1382–1394.

Yongchao Li received the B.Eng. degree in information management and information system from the college of Information Engineering of Northwest A&F University, China, in 2011. Currently, he is working toward the M.Eng. degree at the same university. His research interests include image processing and machine learning.

Cheng Cai received the B.S. degree in information engineering from the Xi'an Jiaotong University, China, the M.E. and Ph.D. degrees in information and communication engineering from Xi'an Jiaotong University, China, in 2001, 2003 and 2008, respectively. From March 2004 to December 2005, he was a Research Assistant in the Multimedia Center at the Hong Kong Polytechnic University, Hong Kong. From June 2008 to July 2008, he was a Visiting Scientist in the Department of Computer Science at the Oldenburg University, Germany. From July 2009 to September 2009, he was a Research Associate in the Multimedia Center at the Hong Kong Polytechnic University, Hong Kong. From July 2010 to August 2010, he was a Senior Research Associate in the Department of Electronic Engineering, City University of Hong Kong, Hong Kong. From December 2012 to March 2013, he was a Postdoctoral Fellow in the Institute of Computer Graphics and Algorithms, Vienna University of Technology, Vienna, Austria. Currently, he was an Associate Professor at the Department of Computer Science, College of Information Engineering, Northwest A&F University, China. His research interests include pattern recognition, signal processing and bioinformatics. He has authored over 30 papers in journals and conferences, and has served as a reviewer for many journals and conferences. He organized a special session on APSIPA ASC 2010 Singapore.

Guoping Qiu received the B.S. degree in electronic measurement and instrumentation from the University of Electronic Science and Technology of China, Chengdu, China, in 1984, and the Ph.D. degree in electrical and electronic engineering from the University of Central Lancashire, Preston, UK, in 1993. He is currently a Reader in the School of Computer Science, University of Nottingham, Nottingham, UK. He has research interests in the broad area of computational visual information processing and has published widely in this area.

Kin-Man Lam received the Associateship in Electronic Engineering with distinction from The Hong Kong Polytechnic University (formerly called Hong Kong Polytechnic) in 1986, the M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College of Science, Technology and Medicine, London, UK, in 1987 (under the S.L. Poa Scholarship for overseas studies), and the Ph.D. degree from the Department of Electrical Engineering, University of Sydney, Sydney, Australia, in August 1996. From 1990 to 1993, he was a Lecturer at the Department of Electronic Engineering, The Hong Kong Polytechnic University. He joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, as an Assistant Professor in October 1996, became an Associate Professor in 1999, and is now a Professor. He has been a member of the organizing committee and program committee of many international conferences. His current research interests include human face recognition, image and video processing, and computer vision. Professor Lam is a Secretary of the 2010 International Conference on Image Processing (ICIP 2010), a Technical Co-Chair of 2010 Pacific-Rim Conference on Multimedia (PCM 2010), and a BoG member of the Asia-Pacific Signal and Information Processing Association (APSIPA). He also serves as an Associate Editor of the IEEE Transactions on Image Processing and International Journal on Image and Video Processing.