Article:
Xu, Chen, Zhang, Zhihong, Wang, Beizhan et al. (2 more authors) (2017) Recovering Variations in Facial Albedo from Low Resolution Images. Pattern Recognition. pp. 1-37. ISSN 0031-3203

https://doi.org/10.1016/j.patcog.2017.09.019

Recovering Variations in Facial Albedo from Low Resolution Images

Xu Chen[a], Zhihong Zhang[a,*], Beizhan Wang[a], Guosheng Hu[b], Edwin R.Hancock[c]

[a]*Xiamen University, Xiamen, China*
[b]*Anyvision company, Belfast, UK.*
[c]*Department of Computer Science, University of York, York, YO10 5GH, UK.*

Abstract

Recovering facial albedo from low quality face images is a challenging task which arises when face recognition is attempted in the wild. Low quality of facial images is usually caused by extrinsic factors such as low resolution and noises, and intrinsic ones such as expressions. Existing research recovers facial albedo by dealing with the extrinsic and intrinsic factors separately. However, it is more natural and potentially more useful to approach albedo recovery by removing the two effects simultaneously. In this paper, we present a novel framework which can recover facial albedo by jointly solving these for both the extrinsic and intrinsic sources of uncertainty. This framework models albedo recovery problem by a joint optimization process which alternatively (1) removes intra-personal variations and (2) performs super resolution. To deal with the intrinsic sources of albedo variability, we use a linear model. To handle extrinsic problems associated with low quality imaging, we use a sparse coding method which is applied to super resolution. The proposed method can also significantly improve the performance of face recognition and clustering in case of very low resolution and in the presence of various facial variations. Extensive experiments and comparisons are conducted on the AR and FERET face databases. Experimental results show the effectiveness of the proposed method.

Keywords: Facial albedo estimation, Low quality facial image, Sparse coding, ADMM

*Corresponding author
Email address: zhihong@xmu.edu.cn (Zhihong Zhang)

1. Introduction

A fundamental challenge in face analysis is to enhance the quality of facial images that are captured using poor quality imaging equipment or limited imaging conditions, and achieve acceptable recognition rates. Factors which can adversely affect this process are different illumination conditions, facial expressions, partial occlusions and low resolution images. These factors not only prove challenging to the human visual system, but also adversely affect the performance of automatic face analysis. In this work, we categorize these sources of image degradation into those that are intrinsic to the subject and their setting (lighting, expression, etc) and those that are extrinsic, and are artifacts of the imaging process or device, i.e. low resolution or noise.

To remove the intrinsic effects and faithfully recover the facial albedo (facial texture) without degradations, to the best of our knowledge, there is no general solution to hand. However, various solutions have been proposed to deal with individual sources of intrinsic variation. Specifically, illumination normalization methods have been used to remove the effects of illumination variations while maintaining facial albedo. These methods project the images to an illumination-free space in either 2D or 3D. In 2D it is usual to perform this projection in the frequency domain[1] [2], while in 3D lighting models from the graphics domain, such as the Phong model [3] or Spherical Harmonics [4], are used. Expression normalization is also performed to convert a face with variations in expression into the one with a neutral expression [5], and then applies machine learning methods for expression transfer. Finally, occlusion modeling is usually tackled using sparse representation methods [6] and is based on the assumption that occlusions are sparse. Each of these methods is singly very effective to remove one source of intrinsic facial variation, but to our knowledge, there is no integrated way to deal with them simultaneously.

To alleviate the problems of low resolution, image super resolution (SR) attempts to increases high-frequency components whilst removing undesirable effects, such as resolution degradation, blur and noise. For an observed low resolution image $\mathbf{y}$, the problem of image SR is generally modeled as $\mathbf{y} = \mathbf{SHx} + e$ with the goal of recov-

ering a high-resolution (HR) image $\mathbf{x}$ from $\mathbf{y}$, where e is a small noise term, $\mathbf{H}$ is a blur filter, and $\mathbf{S}$ represents a down-sampling operator. The dimension of $\mathbf{y}$ is significantly smaller than that of $\mathbf{x}$; thus there are an infinite number of possible HR images $\mathbf{x}$ that can generate the same LR image $\mathbf{y}$. To cope with this ill-posed nature of image restoration, prior knowledge is imperative and of pivotal importance to eliminate the uncertainties in the recovery process. Early studies [7] tend to focus on the priors associated with natural high quality images and using such priors to regularize the HR estimation. Recent research focuses on taking advantage from HR/LR training image pairs and a variety of different methods [8, 9, 10, 11, 12, 13] have been proposed to model the relationship between the HR image and its corresponding LR image. Instead of utilizing a general prior over natural images, such methods implicitly incorporate prior knowledge using a complex mapping function, and in so doing achieve state-of-the-art SR performance. However, by embedding prior information into the learned mapping function between corresponding LR and HR images, such methods can only process well aligned images without any intrinsic variations, which limits their applications to the real world scenarios.

In this paper, we present a novel image enhancement framework which jointly estimates facial albedo and performs super resolution. Figure 1 shows the task we aim to solve. Occlusions as well as variations of expressions and illumination on face images share a similar pattern across different subjects. Figure 2 shows some examples of residuals between normal and abnormal (face with occlusion, expression and illumination variations) face images. The patterns of these residuals can be exploited to develop statistical models of face appearance. In fact, the residuals between normal faces and their corresponding variations have limited variability. In this paper, we propose a model which captures the global structure of normal faces and the variations reading in a residual decomposition model. Such a model not only utilizes prior knowledge concerning global face structure, but also the distribution of residual patterns between normal and abnormal faces. We combine the global and local models into a single objective function, which is optimized to obtain high resolution face images without variations. Such a hybrid model enables us to utilize both global and local information to generate better "hallucinations" in the high resolution domain. Specifically,

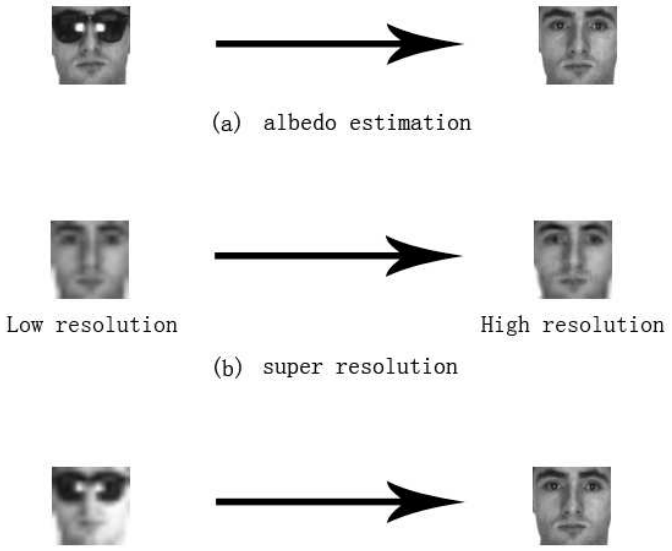


Figure 1: Joint albedo estimation and face super resolution

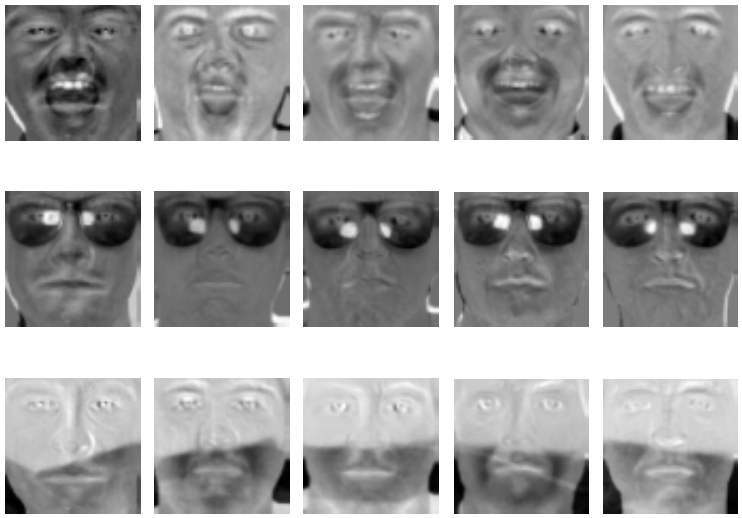


Figure 2: Intra-personal variations.

position-specific patch terms in our model prove to be particularly beneficial since they enable us to generate local structures in HR with sharp edges and fine details. The global term, on other hand, takes full advantage of face structure information. Furthermore, another advantage of the proposed model is that we are able to generate HR face images with normal position, even though the input low resolution face has expression or pose variations. The contributions of this paper are three-fold.

- Existing research independently investigates two tasks: (1) facial albedo estimation and (2) super resolution. To our knowledge, we are the first to propose a face image enhancement framework which can jointly estimate facial albedo and perform face super resolution. It is more effective to simultaneously solve these two tasks, which both aim to recover the facial albedo or texture from low quality images.

- The proposed framework can be modeled as a non-convex optimization problem. We propose an efficient alternating optimization strategy which interleaves removing intrinsic facial variations and performing super resolution.

- Existing albedo estimation methods can only deal with single sources of intrinsic facial image variation, such as illumination variation, because these methods assume the facial image is formed by the non-linear interplay between albedo and the intrinsic sources of facial variation. The non-linearity assumption does not generalize well. For example, the interplay between albedo and illumination cannot be used for both albedo and occlusion estimation. To solve this problem, we propose a unified linear framework, which represents a face images as the sum of facial albedo and intra-personal variations. Via this a linear model, our framework can model more diverse sources of facial image variations.

- Experiments demonstrate that the proposed method can also significantly improve the performance of face recognition and clustering when given very low resolution images with various facial variations.

The paper is organized as follows. In Section 2 we present a discussion of related work. The novel framework is introduced in Section 3. The proposed methods are

evaluated in Section 4. Finally, Section 5 discusses the results and draws the paper to a conclusion.

2. Related Work

In this section, we discuss related work on albedo estimation and face super resolution.

2.1. Albedo Estimation

We first focus on methods which remove the effects of various facial image variations (illumination, expression and occlusion) while maintaining the facial albedo. The removal of illumination, expression and occlusion is a widely studied problem in the face recognition literature. Expression removal and facial morphing, on the other hand, is widely studied in the graphics literature

Illumination normalization is widely employed as a pre-processing step for face recognition because illumination variations could significantly degrade face recognition performance. One popular solution is projecting the facial image into the frequency domain, in which the components of very low and very high frequency are removed. Thus the SNR (Signal Noise Ratio) is better. A typical example is the DoG (difference of Gaussian) filter. However, it is not straightforward to determine how many components should be removed. Another solution is to learn an illumination invariant representation [14, 15, 16, 17]. For example, O. Arandjelović et al [14] combines a weak photometric model with a statistical model to achieve invariance to illumination, pose and user motion pattern variation. Recently, lighting models originally developed in the graphics literature have also been applied to the problem of facial illumination normalization. These models include the Lambertian, Phong [3], Spherical Harmonic [4] models, and usually use a 3D shape prior such as 3D Morphable Model [3] to estimate the lighting conditions (lighting direction and strength) and then remove its effects.

Occlusion and expression variations also affect the visible facial image, and have been well investigated. One well known method for removing facial occlusion is sparse

coding [6]. An observed face is modeled as the summation of facial albedo and occlusion effects. Based on the assumption that occlusions on the face are sparsely distributed, the occlusion term is constrained using ℓ_1-norm. One effective expression normalization method is the 3D morphable model [5] (3DMM), which can capture variations in both albedo and expression. After a 3DMM has been fitted to an input image, the coefficients of expression variations can be set to zero, thus removing the effects of expression variations from the input image.

Although extensive research has been conducted to deal with each particular source on intrinsic facial image variation, very few studies simultaneously deal with more than one source of variation. In this work, we propose an integrated framework which can simultaneously estimate the various sources of facial image degradations detailed in Section 3.

2.2. Face Super-Resolution

In the seminal work [7], prior on the derivatives of the high resolution image is formulated as a function of spatial location, and a pyramid based algorithm is proposed to gradually enhance the resolution of face image. After [7], different face super resolution methods have been proposed to generate better results. Liu et al. [18, 19] proposed a two-step statistical modeling approach that integrates both a global parametric model and a local nonparametric model, and achieved very promising face hallucination results. O. Arandjelović [20] successfully reconstruct the personal subspace in the high-dimensional image space from a low-dimensional input without any assumptions on the nature of appearance that the subspaces represent. Recent studies [8, 21, 9, 10, 11, 22, 23] share a similar idea of using patch-based method to model the prior information of local structure of face images. These methods assume that each patch from the considered images can be well represented using a linear combination of a few atoms from a dictionary. By forcing LR and the corresponding HR patches to have the same sparse coefficients, Yang et al. [21] are the first to apply the idea of sparse representation to the face image SR. The method offline trains HR and LR dictionaries to sparsely decompose HR and LR image patches, respectively. Given a LR patch, a sparse coefficient vector is computed using the LR dictionary by solving a

ℓ_1-norm minimization problem. The desired HR patch is reconstructed by combining the HR dictionary with the same coefficients. The similar intuitive can also be found in [23].

Although the above methods have demonstrated promising performance as generative models for facial images, most of them focus on processing well aligned face images without any significant intrinsic variations. Such strong requirements have limited their applications to real world image data.

3. Methodology

3.1. Albedo Estimation

A face image is generated by the interplay of face albedo and other intrinsic variations (IVs) such as illumination, expression and occlusion. Motivated by [24, 25], in this work, we model the problem of albedo estimation as follow:

$$\mathbf{y} = \mathbf{y}^a + \mathbf{y}^i + \mathbf{n} \tag{1}$$

where $\mathbf{y}$ is an observed face, $\mathbf{y}^a$ and $\mathbf{y}^i$ are albedo and variations respectively, $\mathbf{n}$ is noise. We assume $\mathbf{y}^a$ and $\mathbf{y}^i$ are represented by linear dictionaries, therefore, Eq.(1) can be rewritten as:

$$\mathbf{y} = \mathbf{N}\boldsymbol{\alpha} + \mathbf{V}\boldsymbol{\beta} + \mathbf{n} \tag{2}$$

where $\mathbf{N}$ and $\mathbf{V}$ are dictionaries for albedo and variation, respectively; $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are free parameters. The sparsity prior, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated by solving the cost function:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\| [\mathbf{N}, \mathbf{V}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} - \mathbf{y} \right\|_2^2 + \lambda \left\| \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_1 \tag{3}$$

where λ is a weighting parameter. The solutions of Eq.(3) are denoted as $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. Therefore, the estimate of albedo is obtained by

$$\hat{\mathbf{y}} = \mathbf{N}\hat{\boldsymbol{\alpha}} \tag{4}$$

3.2. Sparse Coding for Face Super-Resolution

Sparse coding approaches for super resolution create a sparse representation in a patch space by training a codebook of dictionary atoms. Yang et al.[8, 21] first proposed an approach for super resolution based on this idea. Specifically, they used a sparsity constraint to jointly train the low resolution (LR) and high resolution (HR) dictionaries, therefore, LR patches and their corresponding HR counterparts can be reconstructed in a sparse manner using the same sparsity representation. Given the learned $\mathbf{D}_l$ and $\mathbf{D}_h$, the HR patch $\mathbf{y}_h$ can be estimated from its corresponding LR patch $\mathbf{y}_l$. It is assumed that $\mathbf{y}_l$ and $\mathbf{y}_h$ share the same linear sparse representation of $\mathbf{D}_l$ and $\mathbf{D}_h$. Given $\mathbf{y}_l$ and $\mathbf{D}_l$, the sparse representation coefficients $\mathbf{c}$ can be obtained by solving:

$$\min_{\mathbf{c}} \|\mathbf{y}_l - \mathbf{D}_l\mathbf{c}\|_2^2 + \lambda \|\mathbf{c}\|_1 \tag{5}$$

As the estimated $\hat{\mathbf{c}}$ is shared with $\mathbf{D}_h$ and $\mathbf{y}_h$, the estimated HR patch is $\hat{\mathbf{y}_h} = \mathbf{D}_h\hat{\mathbf{c}}$.

Motivated by [8, 21], Zeyde et al.[26] proposed another approach which is more efficient and accurate. Instead of jointly training the HR and LR dictionaries, the K-SVD algorithm [27] is used to train the LR dictionary and the pesudo-inverse is used to compute the HR dictionary. To improve the algorithm efficiency, PCA can be used to reduce the dimensionality of the features extracted from LR patches.

3.3. Joint Albedo Estimation and Face Super-Resolution

In this paper we aim to recover a HR facial albedo image from a LR face image subject to various intrinsic variations including illumination and occlusion. To achieve this, we first use a low-rank matrix decomposition processing in order to remove the redundant and noise related elements of $\mathbf{V}$ (the dictionary of variations) in Section 3.3.1. We then present a joint albedo estimation and face super resolution framework. According to this framework, the HR and LR dictionaries $\mathbf{D}_h$ and $\mathbf{D}_l$ are trained offline, and as a result $\mathbf{D}_h$ and $\mathbf{D}_l$ are assumed known in our framework. To jointly solve the albedo estimation and face super resolution problem, we propose an efficient optimization algorithm detailed in Section 3.3.3.

3.3.1. Intra-Personal Variations Dictionary Construction

As introduced in Section 2.1, $\mathbf{N}$ and $\mathbf{V}$ are the dictionaries for albedo and intrinsic image variation, respectively. However, the way that $\mathbf{V}$ is constructed according to [24, 25] will bring not only the expected intrinsic facial image variations, but also undesirable redundant information and noise.

To overcome this problem, we introduce a filtering step based on low-rank matrix decomposition [28]. The problem can be stated as follows.

$$\min_{\mathbf{V},\mathbf{E}} \|\mathbf{V}\|_* + \lambda \|E\|_1 \qquad s.t \quad \mathbf{V}_0 = \mathbf{V} + \mathbf{E} \tag{6}$$

where $\mathbf{V}_0$ is the raw intrinsic facial variation dictionary constructed according to [25], and $\|\cdot\|_*$ represents the nuclear norm. This process decomposes the raw dictionary $\mathbf{V}_0$ into a low-rank matrix $\mathbf{V}$ and a sparse error matrix $\mathbf{E}$. The former represents the expected intrinsic facial image variations, while the latter denotes the redundant information and noise. Figure 3 shows the effect of the filtering procedure, where the middle and third row are examples of facial image variations before and after filtering, respectively.

3.3.2. Problem Formulation

It is a challenging problem to recover a HR facial albedo from a LR image with degradations. To our knowledge, very little research has investigated this task. Estimating facial albedo and conducting face super resolution simultaneously can be formulated as the optimization problem:

$$\min \left(\mathbf{E}^a + \mathbf{E}^s\right) \tag{7}$$

$$\mathbf{E}^a(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \left\| [\mathbf{N}, \mathbf{V}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} - \mathbf{y} \right\|_2^2 + \lambda \left\| \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_1 \tag{8}$$

$$\mathbf{E}^s(\mathbf{c}_{ij}) = \eta \sum_{i,j} \|\mathbf{P}_{ij}\mathbf{FHN}\boldsymbol{\alpha} - \mathbf{D}_l\mathbf{c}_{ij}\|_2^2 + \tau \sum_{i,j} \|\mathbf{c}_{ij}\|_1 \tag{9}$$

Clearly, $\mathbf{E}^a$ is the cost function for albedo estimation while $\mathbf{E}^s$ is that for face super resolution. $\mathbf{y}$ is the LR degraded image. λ, η, τ are weighting parameters; $\mathbf{P}_{ij}$ is a

Figure 3: The effect of low-rank matrix decomposition. First row: the raw images. Middle row: the facial image variations obtained by subtracting the natural image from the raw image. Third row: the filtered variations after applying low-rank matrix decomposition.

matrix which can select one particular patch from the LR image at location (i, j); $\mathbf{H}$ is an interpolation operator; $\mathbf{F}$ is a gradient feature extractor. Finally, the estimated HR image $\hat{y} = D_h\hat{\mathbf{c}} + \mathbf{HN}\hat{\boldsymbol{\alpha}}$, where $\hat{\mathbf{c}}$ and $\hat{\boldsymbol{\alpha}}$ are the optimized solution of Eq.(7).

Note that $\mathbf{E}^a$ is the same as Eq.(3), however, $\mathbf{E}^s$ is a variant of Eq.(5). Specifically, $\mathbf{E}^s$ models the whole image, while Eq.(5) is based on just an image patch.

We combine the tasks of albedo estimation and super resolution in a single framework. In this way, these two tasks can be optimized in a joint manner. On the one hand, the optimization of albedo estimation will have the opportunity to access some useful local information from HR images achieved by super resolution, rather than to use the information from LR images. On the other hand, the albedo estimation can also benefit the super resolution part by removing undesirable facial variations and providing global structure constraint. Optimizing these two problems simultaneously can lead to a better reconstruction than optimizing them separately.

3.3.3. Optimization Procedure

The optimization problem above can be solved by using the *alternating direction method of multipliers* (ADMM) method. More specifically, Eq.(7) can be reformulated as an equivalent problem by introducing the auxiliary variables $\boldsymbol{s}_1, \boldsymbol{s}_2$ and $\boldsymbol{s}_{ij}$:

$$\begin{aligned}
&\min \left\| [\mathbf{N}, \mathbf{V}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} - \mathbf{y} \right\|_2^2 + \lambda \left\| \begin{bmatrix} \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \end{bmatrix} \right\|_1 + \eta \sum_{i,j} \|\mathbf{P}_{ij}\mathbf{FHN}\boldsymbol{\alpha} - \mathbf{D}_l \mathbf{c}_{ij}\|_2^2 + \tau \sum_{i,j} \|\mathbf{s}_{ij}\|_1 \\
&s.t. \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} - \begin{bmatrix} \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \end{bmatrix} = 0, \quad \boldsymbol{c}_{ij} - \boldsymbol{s}_{ij} = 0, \quad \forall i, j
\end{aligned} \tag{10}$$

The problem now can be solved using the method of constrained optimization. The augmented Lagrangian objective function associated with Eq.(10) is given by

$$\begin{aligned}
&\mathbf{L}_\delta(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}_{ij}, \boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_{ij}, \boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_{ij}) = \\
&\left\| [\mathbf{N}, \mathbf{V}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} - \mathbf{y} \right\|_2^2 + \eta \sum_{i,j} \|\mathbf{P}_{ij}\mathbf{FHN}\boldsymbol{\alpha} - \mathbf{D}_l \mathbf{c}_{ij}\|_2^2 + \lambda \left\| \begin{bmatrix} \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \end{bmatrix} \right\|_1 \\
&+ \tau \sum_{i,j} \|\mathbf{s}_{ij}\|_1 - \left\langle \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} - \begin{bmatrix} \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{t}_1 \\ \boldsymbol{t}_2 \end{bmatrix} \right\rangle - \sum_{i,j} \langle \boldsymbol{c}_{ij} - \boldsymbol{s}_{ij}, \boldsymbol{t}_{ij} \rangle \\
&+ \frac{\boldsymbol{\delta}}{2} \left\| \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} - \begin{bmatrix} \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \end{bmatrix} \right\|_2^2 + \frac{\boldsymbol{\delta}}{2} \sum_{i,j} \|\boldsymbol{c}_{ij} - \boldsymbol{s}_{ij}\|_2^2
\end{aligned} \tag{11}$$

where $\langle *, * \rangle$ represents the inner product operation, $\delta > 0$ is a positive penalty parameter and $\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_{ij}$ are dual variables, i.e., the Lagrange multipliers. The ADMM method solves the above problem by first solving for $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}_{ij}\}$ with $\boldsymbol{s}_1, \boldsymbol{s}_2$ and $\boldsymbol{s}_{ij}$ fixed, and then solving $\{\boldsymbol{s}_1, \boldsymbol{s}_2$ and $\boldsymbol{s}_{ij}\}$ with $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{c}_{ij}$ fixed. The iteration proceeds until convergence, while the dual variables $\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_{ij}$ are updated directly at each iteration.

The details of the updating procedure at each iteration are given below.

$$\begin{bmatrix} \boldsymbol{\alpha}^{k+1} \\ \boldsymbol{\beta}^{k+1} \\ \boldsymbol{c}_{ij}^{k+1} \end{bmatrix} = \underset{\alpha,\beta,c_{ij}}{\arg\min} \mathbf{L}_\delta(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}_{ij}, \boldsymbol{s}_1^k, \boldsymbol{s}_2^k, \boldsymbol{s}_{ij}^k, \boldsymbol{t}_1^k, \boldsymbol{t}_2^k, \boldsymbol{t}_{ij}^k) \tag{12a}$$

$$\begin{bmatrix} \boldsymbol{s}_1^{k+1} \\ \boldsymbol{s}_2^{k+1} \\ \boldsymbol{s}_{ij}^{k+1} \end{bmatrix} = \underset{s_1,s_2,s_{ij}}{\arg\min} \mathbf{L}_\delta(\boldsymbol{\alpha}^{k+1}, \boldsymbol{\beta}^{k+1}, \boldsymbol{c}_{ij}^{k+1}, \boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_{ij}, \boldsymbol{t}_1^k, \boldsymbol{t}_2^k, \boldsymbol{t}_{ij}^k) \tag{12b}$$

$$\begin{bmatrix} \boldsymbol{t}_1^{k+1} \\ \boldsymbol{t}_2^{k+1} \\ \boldsymbol{t}_{ij}^{k+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{t}_1^{k} \\ \boldsymbol{t}_2^{k} \\ \boldsymbol{t}_{ij}^{k} \end{bmatrix} - \boldsymbol{\delta} \left(\begin{bmatrix} \boldsymbol{\alpha}^{k+1} \\ \boldsymbol{\beta}^{k+1} \\ \boldsymbol{c}_{ij}^{k+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{s}_1^{k+1} \\ \boldsymbol{s}_2^{k+1} \\ \boldsymbol{s}_{ij}^{k+1} \end{bmatrix} \right) \tag{12c}$$

Note that $\mathbf{L}_\delta$ is convex with respect to $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{c}_{ij}$ while other variables are fixed. For solving Eq.(12a), we calculate the partial derivatives of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{c}_{ij}$ as follows

$$\begin{aligned} \partial_{\boldsymbol{\alpha}} \mathbf{L}_\delta =& 2\boldsymbol{N}^T(\boldsymbol{N}\boldsymbol{\alpha} + \boldsymbol{V}\boldsymbol{\beta} - y) + 2\eta \sum_{i,j} (\boldsymbol{P}_{ij}\boldsymbol{F}\boldsymbol{H}\boldsymbol{N})^T(\boldsymbol{P}_{ij}\boldsymbol{F}\boldsymbol{H}\boldsymbol{N}\boldsymbol{\alpha} - \boldsymbol{D}_l\boldsymbol{c}_{ij}) \\ & - \boldsymbol{t}_1^k + \boldsymbol{\delta}(\boldsymbol{\alpha} - \boldsymbol{s}_1^k) \end{aligned} \tag{13a}$$

$$\partial_{\boldsymbol{\beta}} \mathbf{L}_\delta = 2\boldsymbol{V}^T(\boldsymbol{N}\boldsymbol{\alpha} + \boldsymbol{V}\boldsymbol{\beta} - y) - \boldsymbol{t}_2^k + \boldsymbol{\delta}(\boldsymbol{\beta} - \boldsymbol{s}_2^k) \tag{13b}$$

$$\partial_{\boldsymbol{c}_{ij}} \mathbf{L}_\delta = -2\boldsymbol{\eta}\boldsymbol{D}_l^T(\boldsymbol{P}_{ij}\boldsymbol{F}\boldsymbol{H}\boldsymbol{N}\boldsymbol{\alpha} - \boldsymbol{D}_l\boldsymbol{c}_{ij}) - \boldsymbol{t}_{ij}^k + \boldsymbol{\delta}(\boldsymbol{c}_{ij} - \boldsymbol{s}_{ij}^k) \tag{13c}$$

Let Eq.(13a), (13b) and (13c) be equal to 0, the equations can be formed as $\mathbf{A}\boldsymbol{\theta} = \boldsymbol{b}$, where $\mathbf{A}$ is the coefficient matrix of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{c}_{ij}$ in Eq.(13). $\boldsymbol{\theta} = [\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \boldsymbol{c}_{11}^T, \ldots, \boldsymbol{c}_{mn}^T]^T$. And $\boldsymbol{b}$ is the constant term in Eq.(13). The algebraic solution is given by

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \\ \boldsymbol{c}_{11} \\ \vdots \\ \boldsymbol{c}_{mn} \end{bmatrix} = \mathbf{A}^{-1}\boldsymbol{b} \tag{14}$$

Note that since $\mathbf{A}$ is constant during iterations, $\mathbf{A}^{-1}$ can be computed offline. This

means that at each iteration, we can calculate the new $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{c}_{ij}$ by simply multiplying the precomputed $\mathbf{A}^{-1}$ by $\boldsymbol{b}$ (which changes at each iteration).

To solve Eq.(12b), similarly, we set the partial derivatives of $\boldsymbol{s}_1, \boldsymbol{s}_2$ and $\boldsymbol{s}_{ij}$ equal to 0

$$\partial_{\boldsymbol{s}_1}\mathbf{L}_\delta = \boldsymbol{\lambda}\partial\|\boldsymbol{s}_1\|_1 + \boldsymbol{t}_1^k - \boldsymbol{\delta}(\boldsymbol{\alpha}^{k+1} - \boldsymbol{s}_1) = 0 \tag{15a}$$

$$\partial_{\boldsymbol{s}_2}\mathbf{L}_\delta = \boldsymbol{\lambda}\partial\|\boldsymbol{s}_2\|_1 + \boldsymbol{t}_2^k - \boldsymbol{\delta}(\boldsymbol{\beta}^{k+1} - \boldsymbol{s}_2) = 0 \tag{15b}$$

$$\partial_{\boldsymbol{s}_{ij}}\mathbf{L}_\delta = \boldsymbol{\tau}\partial\|\boldsymbol{s}_{ij}\|_1 + \boldsymbol{t}_{ij}^k - \boldsymbol{\delta}(\boldsymbol{c}_{ij}^{k+1} - \boldsymbol{s}_{ij}) = 0 \tag{15c}$$

Thus we have that

$$\boldsymbol{s}_1^{k+1} = \begin{cases} \frac{1}{\boldsymbol{\delta}}(\boldsymbol{\delta}\boldsymbol{\alpha}^{k+1} - \boldsymbol{t}_1^{k+1} - \boldsymbol{\lambda}) & , \quad \text{if} \quad \boldsymbol{\delta}\boldsymbol{\alpha}^{k+1} - \boldsymbol{t}_1^{k+1} > \boldsymbol{\lambda} \\ \frac{1}{\boldsymbol{\delta}}(\boldsymbol{\delta}\boldsymbol{\alpha}^{k+1} - \boldsymbol{t}_1^{k+1} + \boldsymbol{\lambda}) & , \quad \text{if} \quad \boldsymbol{\delta}\boldsymbol{\alpha}^{k+1} - \boldsymbol{t}_1^{k+1} < -\boldsymbol{\lambda} \\ 0 & , \quad \text{otherwise} \end{cases} \tag{16a}$$

$$\boldsymbol{s}_2^{k+1} = \begin{cases} \frac{1}{\boldsymbol{\delta}}(\boldsymbol{\delta}\boldsymbol{\beta}^{k+1} - \boldsymbol{t}_2^{k+1} - \boldsymbol{\lambda}) & , \quad \text{if} \quad \boldsymbol{\delta}\boldsymbol{\beta}^{k+1} - \boldsymbol{t}_2^{k+1} > \boldsymbol{\lambda} \\ \frac{1}{\boldsymbol{\delta}}(\boldsymbol{\delta}\boldsymbol{\beta}^{k+1} - \boldsymbol{t}_2^{k+1} + \boldsymbol{\lambda}) & , \quad \text{if} \quad \boldsymbol{\delta}\boldsymbol{\beta}^{k+1} - \boldsymbol{t}_2^{k+1} < -\boldsymbol{\lambda} \\ 0 & , \quad \text{otherwise} \end{cases} \tag{16b}$$

$$\boldsymbol{s}_{ij}^{k+1} = \begin{cases} \frac{1}{\boldsymbol{\delta}}(\boldsymbol{\delta}\boldsymbol{c}_{ij}^{k+1} - \boldsymbol{t}_{ij}^{k+1} - \boldsymbol{\tau}) & , \quad \text{if} \quad \boldsymbol{\delta}\boldsymbol{c}_{ij}^{k+1} - \boldsymbol{t}_{ij}^{k+1} > \boldsymbol{\tau} \\ \frac{1}{\boldsymbol{\delta}}(\boldsymbol{\delta}\boldsymbol{c}_{ij}^{k+1} - \boldsymbol{t}_{ij}^{k+1} + \boldsymbol{\tau}) & , \quad \text{if} \quad \boldsymbol{\delta}\boldsymbol{c}_{ij}^{k+1} - \boldsymbol{t}_{ij}^{k+1} < -\boldsymbol{\tau} \\ 0 & , \quad \text{otherwise} \end{cases} \tag{16c}$$

Finally, the dual variables $\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_{ij}$ are updated according to Eq.(12c). For the sake of clarity, a summary of the optimization procedure is shown in Algorithm 1.

4. Experiments

In this section, we first present experiments for super resolution on publicly available databases. We then evaluate our method for face recognition. Finally, we show that our method can also improve the performance of face clustering.

4.1. Super Resolution

To evaluate the performance of the proposed method on super resolution, the benchmark AR face database [29] is used. We also use the FERET database [30] and CAS-

Input: albedo dictionary $\mathbf{N}$, variation dictionary $\mathbf{V}$, LR dictionary $\mathbf{D}_l$, HR dictionary $\mathbf{D}_h$, test sample $\mathbf{y}$, weighting parameters $\lambda, \eta, \tau, \delta$

Initialize: $\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0, \{\mathbf{c}_{ij}^0\}, \boldsymbol{s}_1^0 = \boldsymbol{\alpha}^0, \boldsymbol{s}_2^0 = \boldsymbol{\beta}^0, \boldsymbol{s}_{ij}^0 = \boldsymbol{c}_{ij}^0, \boldsymbol{t}_1^0 = \boldsymbol{t}_2^0 = \boldsymbol{t}_{ij}^0 = 0$

precompute $\mathbf{A}^{-1}$, where $\mathbf{A}$ is the coefficient matrix of Eq.(13)

for $t = 1, 2, \ldots, T$ **do**

  calculate $\boldsymbol{\alpha}^t, \boldsymbol{\beta}^t$ and $\boldsymbol{c}_{ij}^t$ through Eq.(14)

  calculate $\boldsymbol{s}_1^t, \boldsymbol{s}_2^t$ and $\boldsymbol{s}_{ij}^t$ through Eq.(16)

  calculate $\boldsymbol{t}_1^t, \boldsymbol{t}_2^t$ and $\boldsymbol{t}_{ij}^t$ through Eq.(12c)

end

Output: $\hat{y} = D_h \mathbf{c}^T + \mathbf{HN}\boldsymbol{\alpha}^T$

Algorithm 1: Joint Albedo Estimation and Face Super-Resolution

PEAL database [31] to train the face super resolution dictionaries. The face images are aligned, cropped and scaled to the size of 64×64. Figure 4 presents some of sample images selected from each of the three databases. The faces from the AR database (top row) are with various expressions, illuminations and occlusions. While the faces from the FERET database (middle row) and the CAS-PEAL database (bottom row) are mainly with neutral expressions, moderate illumination variations and without occlusions.

4.1.1. Experimental Setup

The AR database is a benchmark face database which contains over 4000 frontal face images with different facial expressions, illuminations and occlusions. They are from 126 subjects (70 men and 56 women). We adopt a 10-fold cross-validation strategy to evaluate the performance of the proposed method. That is, all subjects from the AR database are randomly separated into 10 disjoint subsets. For each experiment we only used faces of persons from one of the subsets for testing while leaving the remaining nine subsets for training. The neutral expression face images from the AR database training set are used to construct the albedo dictionary $\mathbf{N}$, while the remaining AR face images with various facial expressions, illuminations and occlusions are used to construct the variation dictionary $\mathbf{V}$. The construction of albedo and

(a) (b) (c) (d) (e)

Figure 4: Some of sample images from AR database (top), FERET database (middle) and CAS-PERL database (bottom).

variation dictionaries are adopted from ESRC [25] followed by the filtering step introduced in Section 3.3.1. The images from the FERET and CAS-PERL databases are used to train the HR dictionary $\mathbf{D}_h$ and LR dictionary $\mathbf{D}_l$ following [26]. The parameters are empirically initialized and fine-tuned on the validation set as follows: $\boldsymbol{\lambda} = 0.1, \boldsymbol{\eta} = 0.001, \boldsymbol{\tau} = 0.1, \boldsymbol{\delta} = 50, \boldsymbol{T} = 50$.

4.1.2. Performance

To the best of our knowledge, the proposed method is the first attempt to directly recover the neutral HR face image from a LR one with variations. Related works such as [20, 14, 15, 16, 17] only deal with single sources of facial image variation (e.g., illumination variation), or perform super resolution independently [8, 21, 9, 10, 11, 22, 23]. Thus, there are no peer methods for direct comparison. However, to evaluate the performance of the proposed method, we adopt a two-step image recovery strategy as a reference, in which we sequentially perform albedo estimation and super resolution. That is, for a given LR input, we first estimate the albedo of LR inputs using the method discussed in section 3.1. Then we apply several popular super resolution methods (such as the sparse methods Yang et al. [21] and Zeyde et al. [26], ANR [32], A+ [33], etc.)

Table 1: Performance of the proposed method comparing with two-step recover strategy (with A+ super-resolver) in terms of PSNR and SSIM with upscaling $\times 2$.

Image Type	Two-step recover strategy		The Proposed	
	PSNR	SSIM	PSNR	SSIM
Smile	20.66	0.6334	**21.08**	**0.6576**
Anger	20.09	0.6107	**20.47**	**0.6377**
Scream	18.95	0.5564	**19.39**	**0.5884**
Leftward direction light	19.76	0.6149	**20.20**	**0.6464**
Rightward direction light	19.57	0.6166	**20.08**	**0.6479**
Both side lights	19.05	0.5898	**19.46**	**0.6249**
Sunglasses	18.17	0.5496	**18.62**	**0.5866**
Sunglasses & Leftward direction light	17.16	0.5154	**17.94**	**0.6612**
Sunglasses & Rightward direction light	17.13	0.5100	**17.93**	**0.5559**
Scarf	17.26	0.5766	**17.44**	**0.5960**
Scarf & Leftward direction light	16.42	0.5276	**16.91**	**0.5624**
Scarf & Rightward direction light	16.20	0.5184	**16.70**	**0.5562**
Average	18.37	0.5683	**18.85**	**0.6018**

on the result of the albedo estimation procedure to obtain HR face images.

We evaluate the performance of the proposed method by comparing with the two-step image recovery strategy in terms of the *Peak Signal to Noise Ratio* (PSNR) and *Structural Similarity* (SSIM). The images from the test set with various variations (i.e., expressions, illuminations and occlusions) are interpolated and down-sampled to 32×32 (for upscaling $\times 2$) and 16×16 (for upscaling $\times 4$) as LR image inputs, while the neutral expression face images serve as ground truth.

Table 1 and 2 detail the comparison results between the two-step recover strategy (with A+[33] super-resolver) and the proposed method on upscaling $\times 2$ and $\times 4$, respectively. In the tables the bold numbers signify the best performance. In each row the proposed method achieves significantly better results than the two-step recovery strategy in terms of both PSNR and SSIM in all cases. The proposed method achieves

Table 2: Performance of the proposed method comparing with two-step recover strategy (with A+ super-resolver) in terms of PSNR and SSIM with upscaling $\times 4$.

Image Type	Two-step recover strategy		The Proposed	
	PSNR	SSIM	PSNR	SSIM
Smile	20.54	0.6316	**21.02**	**0.6422**
Anger	19.98	0.6179	**20.42**	**0.6305**
Scream	18.95	0.5669	**19.49**	**0.5886**
Leftward direction light	19.63	0.6171	**20.13**	**0.6360**
Rightward direction light	19.48	0.6166	**20.01**	**0.6331**
Both side lights	18.95	0.5964	**19.44**	**0.6177**
Sunglasses	18.14	0.5572	**18.78**	**0.5893**
Sunglasses & Leftward direction light	17.30	0.5359	**18.10**	**0.5727**
Sunglasses & Rightward direction light	17.43	0.5309	**18.14**	**0.5671**
Scarf	17.20	0.5712	**17.37**	**0.5776**
Scarf & Leftward direction light	16.22	0.5301	**16.79**	**0.5552**
Scarf & Rightward direction light	16.06	0.5254	**16.63**	**0.5539**
Average	18.32	0.5748	**18.86**	**0.5970**

Table 3: Summary of p-values for paired t-tests between PSNR and SSIM values obtained by the proposed method and the two-step recover strategy.

$\times 2$		$\times 4$	
PSNR	SSIM	PSNR	SSIM
9.8e-7	5.1e-16	8.8e-8	6.2e-9

Table 4: Performance of the proposed method comparing with two-step recover strategy (with various super-resolver) in terms of PSNR and SSIM.

		Two-step recover strategy with belowing super-resolver						Ours
		Yang et al. [21]	Zeyde et al. [26]	ANR [32]	NE + LS [32]	NE + NNLS [32]	NE + LLE [32]	
$\times 2$	PSNR	16.15	18.44	18.40	18.26	18.25	18.39	**18.85**
	SSIM	0.430	0.578	0.572	0.542	0.542	0.570	**0.602**
	time	2.582	0.239	**0.207**	0.293	1.555	0.394	5.792
$\times 4$	PSNR	15.67	18.49	18.41	17.73	16.78	18.40	**18.86**
	SSIM	0.394	0.580	0.565	0.445	0.334	0.559	**0.597**
	time	0.572	0.133	**0.109**	0.123	0.320	0.138	0.238

on average a PSNR value which is 0.48 better than the two-step recovery strategy with upscaling $\times 2$, and 0.54 better with upscaling $\times 4$. For the average SSIM value, the improvements are 0.0335 and 0.0222 with upscaling $\times 2$ and $\times 4$, respectively. In addition, the paired t-tests results under significance 0.05 in Table 3 imply that our proposed approach statistically improves the performance in terms of both PSNR and SSIM. Investigating the results of different types of input, it is not surprising that the performance of face images with occlusions due to sunglasses and scarfs behave significantly worse than those which do not have such occlusions. Since significant details are missed due to the presence of occlusions, these images are very difficult to recover.

Some representative results are shown in Figures 5 and 6 with upscaling $\times 2$ and $\times 4$, respectively. The first column lists the raw test images with intrinsic variations, and in column (b) there are corresponding LR test inputs. Column (c) contains the HR images recovered by the proposed method. Finally, the images in the last column are the ground truth, i.e., the neutral expression face without illuminations and occlusions.

The detailed PSNR for 10 representative individuals with upscaling $\times 2$ and $\times 4$ are shown in Figures 7 and 8, respectively. Moreover, Figures 9 and 10 show the detailed SSIM values with upscaling $\times 2$ and $\times 4$, respectively.

More comparison results can be found in Table 4. The time consumption for every

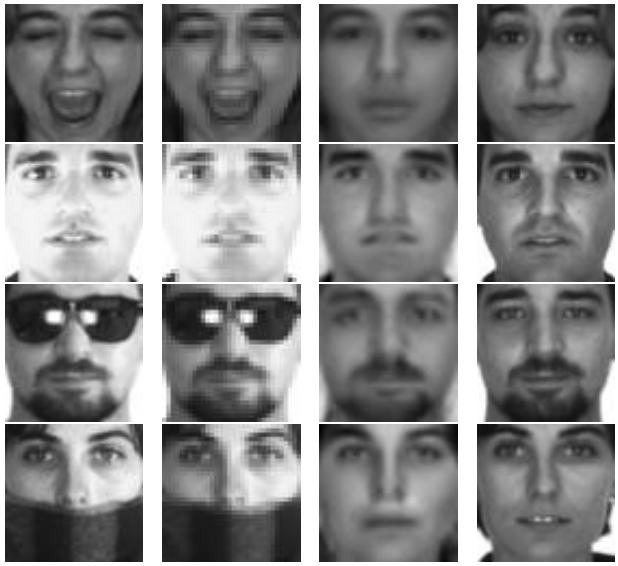
methods are evaluated on Intel E5-2643 (3.4 GHz). As stated above, it is clear that the proposed method performs much better than the two-step recovery strategy with various super-resolvers in terms of both PSNR and SSIM. On the one hand, the two-step recovery strategy optimizes parameters from albedo estimation and super resolution models sequentially, and thus the latter learned parameters (from super resolution part) can not benefit from the learning of former parameters (from albedo estimation part). In addition, although the learned parameters are locally optimal for both the albedo estimation and super resolution models individually, they are not optimal for the integrated model. In contrast, the proposed method jointly learns parameters from both albedo estimation and super resolution, in pursuit of a better result. The albedo estimation and super resolution components of the model can benefit each other during the optimization process. On the other hand, the ADMM method performs optimization iteratively. In each iteration, the parameters are optimized alternately. However, the two-step recovery strategy only optimizes the parameters once. From this point of view, the two-step recovery strategy can be seen as a special case of the proposed method which contains only one iteration during the optimization process.

4.1.3. Time Complexity

In this section, we evaluate the time complexity of the proposed method. As shown in Algorithm 1, each iteration consists of 3 subprocesses. Denoted by n_N, n_V the number of atoms in dictionaries $\mathbf{N}$ and $\mathbf{V}$, respectively, and n_c the number of patches from the LR images. Calculating α, β and c_{ij} takes a matrix multiplication operation with complexity $O(2(n_N + n_V + n_c)^2)$. Similarly, the complexities of the second and third subprocesses are $O(4(n_N + n_V + n_c))$ and $O(3(n_N + n_V + n_c))$, respectively. Thus the final time complexity of the proposed method is $O(T(2n^2 + 7n))$, where $n = n_N + n_V + n_c$ and T is the number of iterations (ignoring the cost of computing A^{-1}, which is done offline).

4.1.4. Convergence Analysis

To evaluate the convergence of the proposed method, we set the maximum number of iterations $\mathbf{T} = 200$ and calculate the value of the cost function (7) as well as the

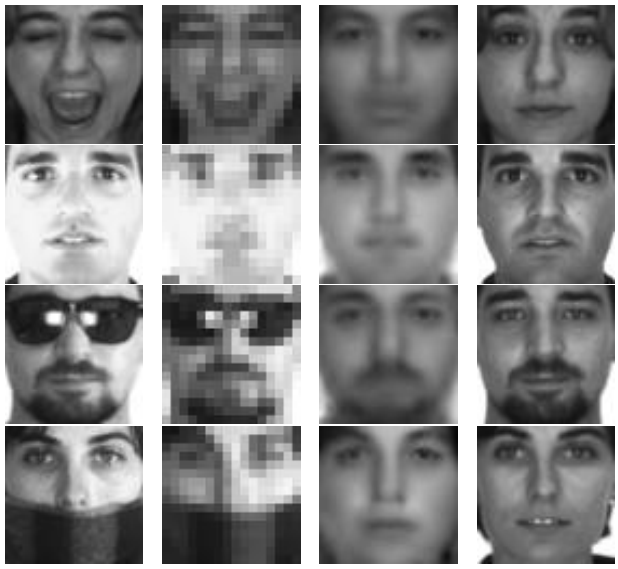


(a) (b) (c) (d)

Figure 5: Representative results with upscaling $\times 2$. (a) Raw test images with intra-personal variations. (b) Input LR images. (c) Images recovered by proposed method. (d) Ground truth.

PSNR and SSIM at each iteration. The average cost, PSNR and SSIM values for different iterations with upscaling $\times 2$ are shown in Figure 11(a). The average cost value falls sharply in the first 20 iterations. Then, the rate of decrease gradually slows. With the decreasing cost value, the average PSNR and SSIM values increase in the first 10-20 iterations and then become stable. These figures suggest that 20˜40 iterations would be enough to achieve desirable performance. More iterations beyond that can hardly bring significant improvement.

Figure 11(b) shows the average cost, PSNR and SSIM values at each iteration with upscaling $\times 4$. The average cost value also falls sharply in the first 20 iterations, which is similar to that with upscaling $\times 2$. However, the decrease does not slow down as quickly as is the case of $\times 2$ after 40 iterations. The increase of the PSNR and SSIM values is also more gentle compared to upscaling $\times 2$. These results imply that upscaling $\times 4$ will take more iterations to converge than upscaling $\times 2$.



(a) (b) (c) (d)

Figure 6: Representative results with upscaling $\times 4$. (a) Raw test images with intra-personal variations. (b) Input LR images. (c) Images recovered by proposed method. (d) Ground truth.

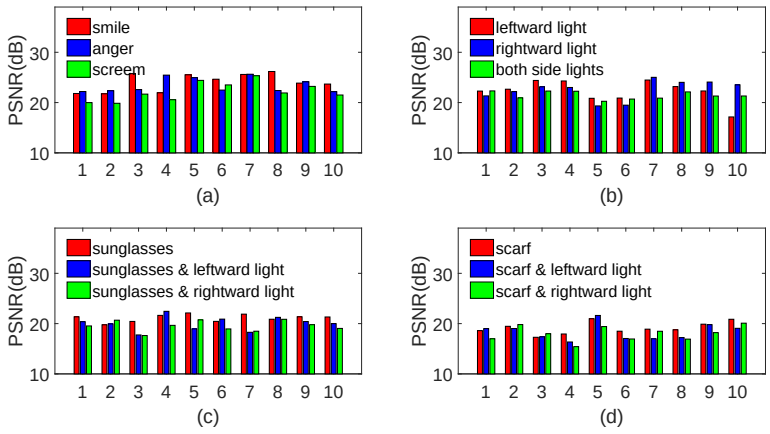


Figure 7: PSNR values of the hallucinated results from ten individuals with upscaling $\times 2$. (a) LR test inputs with expressions. (b) LR test inputs with illuminations. (c) LR test inputs with sun glasses & illuminations. (d) LR test inputs with scarf & illuminations.

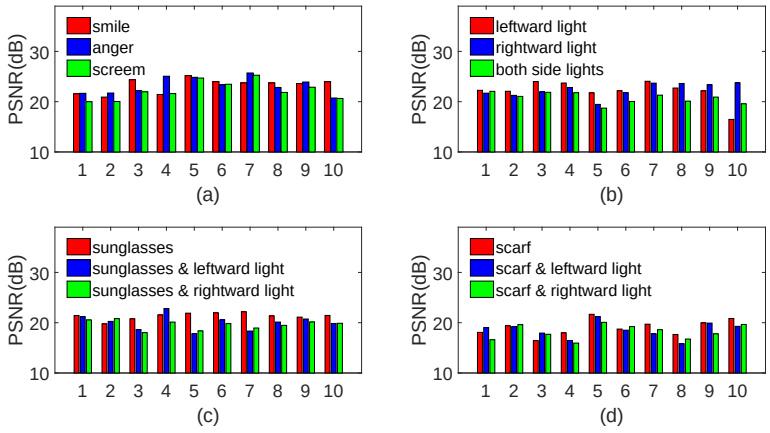


Figure 8: PSNR values of the hallucinated results from ten individuals with upscaling $\times 4$. (a) LR test inputs with expressions. (b) LR test inputs with illuminations. (c) LR test inputs with sun glasses & illuminations. (d) LR test inputs with scarf & illuminations.

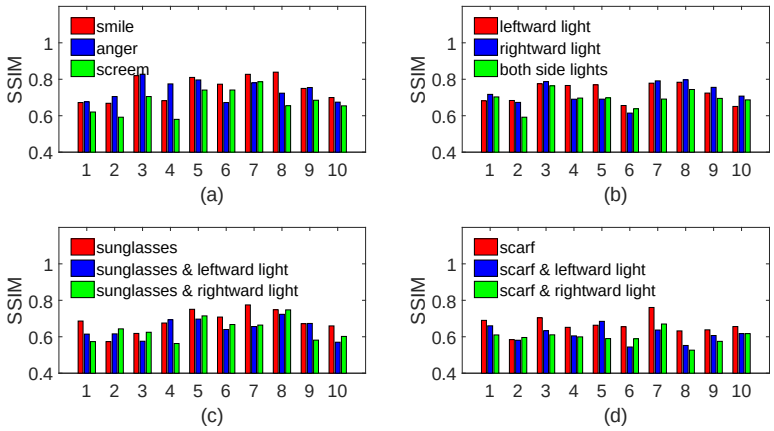


Figure 9: SSIM values of the hallucinated results from ten individuals with upscaling $\times 2$. (a) LR test inputs with expressions. (b) LR test inputs with illuminations. (c) LR test inputs with sun glasses & illuminations. (d) LR test inputs with scarf & illuminations.

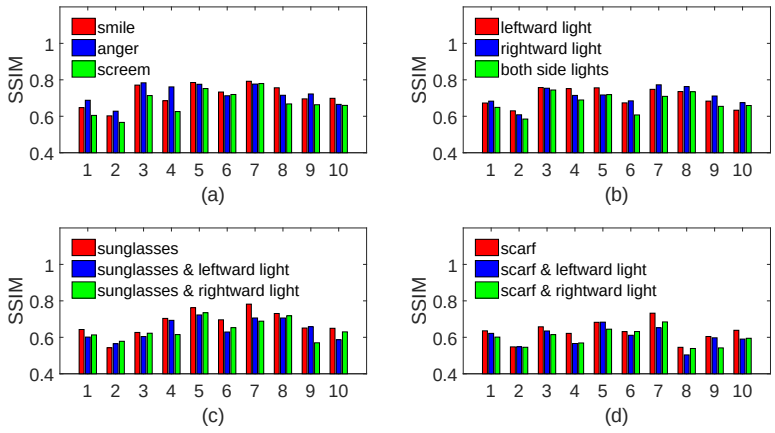


Figure 10: SSIM values of the hallucinated results from ten individuals with upscaling $\times 4$. (a) LR test inputs with expressions. (b) LR test inputs with illuminations. (c) LR test inputs with sun glasses & illuminations. (d) LR test inputs with scarf & illuminations.

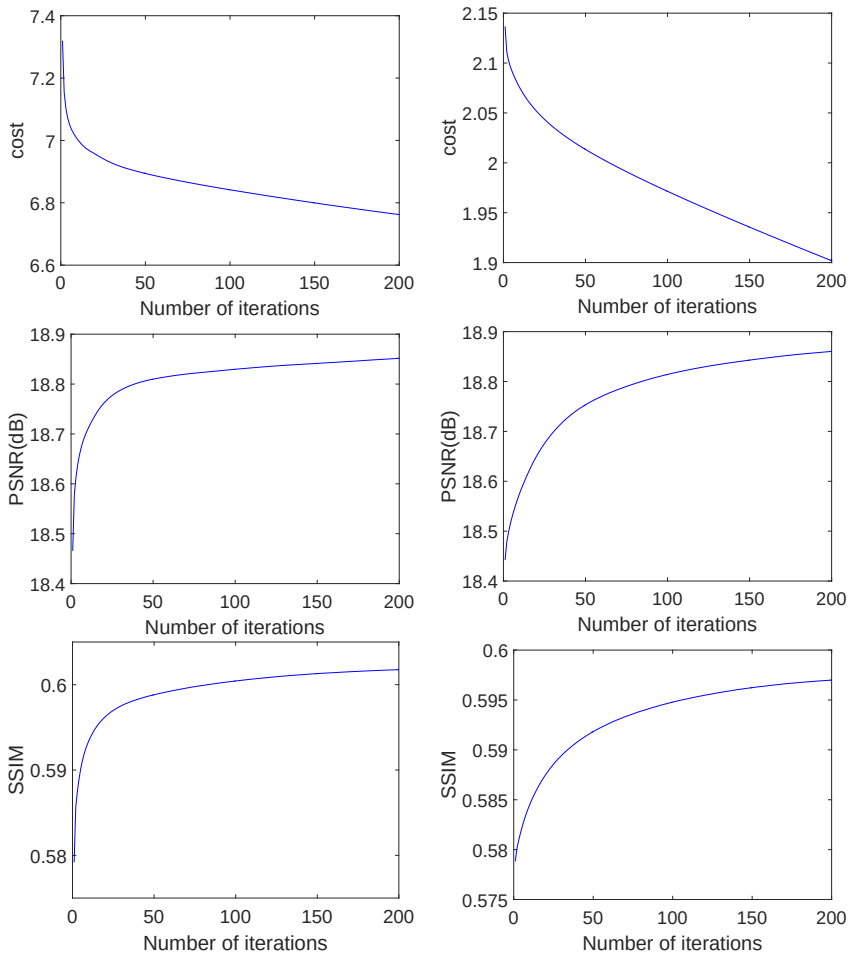
4.2. Face Recognition

In this section, we apply our method to face recognition. More specifically, we apply the same face recognition algorithm on two sets of faces. These two sets are the down-sampled low resolution faces (with size 32×32 for factor $\times 2$ or 16×16 for factor $\times 4$) from the AR database and the corresponding high resolution faces recovered by the proposed method. To make a comparison, the low resolution faces are interpolated and up-sampled to size 64×64, i.e., the size of the recovered high resolution faces.

The 10-fold cross-validation strategy is adopted to evaluate the performance of face recognition. In each fold, face features are extracted from the training set using Fisherfaces method. To be specific, images are first projected into a lower dimensional space using PCA algorithm. Then LDA algorithm is performed to extract the features [34]. [1] The extracted features are used to train a Support Vector Machine (SVM) model.[2] The trained model is then used to classify images from test set.

[1] We use the matlab implementations of PCA and LDA provided by Deng Cai, which is available at http://www.cad.zju.edu.cn/home/dengcai/Data/DimensionReduction.html.

[2] We choose LibSVM [35] for training and classifying. Software available at



(a) $\times 2$ (b) $\times 4$

Figure 11: The average cost (top), PSNR (middle) and SSIM (bottom) values per iteration with upscaling $\times 2$ (a) and $\times 4$ (b).

Table 5: The correct rate of face recognition on raw and recovered images.

	raw images	recovered images
$\times 2$	52.92%	**76.77**%
$\times 4$	20.69%	**68.54**%

Table 5 shows The correct rate of face recognition for raw and recovered images, the bold numbers signify the best performance. It is clear that the raw low resolution images with various expressions, illuminations and occlusions are hard to recognize, especially with a large down-sampling factor ($\times 4$). In contrast, the proposed method successfully recovers images from low resolution as well as to a large extent eliminating the influence of variations, what leads to a significant improvement in recognition performance.

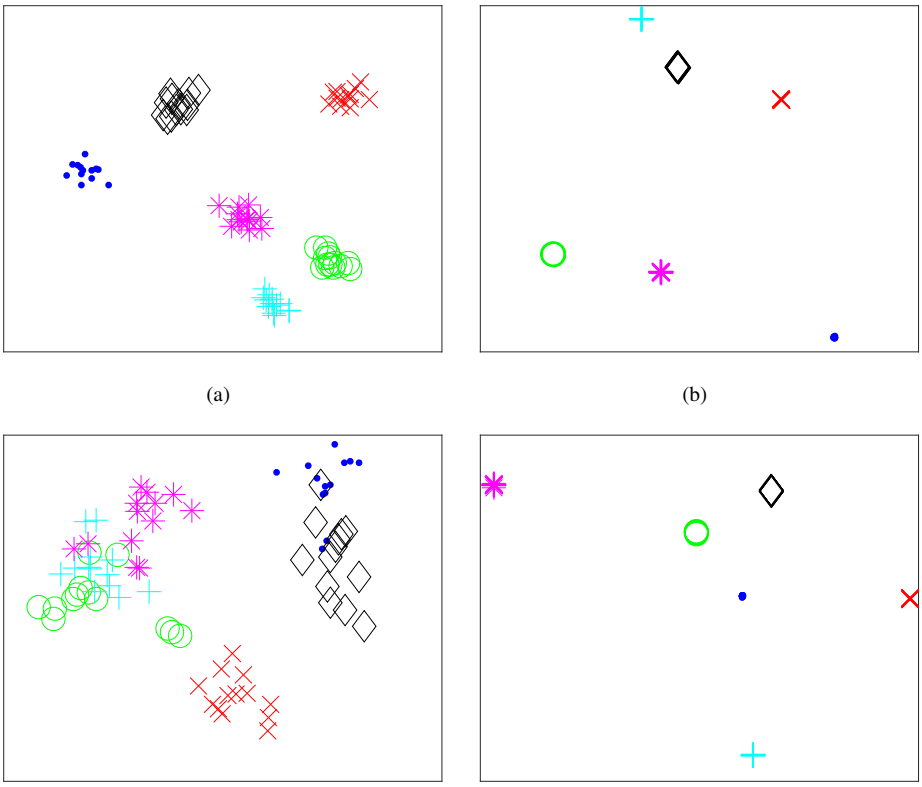
4.3. Clustering

Finally, we demonstrate that our method can also improve significantly the performance of face clustering. The raw low resolution faces and corresponding recovered faces, which are the same as the sets used in the previous face recognition experiment, are used in this experiment. Similarly, features are extracted by using Fisherfaces method. Then, the k-means method is applied to cluster faces using extracted features. The clustering performance is evaluated by comparing the obtained label of each face with ground truth (the identities of faces) in terms of two metrics, the accuracy (AC)[36] and the normalized mutual information metric (MI)[36].

For visualization, we randomly select 6 classes and plot the samples from these classes in a 2D plane. As illustrated in Figure 12, features extracted directly from low resolution images with variations are somewhat hard to differentiate with a large down-sampling factor ($\times 4$). In contrast, features extracted from the recovered faces almost overlap according to the their classes, which is obviously easy to cluster.

The quantitative results are shown in Table 6. The experimental results show that applying clustering method on recovered images performs much better than on raw

http://www.csie.ntu.edu.tw/ cjlin/libsvm/.



(a) (b)

(c) (d)

Figure 12: Samples of raw low resolution faces with down-sampling factor $\times 2$ (a) and $\times 4$ (c). And samples of corresponding recovered faces with down-sampling factor $\times 2$ (b) and $\times 4$ (d).

Table 6: Performance of clustering on raw and recovered images.

	raw images		recovered images	
	AC	MI	AC	MI
$\times 2$	0.851	0.958	**0.954**	**0.987**
$\times 4$	0.771	0.941	**0.943**	**0.984**

images, which demonstrates that the proposed method can also boost the clustering performance.

5. Conclusions and Discussions

The recovery of neutral HR face images (i.e., without any expression, illumination and occlusion) from a LR face image with various expressions, illuminations and occlusions is a challenging problem. The presence of intrinsic facial image variations can cause serious loss of facial details, which is a critical problem in super resolution image processing. In this paper, we have proposed a novel face enhancement framework that jointly estimates facial albedo and performs face super resolution. To the best of our knowledge, this is the first attempt to simultaneously handle these two problems in a single integrated process. Experimental results show that the proposed method achieves better performance than a simple two-step processing strategy (i.e., performing albedo estimation and super resolution sequentially) in terms of both PSNR and SSIM. Moreover, the proposed method can also significantly improve the performance of face recognition and clustering.

However, many future works can be investigated. One major drawback of the proposed method is the time complexity. Since it relies on $l1$-norm regularization terms, the optimization procedure can only be performed iteratively, which brings with it significant demands in terms of time consumption. In super resolution, A+[33] solves this problem by dividing training examples into hundreds of compact clusters. Since the samples in each cluster are densely packed in the search space, the $l1$-norm can be replaced by the $l2$-norm. This means that there is a closed form optimal solution. Thus a projection matrix from LR space to HR space can be learned in the training

phase. In the testing phase, the HR patches can be recovered by directly multiplying the precomputed projection matrix with the input LR patches, which makes it fairly efficient. In fact A+ achieves excellent performance in terms of both image quality and time efficiency. Unfortunately, this strategy can not be used in our case. On the one hand, the input LR images contain a variety of expressions, illuminations and occlusions. Thus it is impossible to ensure which cluster the patches should belonging to before estimating albedo. On the other hand, the albedo estimation should be performed over the entire face image, rather than on the individual patches as A+ does. However, the proposed method still outperforms the two-step recovery strategy, which applies albedo estimation and super resolution (A+) sequentially.

Acknowledgment

This work is supported by National Natural Science Foundation of China (Grant No.61402389) and the Fundamental Research Funds for the Central Universities (No. 20720160073).

References

[1] W. Chen, M. J. Er, S. Wu, Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain, IEEE Trans. on Systems, Man, and Cybernetics, B 36 (2) (2006) 458–466.

[2] S. Du, R. Ward, Wavelet-based illumination normalization for face recognition, in: IEEE International Conference on Image Processing, Vol. 2, IEEE, 2005, pp. II–954.

[3] V. Blanz, T. Vetter, Face recognition based on fitting a 3d morphable model, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (9) (2003) 1063–1074.

[4] O. Aldrian, W. A. Smith, Inverse rendering of faces with a 3d morphable model, IEEE Transactions on Pattern Analysis and Machine Intelligence. 35 (5) (2013) 1080–1093.

[5] X. Zhu, Z. Lei, J. Yan, D. Yi, S. Z. Li, High-fidelity pose and expression normalization for face recognition in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 787–796.

[6] 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 (2) (2009) 210–227.

[7] S. Baker, T. Kanade, Hallucinating faces, in: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition., 2000, pp. 83–88.

[8] J. Yang, H. Tang, Y. Ma, T. Huang, Face hallucination via sparse coding, in: 15th IEEE International Conference on Image Processing, 2008, pp. 1264–1267.

[9] X. Ma, J. Zhang, C. Qi, Hallucinating face by position-patch, Pattern Recognition 43 (6) (2010) 2224–2236.

[10] X. Ma, H. Q. Luong, W. Philips, H. Song, H. Cui, Sparse representation and position prior based face hallucination upon classified over-complete dictionaries, Signal processing 92 (9) (2012) 2066–2074.

[11] C. Huang, Y. Liang, X. Ding, C. Fang, Generalized joint kernel regression and adaptive dictionary learning for single-image super-resolution, Signal processing 103 (2014) 142–154.

[12] C. Jung, L. Jiao, B. Liu, M. Gong, Position-patch based face hallucination using convex optimization, IEEE Signal Processing Letters 18 (6) (2011) 367–370.

[13] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, L. Zhang, Convolutional sparse coding for image super-resolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1823–1831.

[14] O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition 46 (1) (2013) 9–23.

[15] W. Kim, S. Suh, W. Hwang, J.-J. Han, Svd face: illumination-invariant face representation, IEEE Signal Processing Letters 21 (11) (2014) 1336–1340.

[16] H. Hu, Illumination invariant face recognition based on dual-tree complex wavelet transform, IET Computer Vision 9 (2) (2014) 163–173.

[17] A. G. Pai, S. L. Fernandes, K. Nayak, K. Accamma, K. Sushmitha, K. Kumari, et al., Recognizing human faces under varying degree of illumination: A comprehensive survey, in: Electronics and Communication Systems (ICECS), 2015 2nd International Conference on, IEEE, 2015, pp. 577–582.

[18] C. Liu, H.-Y. Shum, C.-S. Zhang, A two-step approach to hallucinating faces: global parametric model and local nonparametric model, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition., Vol. 1, 2001, pp. I–192.

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

[20] O. Arandjelović, Hallucinating optimal high-dimensional subspaces, Pattern Recognition 47 (8) (2014) 2662–2672.

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

[22] M. Aharon, M. Elad, A. Bruckstein, The k-svd: An algorithm for designing of overcomplete dictionaries for sparse representation, IEEE transactions on Signal Processing 54 (11) (2006) 4311–4322.

[23] L. Chang, M. Zhou, Y. Han, X. Deng, Face sketch synthesis via sparse representation, in: 20th International Conference on Pattern Recognition, 2010, pp. 2146–2149.

[24] D. Chen, X. Cao, L. Wang, F. Wen, J. Sun, Bayesian face revisited: A joint formulation, in: European Conference on Computer Vision, 2012, pp. 566–579.

[25] W. Deng, J. Hu, J. Guo, Extended src: Undersampled face recognition via intraclass variant dictionary, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (9) (2012) 1864–1870.

[26] R. Zeyde, M. Elad, M. Protter, On single image scale-up using sparse-representations, in: International conference on curves and surfaces, 2010, pp. 711–730.

[27] M. Aharon, M. Elad, A. Bruckstein, K-svd: An algorithm for designing overcomplete dictionaries for sparse representation, IEEE Transactions on Signal Processing 54 (11) (2006) 4311–4322.

[28] J. Wright, A. Ganesh, S. Rao, Y. G. Peng, Y. Ma, Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization., in: Advances in Neural Information Processing Systems 22: Conference on Neural Information Processing Systems 2009. Proceedings of A Meeting Held 7-10 December 2009, Vancouver, British Columbia, Canada, 2009, p. 20:320:56.

[29] A. M. Martinez, The ar face database, CVC Technical Report 24.

[30] P. J. Phillips, H. Moon, S. A. Rizvi, P. J. Rauss, The feret evaluation methodology for face-recognition algorithms, IEEE Transactions on pattern analysis and machine intelligence 22 (10) (2000) 1090–1104.

[31] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, D. Zhao, The cas-peal large-scale chinese face database and baseline evaluations, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 38 (1) (2008) 149–161.

[32] R. Timofte, V. De Smet, L. Van Gool, Anchored neighborhood regression for fast example-based super-resolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1920–1927.

[33] R. Timofte, V. De Smet, L. Van Gool, A+: Adjusted anchored neighborhood regression for fast super-resolution, in: Asian Conference on Computer Vision, Springer, 2014, pp. 111–126.

[34] P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, IEEE Transactions on pattern analysis and machine intelligence 19 (7) (1997) 711–720.

[35] C. C. Chang, C. J. Lin, Libsvm: A library for support vector machines, "Acm Transactions on Intelligent Systems & Technology" 2 (3) (2011) 27.

[36] D. Cai, X. He, J. Han, Document clustering using locality preserving indexing, "IEEE Transactions on Knowledge & Data Engineering" 17 (12) (2005) 1624–1637.