Multimed Tools Appl (2016) 75:17273–17301
DOI 10.1007/s11042-015-2996-4

Heteroskedasticity tuned mixed-norm sparse regularization for face hallucination

Zhongyuan Wang[1] · **Ruimin Hu**[1] · **Junjun Jiang**[2] · **Zhen Han**[1] · **Zhenfeng Shao**[3]

Received: 11 June 2015 / Accepted: 8 October 2015 / Published online: 22 October 2015

Abstract Face hallucination is typically an ill-posed inverse problem, so it is essential to exploit an effective norm-regularized underlying representation. Due to the under-sparsity or over-sparsity, the widely used regularization methods, such as ridge regression and sparse representation, lead to poor robustness in the presence of noise. In addition, standard forms of penalty functions fail to account for the nature of heteroskedasticity of reconstruction coefficients, thus hardly providing optimal solutions in terms of accuracy and stability. To this end, this paper derives the locally weighted variants of standard regularization representation from Bayesian inference perspective, which impose the similarity constraint within the observed image and training images onto the penalty function. Further, considering the reduced sparseness of noisy images, a moderately sparse regularization method with a mixture of ℓ_1 and ℓ_2 norms is introduced to deal with noise robust face hallucination. New determination methods on weighting function and regularization parameter are particularly explored. Various experimental results on public face databases as well as real-world images validate the effectiveness of proposed method.

✉ Zhongyuan Wang
wzy_hope@163.com

Ruimin Hu
hrm1964@163.com

Junjun Jiang
junjun0595@163.com

Zhen Han
hanzhen_2003@hotmail.com

Zhenfeng Shao
shaozhenfeng@163.com

[1] National Engineering Research Center for Multimedia Software and the School of Computer, Wuhan University, Wuhan 430072, China

[2] School of Computer, China University of Geosciences, Wuhan 430074, China

[3] State key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

Keywords Face hallucination · Sparse regularization · Locally weighted penalty · Mixed norms

1 Introduction

Face super-resolution, or face hallucination, refers to the technique of estimating a high-resolution (HR) face image from low-resolution (LR) face image sequences or a single LR one. Due to constrained imaging conditions in many scenarios, it is hard to capture HR face images, and thus face hallucination is extensively used for pre- and/or post-processing in vision related applications, such as face recognition, video surveillance and human-computer interaction. A large number of theoretical and applicable works on face hallucination have been carried out. It is universally acknowledged that current face hallucination approaches fall into three categories: interpolation, reconstruction, and learning based methods. Among them, learning based methods have aroused great concerns provided that they can provide high magnification factors.

Learning based methods first learn *a priori* information from LR images and their corresponding HR images in the database, and then reconstruct the HR counterpart assisted by the obtained prior information. They can date back to the early work proposed by Freeman et al. [12], who employed a patch-wise Markov network to model the relationship between LR images and the HR counterparts. Thereafter, Baker and Kanade [1] developed a Bayesian approach to infer the missing high-frequency components from a parent structure training samples, and first coined the term "face hallucination". Motivated by their pioneering work, a class of dictionary learning methods has gained popularization recently, in which image patches can be well-approximated as a linear combination of elements from an appropriately designed over-complete dictionary.

All dictionary-learning-based methods are equipped with representation methods to solve optimal coefficients in terms of accuracy, stability and robustness. Least-squares (LS) is the most commonly used representation in face hallucination. Chang et al. [7] proposed a neighbor embedding (NE) based face hallucination method. It uses a fixed number of neighbors for reconstruction, thus usually resulting in blurring due to under- or over-fitting. To alleviate this problem, [2] used a nonnegative constraint on solution to enhance visual results. By incorporating the position priors of image patch, Ma et al. [23] introduced a position-patch based method to estimate a HR image patch using the same position patches of all training face images. To reduce residual errors, Park et al. [27] proposed an example-based method to recursively update a reconstructed HR image. In [28], the Huber norm was adopted for data fidelity term instead of ordinary squared error to improve robustness against outliers. However, face hallucination is an under determined inverse recovery problem, while LS hardly provides a stable and unique solution.

In contrast, regularization methods enable a global unique and stable solution by improving the conditioning of the ill-posed problem. One of the most popular regularization methods is ridge regression (RR). It can produce the solution of minimum energy subject to a squared ℓ_2 norm penalty. Liu et al. [22] proposed to integrate a global parametric principal component analysis model along with a local nonparametric Markov random field model for face hallucination, where RR is used to represent global face images. Motivated by their work, Li and Lin [20] also presented a two-step approach for hallucinating faces with global images reconstructed with a maximum *a posteriori* (MAP) criterion and residual images re-estimated with the MAP criterion. To enhance the subset selection, Tang et al. [29] proposed to learn a

local regression function over the local training set. The local training set is specified in terms of the distances between training samples and a given test sample. Instead of roughly limiting the spatial range of a training set in [29], Ref. [17] used a locality-constrained RR model and gained impressive results. However, the quadratic penalty function of RR implies that images are globally smooth and always leads to over-smooth or under-sharp edge images. A more realistic image model should consider the fact that images are made of smooth regions, separated by sharp edges.

Half-quadratic (HQ) minimization methods [8, 15, 26] are popular in numerous inverse problems such as deblurring, super-resolution. On the basis of empirical observations and theoretical verifications to a limited extent, [8] showed that non-quadratic penalty functions favor edge-preservation and prevent loss of image details. Non-quadratic (especially non-convex) regularization whereas needs computationally intensive nonlinear optimization. Thus, half-quadratic reformulation of non-quadratic regularization was developed, in two different ways: additive form and multiplicative form [15, 26]. Ref. [26] showed several commonly used HQ regularizations, such as the q-th power of ℓ_q norm (denoted by $l_q{}^q=\sum|\cdot|^q$), Huber function, and Welsh function. Specially, $\ell_q{}^q$ regularization, also called bridge regression (BR) introduced by Frank [11], minimizes the linear regression problem subject to a $\ell_q{}^q$ norm penalty, which includes ridge regression with $q=2$ and sparse representation with $q=1$ as special cases. Cetin [6] used a $\ell_q{}^q$ regularization method to enhance the resolution of point targets of SAR images.

Owing to the well-established theory, sparse representation (SR) becomes more appealing than RR or HQ method. Yang et al. [32] are the first to introduce ℓ_1 norm SR to face hallucination problem. Assume that natural images can be sparsely decomposed, they proposed a local patch method over coupled over-complete dictionaries to enhance the facial profile. In [34], a modified version of SR was shown to be more efficient and much faster. To address the biased estimate of LS [23], Jung et al. [18] applied convex optimization to position-patch based approach. In [16], the idea of two-step face hallucination [22] was extended to robust face hallucination for video surveillance, where SR was adopted to synthesize eigenfaces. Zhang et al. [35] presented a dual-dictionary learning method to recover more image details, with both main dictionary and residual dictionary learned by SR.

The SR based methods can capture salient properties of natural images, and yet ℓ_1 norm SR turns out to be over-sparse for face hallucination. This is primarily due to the fact that face hallucination is a regression problem (pursuing prediction accuracy) rather than a classification problem (seeking discriminability in sparse features). Fan and Li [9] studied a class of regularization methods and proved that ℓ_1 norm shrinkage produces biased estimates for the large coefficients and could be suboptimal in terms of estimation risk. Meinshausen et al. [24] again showed the conflict of optimal prediction and consistent variable selection in ℓ_1 norm. Comparing ℓ_1 norm with elastic net (EN) proposed in [36], Li et al. [21] expressly pointed out that the former is much more aggressive in terms of prediction exclusion. Especially, we observe that ℓ_1 norm results in considerable degradation of hallucination performance in the presence of noise, either manually added Gaussian noise or unknown noise caused by a variety of factors (such as environmental conditions, underexposure, and the quality of acquisition devices). In the light of the discussions, we argue that the underlying representation in face hallucination should maintain a reasonable balance between subset selection and regression estimation. In order to capture the salient facial features, it is sufficient and necessary to exert moderate sparse constraints, but sparsity should not be overemphasized.

With the popularization of sparse representation, its weighted variants are also successively developed, such as iterative reweighted ℓ_1 minimization [5] and weighted ℓ_1 minimization with prior support [13, 19]. The main idea is to exploit *a priori* knowledge about the support of coding coefficients so as to favor desirable properties of solution. In particular, Friedlander et al. [13] theoretically proved that if the partial support estimate is at least 50 % accurate, then weighted ℓ_1 minimization outperforms the standard one in accuracy, stability, and robustness. Intuitively, the geometric locality in terms of Euclidean distance can be treated as intrinsic prior within the observed LR image and training images. Alternatively, previous studies on the local learning paradigm [30, 33] revealed that locality of learning methods is very essential to recognition problems (e.g., face recognition) or regression problems (e.g., face hallucination). Therefore, if we enforce a geometric locality constraint on regression coefficients to induce a weighted sparse representation, the improved hallucinated results can be expected.

In this paper, we propose a locally weighted sparse regularization (LWSR) to boost face hallucination performance. It incorporates distance-inducing weights into penalty function to favor the engagement of near training bases. Particularly, we suggest a mixture of ℓ_1 and ℓ_2 norms to handle noise robust face hallucination in practical applications. The major contributions of this paper lie in three-fold:

1) We are the first to observe the heteroskedasticity of regression coefficients and thereby derive a locally weighted penalty under MAP inference framework to promote accuracy and stability of the solution.
2) We introduce a moderately sparse regularization method with a mixture of ℓ_1 and ℓ_2 norms (referred to as $\ell_{1,2}$ norm for short) to model the statistically less sparse nature of noisy images. It outperforms any counterparts in the presence of noise.
3) To perform face hallucination task with proposed representation method in an effective and efficient way, we devise new determination methods for distance and weighting functions as well as regularization parameters.

In our previous work [31], we introduced a weighted adaptive ℓ_q sparse regularization method to model the sparsity behavior of face hallucination. Experimentally, our newly proposed mixed $\ell_{1,2}$ norm representation turns out more pronounced in terms of accuracy and computational efficiency. More importantly, the local weighting mechanism is justified from the perspective of heteroskedasticity in this paper for the first time.

The remainder of this paper is organized as follows. Section 2 introduces locally weighted regularization. Section 3 particularly presents $\ell_{1,2}$ norm regularization for hallucinating noisy images. Section 4 addresses several practical issues in face hallucination. Various experimental results are shown in Section 5. In Section 6, we conclude the paper.

2 Locally weighted ℓ_1 norm regularization

In this section, we focus on the Bayesian inference of locally weighted ℓ_1 regularization. The representative reweighted methods in compressed sensing will be surveyed first, which are also used for anchors in the experiments in Section 5.2.

2.1 Survey of reweighted methods

To simplify presentation, some notations are specified firstly. N is the size of a test sample (an image or an image patch) and M is the number of basis samples in the training set (or dictionary). Vector $\mathbf{x}$ denotes N-dimensional test sample and $\mathbf{Y}\in R^{N\times M}$ represents a training set with i-th column being the sample $\mathbf{Y}_i$. $\mathbf{w}\in R^{M\times 1}$ stands for an unknown coefficient vector, whose entries w_i, $i=1,2,\ldots,M$ are associated with individual bases in the training set.

Various weighted methods have been devised to enforce the desirable properties of standard regularization, compact or robust, for instance. Consider a common form of the weighted ℓ_1 minimization problem:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \left|\left| \mathbf{x}-\mathbf{Y}\mathbf{w} \right|\right|_2^2 + \lambda \left|\left| \mathbf{\Gamma}\mathbf{w} \right|\right|_1 \right\}, \tag{1}$$

where $\mathbf{\Gamma}$ is a diagonal matrix with positive weights $\mathbf{\Gamma}_i$, $i=1,2,\ldots,M$ on the diagonal and zeros elsewhere. Just like its non-weighted counterpart, this convex problem can be recast as a linear program, and so the optimization can be implemented readily using existing algorithms.

Different expressions on weights will lead to weighted penalties with quite distinct functionalities. In [5], Candes et al. proposed an iterative reweighted ℓ_1 minimization (RLM) method to find maximally sparse representation from over-complete dictionaries, in which weights $\mathbf{\Gamma}_i$, $i=1,2,\ldots,M$ are computed from the current solution:

$$\mathbf{\Gamma}_i^{(l+1)} = \frac{1}{\left| w_i^{(l)} \right| + \varepsilon}, \tag{2}$$

which is used for the next iteration such that more focal estimates can be produced as optimization progresses. Parameter ε is introduced to ensure that a zero-valued component does not strictly prohibit a nonzero estimate at the next step. Empirically, ε should be set slightly smaller than the expected nonzero magnitude.

In Refs. [13, 19], similar weighted ℓ_1 minimization methods are used for signal reconstruction from compressed sensing measurements when prior support information is available. The main idea is to choose weights such that the entries expected to be large are penalized less in the weighted objective function. Khajehnejad et al. [19] employed such a weighted ℓ_1 minimization with prior information (WPI) to recover the unknown signal where two different weights are assigned to the elements in the two sets with a respective different probability of being nonzero, namely,

$$\mathbf{\Gamma}_i = \begin{cases} C_1 & \text{if } i\in K_1 \\ C_2 & \text{if } i\in K_2 \end{cases}. \tag{3}$$

In practice, because the support of the sparse signal is usually unavailable, a method was suggested to obtain the approximate support set with two steps. A standard ℓ_1 minimization is first performed, and then based on the output, a set of entries corresponding to the largest certain number of elements in magnitude is identified. The second step involves a weighted ℓ_1 minimization where the entries outside the approximate support set are penalized with a constant weight larger than 1.

However, neither iterative reweighted ℓ_1 minimization nor weighted ℓ_1 minimization using support information takes into consideration the spatial relationship of training set. In contrast, locally weighted least squares (LWLS) combines points near a query point to estimate the

appropriate output by using a distance weighted regression, leading to the following training criterion:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \left|\left| \mathbf{\Gamma}(\mathbf{x} - \mathbf{Y}\mathbf{w}) \right|\right|_2^2 \right\}. \tag{4}$$

Given weighting function K and distance function d as well as query point q, weights are computed by

$$\mathbf{\Gamma}_i = K(d(\mathbf{Y}_i, q)). \tag{5}$$

Employing LWLS, Bose et al. [3] developed a super-resolution and noise filtering algorithm, which shows to be useful in filtering the noise and approximating scattered data simultaneously. However, LWLS only involves minimizing a weighted squared error loss without considering penalty function. Unlike LWLS, we propose to weight penalty term instead, which turns out to be more effective for face hallucination.

2.2 Bayesian inference of proposed weighted method

From Bayesian perspective, standard ℓ_1 norm regularization corresponds to regularized least squares with Laplace prior imposed on coefficients. All entries in the coefficient vector are assumed to share quite identical variance, so called homoscedasticity in statistics. However, Heteroscedasticity is a major concern of regression related problems. A collection of random variables is heteroskedastic when there are sub-populations that have different variances from others. For example, random variables of larger values often have errors of higher variances, leading to weighted least-square method accounting for the presence of heteroscedasticity on error statistics.

To examine whether heteroscedasticity exists in sparse representation, we order its coefficients with respect to Euclidean distances and then calculate their standard deviations at different distances. Euclidean distances are measured between the input patch and basis patches in training set. As shown in Fig. 1, the coefficients of near bases have larger standard deviations than others, with the standard variances monotonically decreasing as growing Euclidean distances. By a simple function fitting, this variation trend can even be approximated with $f\ (d) = \frac{0.0022}{d}$, where d denotes distance. Evidently, the solution of linear model in face hallucination problem obeys heteroskedasticity rather than homoskedasticity.

Given the observation vector $\mathbf{x}$, MAP technique is often used to estimate the coefficient vector $\mathbf{w}$:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}} \left\{ \log P\left(\mathbf{w}\middle|\mathbf{x}\right) \right\}, \tag{6}$$

where $\log P(\mathbf{w}|\mathbf{x})$ is the log-likelihood function. It follows from Bayesian theorem that the following holds:

$$P\left(\mathbf{w}\middle|\mathbf{x}\right) = \frac{P(\mathbf{w}\mathbf{x})}{P(\mathbf{x})} = \frac{P(\mathbf{w})P\left(\mathbf{x}\middle|\mathbf{w}\right)}{P(\mathbf{x})}, \tag{7}$$

or,

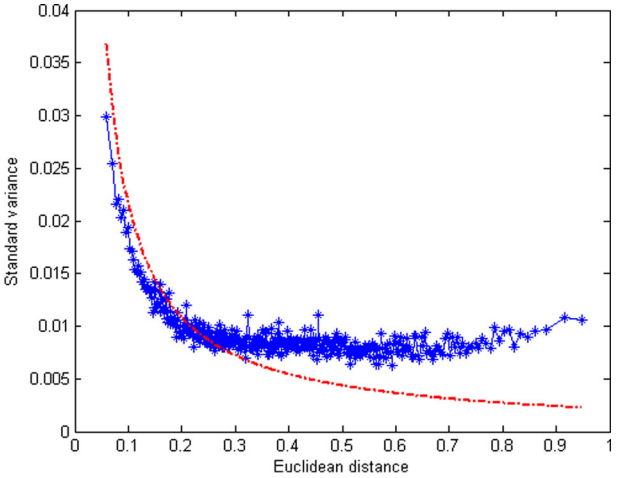


Fig. 1 Standard deviation of sparse representation coefficients declines with increasing Euclidean distance, which can be approximated with the inverse function of the distance as illustrated in dashed red line

$$\log P\left(\mathbf{w}\middle|\mathbf{x}\right) = \log P(\mathbf{w}) + \log P\left(\mathbf{x}\middle|\mathbf{w}\right) - \log P(\mathbf{x}). \tag{8}$$

Since the third term of the log-likelihood function is constant, it can be eliminated from the optimization. Consequently, MAP estimator is alternatively given by

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}} \left\{ 1ogP(\mathbf{w}) + \log P\left(\mathbf{x}\middle|\mathbf{w}\right) \right\}. \tag{9}$$

To solve (9), the conditional probability $P(\mathbf{x}|\mathbf{w})$ and the prior probability $P(\mathbf{w})$ must be specified in advance. As usually assumed, the observation vector $\mathbf{x}$ is corrupted by zero-mean i.i.d. Gaussian noise, we have

$$P\left(\mathbf{x}\middle|\mathbf{w}\right) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{Y}\mathbf{w}\|_2^2\right), \tag{10}$$

where σ^2 describes noise level.

The coefficient vector $\mathbf{w}=[w_1, w_2, \ldots, w_M]^T$ is assumed to obey independent zero-mean multivariate Laplace distribution:

$$P(\mathbf{w}) = \prod_{i=1}^{M} \left\{ \frac{1}{2\mu_i} \exp\left(-\frac{|w_i|}{\mu_i}\right) \right\}, \tag{11}$$

where $\mu_i = \frac{\sigma_i}{\sqrt{2}}$ is a scale parameter indicating the diversity and σ_i describes the standard variance of individual entry. As discussed before, coefficients are highly related to distances and near bases take large magnitudes. Because w_i is viewed as a zero-mean random variable, σ_i or μ_i actually describes the coefficient energy. Therefore, we assume that scale parameter μ_i is inversely proportional to distance. For simplification,

let $\mu_i = \frac{\mu}{d_i}$, where d_i denotes Euclidean distance and μ is a common scale parameter, we can rewrite (11) as

$$P(\mathbf{w}) = \frac{\prod_{i=1}^{M} d_i}{(2\mu)^M} \exp\left\{ -\frac{\sum_{i=1}^{M} |d_i w_i|}{\mu} \right\}. \tag{12}$$

According to (10) and (12), we have

$$\begin{aligned} logP(\mathbf{w}) + \log P\left(\mathbf{x}\middle|\mathbf{w}\right) = N\log\frac{1}{\sqrt{2\pi}\sigma} + M\log\frac{1}{2\mu} \\ + \sum_{i=1}^{M} \log(d_i) - \frac{1}{2\sigma^2}\left\{ \|\mathbf{x}-\mathbf{Y}\mathbf{w}\|_2^2 + \lambda\sum_{i=1}^{M} |d_i w_i| \right\}, \end{aligned} \tag{13}$$

where $\lambda = \frac{2\sigma^2}{\mu}$ is the regularization parameter. From (13), the objective in (9) is equivalent to minimizing the cost function $\{\|\mathbf{x}-\mathbf{Y}\mathbf{w}\|_2^2 + \lambda\sum_{i=1}^{M} |d_i w_i|\}$. Hence the ultimate optimization objective becomes

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \left|\left|\mathbf{x}-\mathbf{Y}\mathbf{w}\right|\right|_2^2 + \lambda\sum_{i=1}^{M} |d_i w_i| \right\}. \tag{14}$$

Let ℓ_1 norm $||\mathbf{w}||_1 = \sum_{i=1}^{M} |w_i|$ and $\mathbf{D}$ denote the diagonal matrix with diagonal elements given by $\mathbf{D}_{ii}=d_i$, (14) can be rewritten as

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \left|\left|\mathbf{x}-\mathbf{Y}\mathbf{w}\right|\right|_2^2 + \lambda\left|\left|\mathbf{D}\mathbf{w}\right|\right|_1 \right\}, \tag{15}$$

where $\lambda \geq 0$ is an appropriately chosen regularization parameter, controlling the tradeoff between the reconstruction error and the regularization penalty. $\mathbf{D}$ is a diagonal weighting matrix with diagonal elements being Euclidean distances (or derivatives from distances). $||\mathbf{D}\mathbf{w}||_1$ is actually the weighted variant of ℓ_1 norm.

As for optimization on LWSR, by simple algebra, (15) can be turned into an ordinary ℓ_1 minimization problem. Let $\mathbf{w}' = \mathbf{D}\mathbf{w}$, and thus $\mathbf{w} = \mathbf{D}^{-1}\mathbf{w}'$, (15) can be rewritten as

$$\mathbf{w}^* = \underset{\mathbf{w}'}{\operatorname{argmin}} \left\{ \left|\left|\mathbf{x}-\mathbf{Y}\mathbf{D}^{-1}\mathbf{w}'\right|\right|_2^2 + \lambda\left|\left|\mathbf{w}'\right|\right|_1 \right\}, \tag{16}$$

which can hence be solved using popular ℓ_1 minimization numerical algorithms.

We further use a numerical simulation to show how LWSR redistributes sparse coefficients, in which the coefficients are learned by standard and weighted ℓ_1 regularization methods over a dictionary of 360 basis images. Figure 2 shows the cumulative percentage of coefficient energy versus ordered bases in ascending distance. As for weighted case, the 50 nearest bases cover more than 90 % of total energy while the rest bases carry a little. Especially, the curve remains horizontal ultimately, which shows these associated bases do not carry any energy at all. However, the cumulative percentage curve for standard ℓ_1 norm keeps steadily increasing

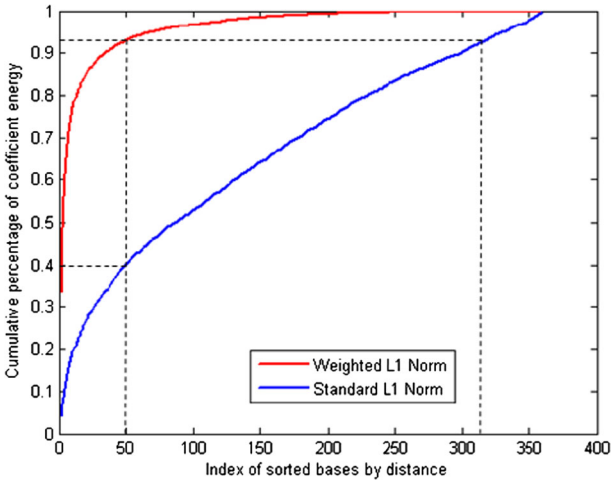


Fig. 2 Comparison of the cumulative percentage of energy in coefficient domain. The training bases are ordered in increasing distance

almost with a same slope, which shows that the bases at different distances make no significant differences in terms of magnitudes. The 50 nearest bases only account for below 40 % of the total energy and the proportion of 90 % allows over 300 bases. This simulated result manifests that LWSR enables coefficient energy concentrated on the bases close to the observed sample, thus explicitly encouraging the engagement of highly similar bases in reconstruction.

3 $\ell_{1,2}$ norm regularization for noisy face hallucination

As shown previously, regularization methods are highly associated with specific prior assumptions on solution. Laplace prior in (11) and Gaussian prior expressed in the follow-up (17) are two well-known ones proposed so far, which correspond to the ℓ_1 norm sparse representation and squared ℓ_2 norm ridge regression, respectively. In this section, we develop two new prior models and extend the regularization from ℓ_1 norm to $\ell_{1,2}$ norm, especially for robustly hallucinating face images in the presence of noise.

$$P(\mathbf{w}) = \frac{1}{(2\mu)^M} \exp\left(- \frac{\left|\left|\mathbf{w}\right|\right|_2^2}{\mu} \right), \tag{17}$$

where squared ℓ_2 norm is the form of $||\mathbf{w}||_2^2 = \sum_{i=1}^{M} |w_i|^2$.

3.1 Statistical properties of coefficients

Intuitively, noisy images may contain less sparsity than noiseless ones. To confirm this issue, three kinds of images, namely, original noise free images, simulated noisy images and real-world images are used to show sparse statistics. Original images are chosen from public face database FEI [10]. Simulated noisy images are generated by adding zero-mean Gaussian white noise of different standard deviations to the original images. The real-world images are taken by ordinary surveillance camera under low-light conditions and are thus noisy and blurring.

We learn coefficients using ordinary linear regression instead of sparse representation (the generation of coefficients is free from the penalty term). When sufficient coefficient samples are collected, we plot their histograms in Fig. 3.

In Fig. 3, peaks at or close to zero decrease steadily as the noise level grows strong. As usually stated, a signal is sparse if most entries of its coefficient vector are zero or close to zero. Therefore, this phenomenon shows that images tend to be less sparse with the increasing amount of noise.

The under-sparse properties of noisy images can be well-comprehended by analogy with their characteristics in frequency domain. Noisy images often contain richer frequency spectrum and thus are difficult to describe in a few low-frequency components. Sparse transform may be viewed as a generalization of the Fourier transform, except for that its bases are no longer fixed orthogonal Fourier basis functions but nondeterministic sample vectors from specific training sets. Transform coefficients in both are used to describe the linear combination relationship of bases. Sparsity is characterized by the ratio of coefficients being zero or close to zero. The shrinkage of sparseness is roughly comparable to the reduction of the proportion of low frequency components suffering from noise.

Alternatively, the under-sparsity of noisy images can be justified from the perspective of the solution of linear regression model. It is known that ordinary linear regression problem $\mathbf{w}^* = \text{argmin}_{\mathbf{w}} \left\{ ||\mathbf{x}-\mathbf{Y}\mathbf{w}||_2^2 \right\}$ has a closed-form solution $\mathbf{w}^*=(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{x}$. We can see that the magnitude of an individual entry in solution is proportional to the correlation term $\mathbf{Y}^T\mathbf{x}$. Under normal circumstances, a small number of the most relevant bases will result in large non-zero coefficients (exactly a manifestation of sparsity) while most of the bases with weak relevance generate zero coefficients. With the emergence and intensification of the noise, the relevance measured by correlation function will become far weaker than ever. Consequently, the situation in which the observed sample is extremely related to the minority in training set no longer remains. Nearly equivalent relationship in terms of relevance with all bases will be bridged, instead. Thereby, the peaks in coefficient distributions in Fig. 3 gradually decrease as the noise grows strong.

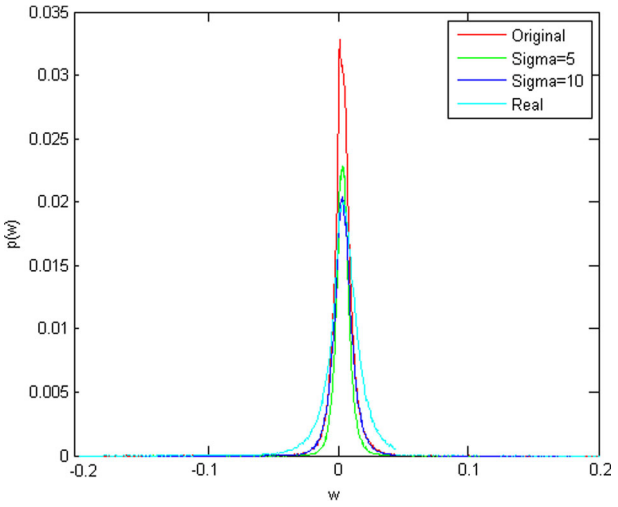


Fig. 3 Coefficient histograms on the images with different noise levels

3.2 Prior models

Since the sparsity in solution space varies with noise intensity, a question naturally arises: does Laplace distribution sufficiently describe coefficient prior in both noise free and noisy scenarios? To answer this question, we intend to use Laplace probability density function (pdf) to fit the actual distributions in Fig. 3 and then examine the consistency between actual and fitted distributions. Besides, we would like to try Gaussian fitting in (17) as well. To figure out the most accurate fitting, we suggest a new pdf as follows:

$$P(\mathbf{w}) = \frac{1}{(2\mu)^M} \exp\left(-\frac{\|\mathbf{w}\|_2}{\mu}\right), \tag{18}$$

which is herein referred to as ℓ_2 distribution with scale parameter $\mu>0$ and ℓ_2 norm $||\mathbf{w}||_2 = \left(\sum_{i=1}^{M} |w_i|^2\right)^{1/2}$. Let $\mathbf{w}_i$, $i=1,2,\ldots,L$ be a set of coefficient vectors, where L is the number of vectors, the maximum likelihood (ML) estimate of scale parameter μ is given by

$$\mu = \frac{\sum_{i=1}^{L} \|\mathbf{w}_i\|_2}{LM}. \tag{19}$$

Further, the multiplicative combination of Laplace and ℓ_2 distributions will lead to another pdf in the following form:

$$P(\mathbf{w}) = \frac{1}{(2\mu)^M} \exp\left(-\frac{(1\text{-}\alpha)\|\mathbf{w}\|_1 + \alpha\|\mathbf{w}\|_2}{\mu}\right), \tag{20}$$

which is called $\ell_{1,2}$ distribution with $\ell_{1,2}$ norm being combination form of $|(1\text{-}\alpha)\|\mathbf{w}\|_1+\alpha\|\mathbf{w}\|_2$. Parameter α, $0\le\alpha\le1$, controls the fraction of ℓ_2 in the mixture distribution. We empirically set it by maximizing the match of the fitted $\ell_{1,2}$ distribution with the actual one, typically 0.9. Scale parameter μ is estimated by

$$\mu = \frac{(1-\alpha)\sum_{i=1}^{L} \|\mathbf{w}_i\|_1 + \alpha\sum_{i=1}^{L} \|\mathbf{w}_i\|_2}{LM}. \tag{21}$$

In the follow-up, we examine their respective fitting performance to the actual distributions under different noise levels. To obtain the fitted probability functions, we collect an image set containing 400 face images from FEI database [10]. Among them, 360 images are used for basis images and the remaining 40 for testing. Each image is divided into 180 small patches for performing patch-wise training. Thus, each patch is associated with a 360-dimensional coefficient vector. By a simple algebra, the total length of coefficients is $40\times180\times360=2592000=2.592$ million. Four fitted distributions are evaluated, namely, Laplace, Gaussian, ℓ_2 and $\ell_{1,2}$, with their parameters estimated with ML method.

As seen in Fig. 4, the sharply-peaked Laplace distribution quite agrees with the actual one under the ideal noise free conditions. As the noise level increases, in other words, sparsity declines, Laplace does not fulfill the best approximation any more but seems over-sparse. In

contrast, Gaussian, ℓ_2 and $\ell_{1,2}$ priors gradually approach the actual distributions of noisy images. Among them, $\ell_{1,2}$ is the most fitted one while Gaussian is the least fitted. Even for the case of noise free, $\ell_{1,2}$ still gives perfect fitting since it is a compromise of ℓ_1 and ℓ_2. Gaussian prior seems far under-sparse for both noiseless and noisy images, while ℓ_2 is relatively neutral in terms of sparseness. These statistical observations indicate that $\ell_{1,2}$ can appropriately model the latent prior of unknown coefficients in the presence or absence of noise. We can draw the same conclusion when we experience other kinds of noise such as pepper salt, or change the noise levels.

3.3 $\ell_{1,2}$ norm regularization

Under MAP framework, Gaussian prior corresponds to the well-known ridge regression. Similarly, substituting the priors in (18) and (20) into MAP estimator in (9), we can deduce the ℓ_2 regularization in (22) and $\ell_{1,2}$ regularization in (23), respectively.

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \left|\left|\mathbf{x}-\mathbf{Y}\mathbf{w}\right|\right|_2^2 + \lambda \left|\left|\mathbf{w}\right|\right|_2 \right\}, \tag{22}$$

where regularization parameter $\lambda = \frac{2\sigma^2}{\mu}$.

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \left|\left|\mathbf{x}-\mathbf{Y}\mathbf{w}\right|\right|_2^2 + \lambda_1 \left|\left|\mathbf{w}\right|\right|_1 + \lambda_2 \left|\left|\mathbf{w}\right|\right|_2 \right\}, \tag{23}$$

where regularization parameters are given by

$$\lambda_1 = \frac{2(1-\alpha)\sigma^2}{\mu},\ \lambda_2 = \frac{2\alpha\sigma^2}{\mu}. \tag{24}$$

The mixture structure in (23) enables us to adaptively select ℓ_1 or ℓ_2 penalty and thus overcomes disadvantages using either of them.

We then briefly investigate the sparsity in a qualitative way. In [9], authors proved that the penalty functions have to be singular at origin so as to produce sparse solutions. Following this finding, the ridge regression turns out non-sparse since its squared ℓ_2 norm is differentiable at zero. As for ℓ_2 norm, it follows from $||\mathbf{w}||_2^2 = \sum_{i=1}^{M} |w_i|^2 \le \left(\sum_{i=1}^{M} |w_i|^2 + 2 \sum_{i,j,i\neq j}^{M} |w_i||w_j| \right) = ||\mathbf{w}||_1^2$ that $\|\mathbf{w}\|_2 \le \|\mathbf{w}\|_1$ holds, which implies that ℓ_2 is less sparse than ℓ_1. Since $\ell_{1,2}$ just compromises ℓ_1 and ℓ_2 in terms of sparsity, they allows to induce sparsity more aggressively in an order of ℓ_1, $\ell_{1,2}$ and ℓ_2,. This proposition quite agrees with the probabilistic observation in Fig. 4. Against the aggressive sparsity of ℓ_1 norm, the moderately sparse $\ell_{1,2}$ regularization is expected to be favorable to noise robust face hallucination.

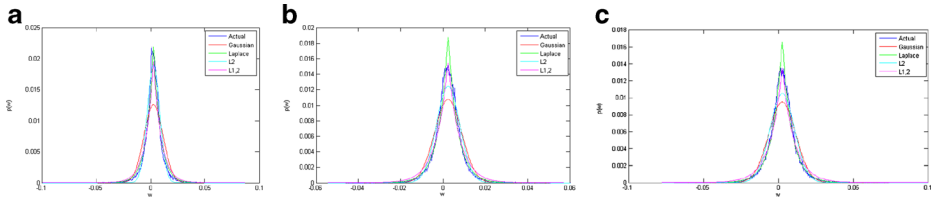


Fig. 4 Actual and fitted distributions of coefficients on images with different noise levels. **a** Original noiseless images. **b** Noisy images with σ=5 Gaussian noise. **c** Noisy images with σ=10 Gaussian noise

Recall the idea of local weighting presented in Section 2, which can be extended readily to ℓ_2 regularization and $\ell_{1,2}$ regularization. Accordingly, the corresponding weighted variants in (22) and (23) are expressed as

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \left|\left| \mathbf{x}-\mathbf{Y}\mathbf{w} \right|\right|_2^2 + \lambda \left|\left| \mathbf{D}\mathbf{w} \right|\right|_2 \right\}, \tag{25}$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \left|\left| \mathbf{x}-\mathbf{Y}\mathbf{w} \right|\right|_2^2 + \lambda_1 \left|\left| \mathbf{D}\mathbf{w} \right|\right|_1 + \lambda_2 \left|\left| \mathbf{D}\mathbf{w} \right|\right|_2 \right\}. \tag{26}$$

3.4 Numerical solution

To facilitate optimization on (25) with Lagrangian multiplier method, we rewrite its objective function as

$$L(\mathbf{w}) = \frac{1}{2} \left|\left| \mathbf{x}-\mathbf{Y}\mathbf{w} \right|\right|_2^2 + \lambda \left|\left| \mathbf{D}\mathbf{w} \right|\right|_2 . \tag{27}$$

In spite of non-differentiability of ℓ_2 norm at origin, the solution of face hallucination problem cannot be zeroed vector. So, the partial derivative to (27) can be safely deduced as $\frac{\partial L}{\partial \mathbf{w}} = \mathbf{Y}^T(\mathbf{Y}\mathbf{w}-\mathbf{x}) + \lambda \frac{\mathbf{D}^2\mathbf{w}}{||\mathbf{D}\mathbf{w}||_2}$. Further, let $\frac{\partial L}{\partial \mathbf{w}} = 0$, we have

$$\mathbf{w} = \left(\mathbf{Y}^T\mathbf{Y} + \frac{\lambda \mathbf{D}^2}{\left|\left| \mathbf{D}\mathbf{w} \right|\right|_2} \right)^{-1} \mathbf{Y}^T\mathbf{x}. \tag{28}$$

The optimal solution $\mathbf{w}^*$ to (25) can thus be obtained by iteratively updating (28). In practice, the initial value of iteration procedure is set to $(\mathbf{Y}^T\mathbf{Y})/(\mathbf{Y}^T\mathbf{x})$ instead of zeroed vector due to non-differentiability of ℓ_2 norm at zero. Expression $(\mathbf{Y}^T\mathbf{Y})/(\mathbf{Y}^T\mathbf{x})$ is just the solution of ordinary linear regression. Note that, the solution to (22) can be seen as a special case in which weighted matrix $\mathbf{D}$ is an identity matrix.

Again, we derive the numerical solution to (26). Because $\frac{\partial \|\mathbf{D}\mathbf{w}\|_1}{\partial \mathbf{w}} = \mathbf{Q}\mathbf{w}$, where $\mathbf{Q}$ is a $M \times M$ diagonal matrix with $\mathbf{Q}_{ii} = \frac{d_i}{|w_i|}$, $1 \le i \le M$, quite different from ℓ_2 norm, ℓ_1 norm is non-differentiable at any zero entries. To avoid any singular problems, we introduce a smooth approximation $\mathbf{Q}_{ii} = \frac{d_i}{|w_i|+\varepsilon}$, where $\varepsilon > 0$ is a small constant. Similarly, we get the following iterative formula:

$$\mathbf{w} = \left(\mathbf{Y}^T\mathbf{Y} + \lambda_1\mathbf{Q} + \frac{\lambda_2 \mathbf{D}^2}{\left|\left| \mathbf{D}\mathbf{w} \right|\right|_2} \right)^{-1} \mathbf{Y}^T\mathbf{x}. \tag{29}$$

In practice, the above two iterative algorithms are found to be very stable and usually reach reasonable convergence tolerance within a few iterations.

4 Face hallucination via LWSR

In this section, we particularly address several practical issues on face hallucination in order that the proposed underlying representation can be fully explored. We first discuss the

determination methods on distance and weighting functions as well as regularization parameters, and then formulate a unified face hallucination algorithm based on studied techniques.

4.1 Distance function

LWSR introduces distance-inducing weights into penalty function. However, when the illumination level of test image significantly deviates from those of training images, the efficacy of LWSR may be restricted. Experimentally, it is unexpectedly observed that the peak signal noise ratio (PSNR) of LWSR method is not higher than that of SR any more when LR test images are reduced or raised in illumination, especially for reduced illumination. This degradation is primarily due to the fact that the ordinary Euclidean distance cannot truthfully measure similarity under varied illumination conditions. To address this issue, LWSR should account for the possible illumination deviations among test image and basis images in practical applications. By incorporating illumination compensation into measurement of Euclidean distance, we thereby form an illumination-calibrated Euclidean distance metric:

$$d_m(i,j) = \left|\left|g\mathbf{X}(i,j) - \mathbf{Y}_m(i,j)\right|\right|_2, \tag{30}$$

where $g = \sqrt{\frac{(\mathbf{Y}_m)^T\mathbf{Y}_m}{\mathbf{X}^T\mathbf{X}}}$ is a gain factor, $\mathbf{X}$ denotes test image and $\mathbf{Y}_m$ stands for the m-th basis image in training set.

4.2 Weighting function

The performance of locally weighted sparse regularization is closely related to the construction of weighting matrix. As presented before, weighting matrix $\mathbf{D}$ is composed of Euclidean distances or their derivatives, but how to figure out the derivatives is an unresolved issue. Without loss of generality, we use a mapping function $\phi(d): R \rightarrow R$ to convert Euclidean distance d into a new value. Such a mapping function is referred to as weighting function in this paper. Two basic issues should be taken account into when designing a weighting function. First, its maximum outcome should emerge at a maximum distance. Second, it should grow smoothly as the distance increases.

To devise an ideal weighting function, we begin with the investigation into distance statistics. Figure 5 shows that distance histogram is dominated by small components, which can be well approximated by gamma distribution

$$P(d) = \frac{1}{\Gamma(k)\theta^k} d^{\,k-1} \exp\left(-\frac{d}{\theta}\right), \tag{31}$$

where $k>0$ is a shape parameter and $\theta>0$ is a scale parameter. According to the property of gamma distribution, these two parameters can be estimated from mean μ and variance σ^2 of samples: $k = \frac{\mu^2}{\sigma^2}$, $\theta = \frac{\sigma^2}{\mu}$. The distance value related to peak (formally called mode in statistics) is given by $M=(k-1)\theta$.

Provided that the gray level is normalized and the image is divided into low-dimensional patches ($3\times3=9$ pixels), as shown in Fig. 5, the samples with distances less than 1 possess an overwhelming proportion. Following the paradigm of non-uniform conversion, fine quantization interval should be given to the region of higher probability density. Thus, $\phi(d)$ should

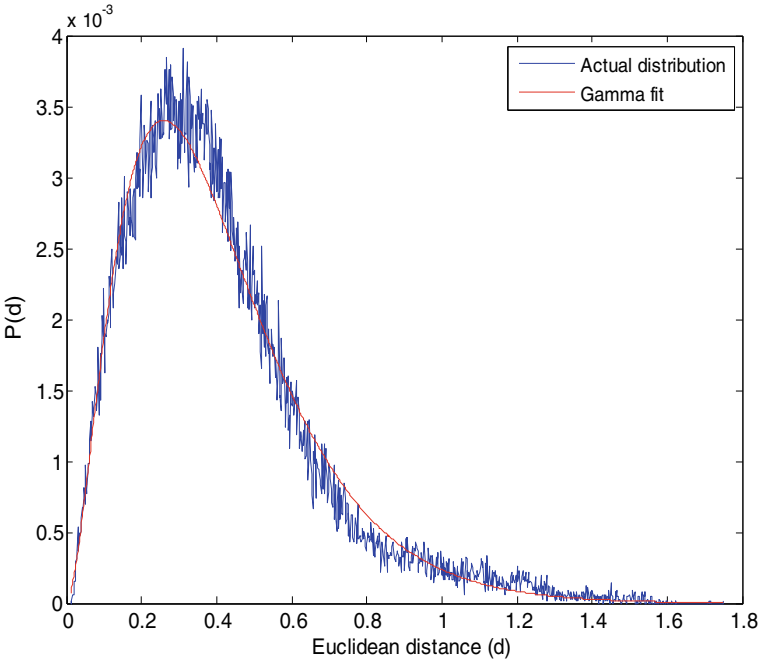


Fig. 5 Probability distribution of Euclidean distances

perform a non-linear rather than a uniform mapping so as to achieve fine conversion for small distances. Specifically, $\phi(d)$ varies more sensitively to small distances than to large ones. We consider a simple weighting function which just raises the distance to a power:

$$\phi(d) = d^n, \tag{32}$$

where power term n is a real number larger than zero, controlling the sensitivity or discriminability of mapping. We discuss the choice of optimal n in the following.

Mathematically, the sensitive response of the function to a given variable is equivalent to a large slope with respect to this variable. We consider derivative $\phi'(d)=nd^{n-1}$. When $n<1$, the derivative monotonically decreases versus d with $\phi(d)$ more sensitive to small d. However, if the power term n is too small, most of the distances will take small slope while only a very minor proportion of small distances allow very large slope. More importantly, this small portion of distances allowing large slope may occur at a low probability in the distance distribution, as shown in Fig. 5. To prevent such improper mapping, we should find an optimal value for n, which is responsible for the mapping sensitivity and distribution density. For a given d value, we further consider the curve shape where the slope function $\phi'(d)$ varies with respect to n. Clearly, $\phi'(d)$ is a concave function with a maxima, which first increases and then decreases as n increases. The n value corresponding to maxima is the desirable optimal power term. Because $\phi'(d)$ depends on a specific d value, we have to first find out a reliable d. Theoretically, we should use the mode parameter of gamma distribution (i.e., the distance value at peak) so that the distances which are most likely to occur enjoy the highest mapping sensitivity.

In methodology, we derive $\phi^{''}(d)=d^{n-1}(1+n\ln d)$ from the slope function $\phi^{'}(d)$. The maxima of $\phi^{'}(d)$ corresponds to $\phi^{''}(d)=0$, leading to $n=-1/\ln d$. Using aforementioned way, we get the mode value of gamma distribution, denoted as d_0, and then the optimal power term n_0 is obtained by

$$n_0 = -1/\ln d_0. \tag{33}$$

4.3 Regularization parameter

Regularization parameter has a significant influence on the learning performance of underlying representation methods. Generally, too large values will lead to low regression precision or under-fitting while too small values often result in poor robustness or over-fitting. Statistical literatures suggest some generic parameter identification techniques, such as L-curve [4] and cross-validation [25]. However, they are seldom applied to practical face hallucination because they need prior knowledge about the range of parameter and spend a large amount of computation. In practice, previous face hallucination methods [16–18, 32] manually choose regularization parameter, which keeps tuning parameter until the best results are attained. This empirical practice is relatively reliable, but it is cumbersome and computationally expensive, suffering from repeated probing.

As a byproduct of MAP inference, (24) gives deterministic estimated values for λ_1 and λ_2. Therefore, we attempt to use these estimated values to perform optimization. To validate the effectiveness, we testify four cases including noise free images, noisy images with two different Gaussian noise levels, and the real-world images. In this verification, not only estimated values are compared with tuned values, but also the hallucinated results with respect to each value are evaluated in PSNR and structural similarity (SSIM) (because the reference HR images are unavailable in real-world scenarios, PSNR and SSIM are not applicable). In the empirical tuning method, the candidate parameters are experienced in a trial-and-error tactic and the one corresponding to the maximum PSNR or the best visual effect will be used. The numerical results are tabulated in Table 1, which is computationally based on the statistics of coefficients generated by linear regression method. It is clear that the estimated values of λ_1 and λ_2 are highly close to tuned values and maintain the same hallucination performance in terms of PSNR and SSIM. Note that, we need to carry out an extra step to supply (24) with statistics of coefficients.

4.4 Face hallucination algorithm

For face hallucination, the training set is composed of HR and LR face image pairs. The primary task is to reconstruct the HR face from the input LR counterpart over training set.

First of all, we divide the training face images and the input LR face image into mutually overlapping small patches using the same dividing scheme as [23]. Overlapping region is used to smooth blocking artifacts suffering from patch-wise manipulation. The regularization parameters, distances and weights are pre-computed with Eqs. (24), (30) and (32) prior to formal LWSR optimization. For each input LR image patch, it is approximated by a linear combination of the LR patches from training set at the same position, where the optimal linear combination coefficients are resolved with (26) with the help of the iterative algorithm described in Section 3.4. Since the LR patches and the HR patches share the same manifold topology [7], a new HR patch of the same position can be synthesized by keeping the

coefficients and replacing the LR training image patches with the corresponding HR ones. By concatenating all the HR patches to their corresponding positions and averaging pixel values in the overlapping regions, we can get an estimate of the HR face.

5 Experiments and results

In this section, we conduct face hallucination experiments to extensively testify proposed LWSR. Experiments are performed on two public face databases: FEI [10] and CAS-PEAL-R1 [14]. A brief description of the datasets is provided along with the details of the experiments in Section 5.1. In Section 5.2, local weighting on penalty is verified against other representative weighted methods. Then, we validate the robustness against noise of several sparse penalties in Section 5.3. The objective metrics, i.e., PSNR and SSIM index, will be reported. Additionally, subjective performance is compared with the state-of-the-art methods such as Chang's neighbor embedding (NE) [7] and Yang's sparse representation (SR) [32] in Sections 5.4 and 5.5, where simulated LR images and real-world LR images are under test, respectively.

5.1 Datasets and parameter settings

FEI face database The public FEI face database contains 400 images covering both genders, different races and facial expressions. Among them, 360 images are randomly chosen as the training set, leaving the remaining 40 images for testing. All the test images are absent completely in the training set. All the HR images are cropped to 120×100 pixels, and then the LR images with 30×25 pixels are generated by smoothing and down-sampling by a factor of 4.

CAS-PEAL-R1 face database CAS-PEAL-R1 is a Chinese face dataset, containing 30,871 images of 1040 subjects. We only use the neutral expression and normal illumination face of each subject from the frontal subset for experiments. In all the 1040 frontal face images, we randomly select 1000 images for training and leave the other 40 images for testing. All the images are cropped to 112×100 pixels, thus the size of LR face images are in 28×25 pixels.

Test images Experiments are conducted on simulated LR images and real-world LR images. Simulation experiments consist of noiseless images and noisy images. We view the original images in databases as noise free ones. To obtain noise corrupted images, zero-mean Gaussian noise with two different standard variances ($\sigma=5$ and $\sigma=10$) are added to original images for simulating noisy images. The real-world images are from realistic scenarios, contaminated by unknown types of noise. Since training images in databases have already been aligned, they should be pre-aligned by the positions of five manually selected feature points.

Parameter settings As stated in [17], patch size has a considerable influence on the super-resolved results. Empirically, we set the HR patch size as 12×12 pixels and the overlap between neighbor patches as 4 pixels, while the corresponding LR patch size is 3×3 pixels with an overlap of 1 pixel.

For the sake of fair comparison, we tune the parameters for all comparative methods to their best results. The number of the neighbors K in NE is set to 50. The regularization parameter in regularized methods is firstly assigned a preliminary value using the estimation way in

Section 4.3, and then slightly refined to pursue the best performance. The $\ell_{1,2}$ regularization uses the parameters λ_1 and λ_2 in Table 1, and the parameter λ for different classes of images in ℓ_1 norm SR is tabulated in Table 2.

5.2 Effects of different weighted methods

Various techniques are proposed to weight penalty and error terms in regularization methods [3, 5, 13, 19]. To validate the superiority of our locally weighted method, we select several representative weighted methods as the anchors: locally weighted least squares (LWLS) [3], reweighted ℓ_1 minimization (RLM) [5] and weighted ℓ_1 with prior information (WPI) [19]. Among them, RLM iteratively updates ℓ_1 penalty weights from the current solution so that maximally sparse representation can be found as optimization progresses, while WPI assigns two different weights to the elements in two sets by different probabilities of being nonzero. In contrast, LWLS combines the points near a query point to estimate the appropriate outcome through a distance weighted regression.

For fair comparison, all weighted methods are implemented under ℓ_1 norm SR. Hence LWLS weights error term while retains standard ℓ_1 norm penalty term for maintaining the same SR structure. RLM executes in an optimal reiteration times of 2. In WPI, we construct approximate support set by identifying the nearest bases instead of the largest elements in magnitude as suggested in [19] because the former turns out to be more helpful. The standard ℓ_1 norm SR is used for benchmark in this experiment.

The results in PSNR and SSIM are shown in Table 3, which indicate that all weighted variants really offer the better results than original non-weighted standard SR. Comparatively, the biggest gain is given by LWSR as the locally weighted penalty promotes the accuracy and stability of solution. More precisely, the distance-inducing weighting matrix imposed on the penalty function encourages the near bases to cover large coefficients in magnitude. On the other hand, the local weighting imposed on error term in LWLS only leads to a minor gain, partially because face hallucination is not a classic regression analysis problem after all. Classic regression analysis benefits from local learning on error estimation in that it usually stores the training data in memory and finds relevant data in the database to answer a particular query. The role of iterative reweighted ℓ_1 minimization is to further sparsify the solution, yet the gain by RLM is small. This evidently manifests that excessive pursuit for sparsity is not

Table 1 Determination of regularization parameters

Images	Methods	λ_1	λ_2	PSNR	SSIM
Noise free	Estimation	4.8e-5	4.2e-4	32.54	0.9137
	Tuning	1.0e-4	1.0e-4	32.56	0.9138
σ=5 Gaussian	Estimation	5.8e-3	5.2e-2	30.28	0.8682
	Tuning	5.0e-3	5.0e-2	30.30	0.8683
σ=10 Gaussian	Estimation	1.3e-2	1.1e-1	28.43	0.8230
	Tuning	1.5e-2	1.0e-1	28.42	0.8227
Real-world	Estimation	1.2e-2	1.1e-1	n/a	
	Tuning	1.0e-2	2.0e-1		

Table 2 Regularization parameter of SR

Images	Noise free	σ=5	σ=10	Real-world
λ	1.0e-4	5.0e-3	1.5e-2	5.0e-3

only unnecessary but also useless for face hallucination problem. In contrast, the fact that WPI offers a considerable gain shows that external prior information is very useful for sparsity related applications.

5.3 Effects of different regularization methods

In Section 3, ℓ_2 or $\ell_{1,2}$ is justified to be more favorable than ℓ_1 for face hallucination in the presence of noise. To verify the effects of different regularization methods experimentally, we conduct another quantitative test to evaluate their PSNR and SSIM. Various typical regularization penalty functions commonly used in image super-resolution are included, such as squared ℓ_2 norm (referred to as ${\ell_2}^2$) [17] in ridge regression, ${\ell_q}^q$ norm [6] in bridge regression and ℓ_q norm [31]. Particularly, [17, 31] are our previously proposed ones. Note that ${\ell_q}^q$ differs from ℓ_q in that it raises the latter to a power of q. Since the role of locality-inducing weighting has been sufficiently verified in the above subsection, in Table 4, we merely compare their respective locally weighted versions.

From experimental results in Table 4, we remark that:

1) ℓ_1 gives better results than ℓ_2 under the conditions of noise free, but the outcome is just the opposite in the presence of noise. As shown in Fig. 4 in Section 3.2, these results exactly agree with their respective prior approximations to actual distributions.
2) Among all penalty functions, ${\ell_2}^2$ offers the worst results for noise free case. This is partially due to the fact that ${\ell_2}^2$ is a non-sparse model as pointed out by Fan and Li [9]. On the other hand, as the regression coefficients of noisy images contain less sparseness, ${\ell_2}^2$ is slightly better than aggressively sparse ℓ_1, but is still inferior to ℓ_2.
3) Both ${\ell_q}^q$ and ℓ_q give quite the same results as ℓ_1 SR for noise free images as the optimal parameter q is tuned to 1, which again shows that ℓ_1 SR indeed fits noiseless images. As a generalization of ${\ell_2}^2$, ${\ell_q}^q$ leads to slightly improved results for noisy images. So does the comparison of ℓ_2 with ℓ_q. Meanwhile, unlike strict ℓ_q norm, ${\ell_q}^q$ penalty results in lack of sparsity when $q>1$, so ${\ell_q}^q$ is inferior to ℓ_q.
4) Regardless of noisy or noiseless images, $\ell_{1,2}$ always achieves the best results. Especially, it substantially outperforms any counterparts in noisy scenarios. Moreover, different from sole ℓ_2, $\ell_{1,2}$ does not lead to performance degradation relative to ℓ_1 for noise free case, owning to a compromise of ℓ_1 and ℓ_2.

Table 3 Comparison of different weighted methods

Metrics	SR	LWLS	RLM	WPI	LWSR
PSNR	32.17	32.28	32.32	32.57	**32.93**
SSIM	0.9069	0.9087	0.9091	0.9118	**0.9172**

Bold texts show the largest values

Table 4 Results by different regularization methods

Images	Metrics	ℓ_1	ℓ_2^2	ℓ_q^q	ℓ_2	ℓ_q	$\ell_{1,2}$
Noise free	PSNR	**32.93**	32.76	**32.93**	32.80	**32.93**	**32.93**
	SSIM	**0.9172**	0.9148	**0.9172**	0.9155	**0.9172**	**0.9172**
σ=5 Gaussian	PSNR	30.15	30.29	30.36	30.45	30.47	**30.61**
	SSIM	0.8536	0.8601	0.8628	0.8684	0.8688	**0.8714**
σ=10 Gaussian	PSNR	27.98	28.33	28.41	28.56	28.56	**28.67**
	SSIM	0.7911	0.8134	0.8147	0.8224	0.8224	**0.8253**

Bold texts show the largest values

$\ell_{1,2}$ though yieds small improvements against the second best method ℓ_q [31], yet it enjoys far lower computational complexity. To illustrate this point more clearly, we provide a bit more investigation into their solutions. As shown in (29), the numerical optimization on $\ell_{1,2}$ minimization involves computations of ℓ_1 norm $||\mathbf{w}||_1 = \sum_{i=1}^{M} |w_i|$ and ℓ_2 norm $||\mathbf{w}||_2 = \left(\sum_{i=1}^{M} |w_i|^2\right)^{1/2}$. Similarly, ℓ_q minimization problem iteratively computes ℓ_q norm $||\mathbf{w}||_q = \left(\sum_{i=1}^{M} |w_i|^q\right)^{1/q}$. The former mainly handles multiplicative calculation: $w_i \times w_i$, while the latter has to deal with more computationally expensive power function: $|w_i|^q$. Experimentally, hallucination task employing $\ell_{1,2}$ regularization is roughly 6 times faster than one employing ℓ_q regularization (53 vs. 379 in cpu-time unit), either on FEI database or on CAS-PEAL-R1 database.

In short, $\ell_{1,2}$ is far superior over either ℓ_1 or ℓ_2 in terms of ubiquity and robustness. In the subsequent subjective experiments, $\ell_{1,2}$ eventually turns out to be the most promising and applicable representation for face hallucination task.

5.4 Simulation results

In subjective tests, our methods are evaluated against Chang's NE [7], Yang's SR [32] and plain Bicubic interpolation. Our two locally weighted sparse regularization variants such as ℓ_1 norm (WSR hereafter for short) and $\ell_{1,2}$ norm (WL1,2 hereafter for short) get involved in this comparison. The experiments are carried out on FEI and CAS-PEAL-R1databases with both noise free and noisy images. Some of randomly selected subjective results on CAS-PEAL-R1 and FEI databases are shown in Figs. 6 and 7, respectively.

Similar performance can be seen in Figs. 6 and 7. For noise free case, we only discern slight visual differences among the results of NE, SR, WSR and WL1,2 except that Bicubic generates heavy blurring effects. For the cases of noise, SR based methods (SR, WSR and WL1,2) can remove noise more thoroughly than NE because NE relies on ordinary least squares. Comparatively, WSR partially reproduces facial features more clearly (see mouths) than SR as it retains relatively more high-frequency components. By a careful examination, the

Fig. 6 Comparison of the results of different methods on CAS-PEAL-R1 face database: **a** Bicubic interpolation; **b** Chang's NE; **c** Yang's SR; **d** Proposed WSR; **e** Proposed WL1,2; **f** Original HR faces (ground truth). Top 3 rows for noise free images; middle 3 rows for noisy images with σ=5; bottom 3 rows for noisy images with σ=10 ▶


a
b
c
d
e
f


a
b
c
d
e
f

Fig. 7 Comparison of the results of different methods on FEI face database: **a** Bicubic interpolation; **b** Chang's NE; **c** Yang's SR; **d** Proposed WSR; **e** Proposed WL1,2; **f** Original HR faces (ground truth). Top 3 rows for noise free images; middle 3 rows for noisy images with $\sigma=5$; bottom 3 rows for noisy images with $\sigma=10$

hallucinated images by WL1,2 look smooth and clean while SR or WSR still exposes some un-smoothed noisy artifacts, especially on cheeks and noses. In addition, with the increasing amount of interference noise, ranging from $\sigma=5$ to $\sigma=10$, the visual perception variations among different methods seems more distinguishable. Note that, the results on CAS-PEAL-R1 database are roughly cleaner than those on FEI partially because more training images are used for CAS-PEAL-R1. This experiment again confirms that relatively conservative sparse representation greatly benefits to noise suppression in the process of hallucination.

5.5 Experiments on real-world images

The input LR face images of all the above experiments are formed by smoothing and down-sampling HR images, which cannot indicate the true spatial degradation relationship between the HR image and the degraded LR one. In a realistic condition, it is too difficult for us to simulate the image degradation process or know how different types of image degradation processes affect the statistics of images. In order to further testify the efficacy of our method, we perform two more experiments with real-world surveillance images and a photo picture.

First of all, robustness against noise is verified with low-quality images under realistic surveillance imaging conditions. They are captured by a commercial surveillance camera in a low-lighting environment, where the persons are far from the camera. Hence they unavoidably contain noise and blurring effects. This experiment is conducted on CAS-PEAL-R1 database.

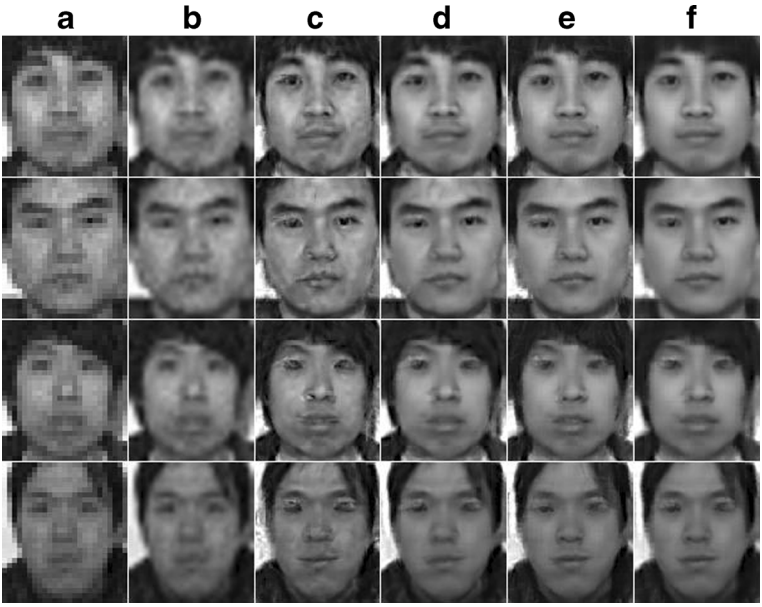


Fig. 8 Comparison of the results of different methods on real-world surveillance images: **a** Input LR faces; **b** Bicubic interpolation; **c** Chang's NE; **d** Yang's SR; **e** Proposed WSR; **f** Proposed WL1,2

Fig. 9 The first row is the raw photo picture and the second row is the extracted faces used for testing

Firstly, we manually extract the faces of interest and crop them to the size of 28×25 pixels. And then, the images are manually aligned to training faces with respect to the eye, mouth and nose positions. Finally, the input LR faces are generated by converting the cropped faces to grayscale. Figure 8 shows the super-resolved results by our proposed and anchor methods.

We can obviously observe that WL1,2 offers the best results in the sense of de-noising and visual fidelity. More specifically, W1,2 can preserve the same high fidelity for facial features as Bicubic interpolation, and meanwhile avoids the inherent over-smoothness caused by up-scaling. The visual quality by NE, SR and WSR appears overall similar although WSR can eliminate considerably more annoying noise artifacts. Admittedly, the eyes in the last two rows show slight deformation, even for the results hallucinated by WL1,2, which is primarily due to the inaccurate alignment. The distinction between WSR and WL1,2 once again confirms that sparsity of the underlying representation should not be over-emphasized in practical face hallucination applications. Experimentally, our proposed method can yield acceptable results even though the test images are generated in poor imaging conditions or in the presence of heavy non-Gaussian noise.

In the second experiment, a color photo picture of MPEG meeting participants is used for testing. The raw photo and the extracted faces of interest are shown in Fig. 9. Seen from the second row of Fig. 9, the raw faces expose somewhat noise. This experiment is performed on FEI database so that LR input images are required in 30×25 pixels. Because the extracted raw faces are slightly larger than 30×25 pixels in size, a resizing is performed prior to pre-alignment to training faces. In accordance with training sets, the raw faces in RGB format have to been converted to YCbCr format. Since humans are more sensitive to illuminance changes, we only perform super-resolution reconstruction in the luminance component. We therefore interpolate the color components using plain Bicubic interpolation. The super-resolved results for a collection of test images are shown in Fig. 10.

Compared with other four methods, we can see that WL1,2 is much more robust against noise of unknown types or levels while maintaining sufficient visual quality. In essence, WL1,2 regularization enjoys both advantages of saliency features readily grasped by ℓ_1 and moderate smoothness against noise contributed by ℓ_2. Evidently, our proposed WL1,2 method can produce reasonable results even though the test images are drastically different from the training samples.

Fig. 10 Comparison of the results of different methods on real-world photo picture: **a** Input LR faces; **b** Bicubic ▶ interpolation; **c** Chang's NE; **d** Yang's SR; **e** Proposed WSR; **f** Proposed WL1,2. (Note that the effect is more pronounced if the figure of electronic version is zoomed)


a
b
c
d
e
f

6 Conclusion

In this paper, we have proposed a novel locally weighted sparse regularization technique to boost face hallucination performance. Our technique takes into account two primary statistics of coefficients in sparse representation domain, i.e., heteroskedasticity and under-sparsity of noisy images. Accordingly, distance-inducing weighting is enforced on penalty function of sparse regularization to favor the locality of dictionary learning. A penalty of ℓ_1 and ℓ_2 mixed norms with conservative sparseness is introduced to model the less-sparse nature of noisy images. The resulting distance-weighted $\ell_{1,2}$ norm regularization can significantly promote the accuracy, stability and robustness of solution. Extensive experimental results on public face databases and real-world images show its superiority over the state-of-the-art methods for face hallucination in terms of PSNR, SSIM and subjective visual quality.

Acknowledgments This work was supported by the National Natural Science Foundation of China (61172173,61172174,61501413,61502354), the Fundamental Research Funds for the Central Universities (2042014kf0286, 2042014kf0212), and Natural Science Fund of Hubei Province (2015CFB406).

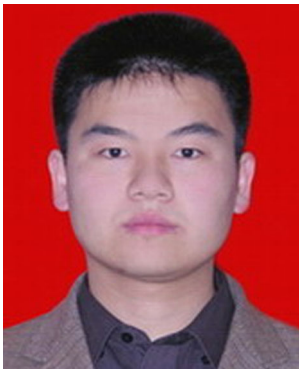
References

1. Baker S, Kanade T (2002) Limits on super-resolution and how to break them. IEEE Trans Pattern Anal Mach Intell 24(9):1167–1183
2. Bevilacqua M, Roumy A, Guillemot C, Alberi ML (2012) Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: British Mach Vis Conf pp. 1–10
3. Bose NK, Ahuja NA (2006) Super-resolution and noise filtering using moving least squares. IEEE Trans Image Process 15(8):2239–2248
4. Bose NK, Koo J (2001) Advances in super-resolution using L-curve. In: Proc IEEE Sym on Circuits and Syst pp. 433–436
5. Candès EJ, Wakin MB, Boyd S (2008) Enhancing sparsity by reweighted ℓ1 minimization. J Fourier Anal App 14(5):877–905
6. Cetin M, Karl WC (2001) Feature-enhanced synthetic aperture radar image formation based on non-quadratic regularization. IEEE Trans Image Process 10(4):623–631
7. Chang H, Yeung DY, Xiong YM (2004) Super-resolution through neighbor embedding. In: Proc IEEE Conf Comput Vis Pattern Recognit pp. 275–282
8. Charbonnier P, Blanc-Feraud L, Aubert G, Barlaud M (1997) Deterministic edge-Preserving regularization in computed imaging. IEEE Trans Image Process 6(2):298–311
9. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96(456):1348–1360
10. FEI Face Database, http://fei.edu.br/~cet/facedatabase.html
11. Frank IE, Friedman JH (1993) A statistical view of some chemometrics regression tools. Technometrics 35: 109–135
12. Freeman W, Pasztor E, Carmichael O (2000) Learning low-level vision. Int J Comput Vis 40(1):25–47
13. Friedlander MP, Mansour H, Saab R, Yılmaz O (2012) Recovering compressively sampled signals using partial support information. IEEE Trans Inf Theory 58(2):1122–1134
14. Gao W, Cao B, Shan SG, Chen XL, Zhou DL, Zhang XH, Zhao DB (2008) The CAS-PEAL large-scale Chinese face database and baseline evaluations. IEEE Trans SMC-A 38(1):149–161
15. He R, Zheng W-S, Tan T, Sun Z (2014) Half-quadratic based iterative minimization for robust sparse representation. IEEE Trans Pattern Anal Mach Intell 36(2):261–275
16. Jia Z, Wang H, Xiong Z (2011) Fast face hallucination with sparse representation for video surveillance. In: Proc Asian Conf Pattern Recognit pp. 179–183
17. Jiang J, Hu R, Wang Z, Han Z (2014) Noise robust face hallucination via locality-constrained representation. IEEE Trans Multimed 16(5):1268–1281

18. Jung C, Jiao L, Liu B, Gong M (2011) Position-patch based face hallucination using convex optimization. IEEE Signal Process Lett 18(6):367–370
19. Khajehnejad MA, Xu W, Avestimehr AS, Hassibi B (2009) Weighted ℓ1 minimization for sparse recovery with prior information. In: Proc IEEE Sym Inf Theory pp. 483–487
20. Li Y, Lin X (2004) An improved two-step approach to hallucinating faces. In: Proc IEEE Conf Image and Graphics pp. 298–301
21. Li Q, Lin N (2010) The Bayesian elastic net. Bayesian Anal 5(1):151–170
22. Liu C, Shum H, Freeman W (2007) Face hallucination: theory and practice. Int J Comput Vis 7(1):15–134
23. Ma X, Zhang J, Qi C (2010) Hallucinating face by position-patch. Pattern Recognit 43(6):3178–3194
24. Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the Lasso. Ann Stat 34(3):1436–1462
25. Nguyen N, Milanfar P, Golub G (2001) A computationally efficient super resolution image reconstruction algorithm. IEEE Trans Image Process 10(3):573–583
26. Nikolova M, Ng MK (2005) Analysis of half-quadratic minimization methods for signal and image recovery. SIAM J Sci Comput 27:937–966
27. Park J-S, Lee S-W (2008) An example-based face hallucination method for single-frame, low-resolution facial images. IEEE Trans Image Process 17(10):1806–1816
28. Suo F, Hu F, Zhu G (2011) Robust super-resolution reconstruction based on adaptive regularization. In: Proc. Int. Conf. Wireless Comm. Signal Process pp. 1–4
29. Tang Y, Yan P, Yuan Y, Li X (2011) Single-image super-resolution via local learning. Int J Mach Learn Cybern 2(1):15–23
30. Vapnik V, Bottou L (1993) Local algorithms for pattern recognition and dependencies estimation. Neural Comput 5(6):893–909
31. Wang Z, Hu R, Wang S, Jiang J (2014) Face hallucination via weighted adaptive sparse regularization. IEEE Trans Circ Syst Video Technol 24(5):802–813
32. Yang J, Tang H, Ma Y, Huang T (2010) Image super-resolution via sparse representation. IEEE Trans Image Process 19(11):2861–2873
33. Yu K, Zhang T, Gong Y (2009) Nonlinear learning using local coordinate coding. In: Proc Neural Inf Process Syst pp. 2223–2231
34. Zeyde R, Elad M, Protter M (2010) On single image scale-up using sparse-representations. Int. Conf. Curves and Surfaces pp. 711–730
35. Zhang J, Zhao C, Xiong R, Ma S, Zhao D (2012) Image super-resolution via dual-dictionary learning and sparse representation. In: IEEE Sym on Circuits and Syst pp. 1688–1691
36. Zou H, Hastie T (2005) Regularization and variables election via the elastic net. J R Stat Soc B 67(2):301–320

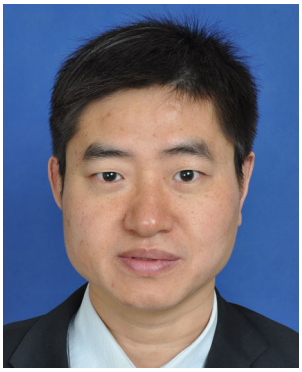
Zhongyuan Wang received the B.S. degree and M.S degree in computer science from Wuhan University, Wuhan, China, in 1995 and 2001, and he received the Ph.D. degree in Communication and Information System in Wuhan University in 2008. From 2001, he worked as a member of research staff in National Multimedia Software Engineering Research Center of Wuhan University. His research interests include video compression, multimedia communications.

Ruimin Hu received the B.S and M.S degrees from Nanjing University of Posts and Telecommunications, Nanjing, China, in 1984 and in 1990, and Ph.D degree in Communication and Electronic System from Huazhong University of Science and Technology, Wuhan, China, in 1994. Dr. Hu is the director of National Engineering Research Center for Multimedia Software, Wuhan University and Key Laboratory of Multimedia Network Communication Engineering in Hubei province. He is Executive Chairman of the Audio Video coding Standard (AVS) workgroup of China in Audio Section. He has published two books and over 100 scientific papers. His research interests include audio/video coding and decoding, video surveillance and multimedia data processing.



Junjun Jiang received the B.S. degree in Information and Computing Science from School of Mathematical Sciences, Huaqiao University, Quanzhou, China, in 2009, and the Ph.D. degree in communication and information system from School of Computer, Wuhan University, Wuhan, China, in 2014. He is currently an Associate Professor with the School of Computer Science, China University of Geosciences. His research interests include applications of image processing and pattern recognition in video surveillance, image super-resolution, image interpolation, and face recognition.

Zhen Han received the B.S degree in computer science and technology and Ph.D degree in computer application technology from Wuhan University, Wuhan, China, in 2002 and in 2009 respectively. Now he is a lecturer in school of computer, Wuhan.



Zhenfeng Shao received his PhD degree from Wuhan University, China, in 2004. He is now professor of the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, China. His research interests are image retrieval, image fusion, and urban remote sensing application.