

© 2010 Jianchao Yang

IMAGE SUPER-RESOLUTION VIA SPARSE REPRESENTATION

BY

JIANCHAO YANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Adviser:

Professor Thomas S. Huang

ABSTRACT

This thesis presents a new approach to single-image super-resolution (SR), based on sparse signal recovery. Research on image statistics suggests that image patches can be well represented as a sparse linear combination of elements from an appropriately chosen over-complete dictionary. Inspired by this observation, we seek a sparse representation for each patch of the low-resolution input, and then use the coefficients of this representation to generate the high-resolution output. Theoretical results from compressed sensing suggest that under mild conditions, the sparse representation can be correctly recovered from the downsampled signals. By jointly training two dictionaries for the low- and high-resolution image patches, we can enforce the similarity of sparse representations between the low- and high-resolution image patch pairs with respect to their own dictionaries. Therefore, the sparse representation of a low-resolution image patch can be applied with the dictionary of high-resolution image patches to generate a high-resolution image patch. Compared to previous approaches, which simply sample a large amount of raw image patch pairs, the learned dictionary pair is a more compact representation of the patch pairs, and, therefore, reduces the computation cost substantially. The effectiveness of such a sparsity prior is demonstrated on both general image super-resolution and the special case of face hallucination. In both cases, our algorithm can generate high-resolution images that are competitive or superior in quality to images produced by other similar SR methods, but with much faster processing speed.

ACKNOWLEDGMENTS

I would like to thank my adviser, Professor Thomas S. Huang, for his advice, support, and intellectual guidance. I would also like to thank Professor Yi Ma and Professor Xu Han (University of Missouri) for discussions of the techniques presented in this thesis. Also, grateful thanks to John Wright and Hao Tang for having discussions about the approach and helping on experiments.

Last, but not least, neither the research work summarized in this thesis nor anything else that I have ever achieved in my life would be possible without the unconditional love and steadfast support from my family in China.

TABLE OF CONTENTS

LIST OF TABLES	vi
LIST OF FIGURES	vii
LIST OF SYMBOLS	ix
CHAPTER 1 INTRODUCTION	1
1.1 What Is Super-resolution?	1
1.2 Super-resolution Techniques	2
1.3 Signal Recovery Based on Sparse Representation	4
1.4 Organization of This Thesis	7
1.5 Notations	7
CHAPTER 2 SUPER-RESOLUTION VIA SPARSE REPRESENTATION	9
2.1 Super-resolution Constraints	9
2.1.1 Reconstruction constraint	9
2.1.2 Sparsity prior	10
2.2 Generic Image Super-resolution from Sparsity	11
2.2.1 Local model from sparse representation	11
2.2.2 Enforcing global reconstruction constraint	13
2.2.3 Global optimization interpretation	15
2.3 Face Super-resolution from Sparsity	16
2.3.1 Nonnegative matrix factorization modeling	17
2.3.2 Two step face super-resolution	18
CHAPTER 3 LEARNING THE DICTIONARY PAIR	20
3.1 Single Dictionary Training	20
3.2 Joint Dictionary Training	22
3.3 Feature Representation for Low-resolution Image Patches	23
CHAPTER 4 EXPERIMENT EVALUATION AND ANALYSIS	26
4.1 Single Image Super-resolution	27
4.1.1 Generic image super-resolution	27
4.1.2 Face super-resolution	29

4.2	Effects of Dictionary Size	31
4.3	Robustness to Noise	33
4.4	Effects of Global Constraints	35
CHAPTER 5 CONCLUSION		38
REFERENCES		39

LIST OF TABLES

4.1	The RMSEs of the reconstructed images using dictionaries of different sizes, and using the raw image patches directly from which the dictionaries are trained.	32
4.2	The RMSEs of the reconstructed images from different levels of noisy inputs.	34
4.3	The global constraint in the second step further refines the results from local sparse model in the first step and reduces RMSEs.	37

LIST OF FIGURES

1.1	Reconstruction of a feline face with magnification factor 2. Left: result by our method. Right: the original image. There is little noticeable difference.	6
3.1	The high-resolution image patch dictionary trained by Eq. (3.8) using 100,000 high- and low-resolution image patch pairs sampled from the generic training images. A total of 512 dictionary atoms are learned with each atom of size 9×9	24
4.1	The flower and girl image magnified by a factor of three. Left to right: input, bicubic interpolation, neighbor embedding, our method, and the original.	28
4.2	Results on an image of the Parthenon with magnification factor three. Top row: low-resolution input, bicubic interpolation, back projection. Bottom row: neighbor embedding, soft edge prior, and our method.	28
4.3	Example training faces for the face super-resolution algorithm. The training images include faces of both genders, different ages, different races, and various facial expressions.	29
4.4	The comparison between the two-step face hallucination algorithm with the generic image super-resolution algorithm applied to low-resolution face images. From left to right: input image, super-resolution result using the two-step approach, and super-resolution result using the generic approach.	30
4.5	Results of our algorithm compared to other methods. From left to right columns: (a) low-resolution input; (b) bicubic interpolation; (c) back-projection; (c) global model via NMF followed by bilateral filtering; (d) global model combined with local model via sparse representation; (f) original.	31
4.6	The effects of dictionary size on the super-resolution reconstruction of Lena. From left to right: dictionary size 256, 512, 1024, and 2048 and the whole sampled patch set.	32
4.7	The computation time on “Girl” image with dictionaries of different sizes (in seconds).	33

4.8	The effects of λ on the recovered image given the input. From left to right, $\lambda = 0.01, 0.05, 0.1, 0.2, 0.3$. The larger λ is, the smoother the result image gets. Note that the results are generated from the local model only.	35
4.9	Performance evaluation of our proposed algorithm on noisy data. Noise level (standard deviation of Gaussian noise) from left to right columns: 0, 4, 6 and 8. Top row: input images. Middle row: recovered images using NE ($k = 13, 12, 9, 7$). Bottom row: recovered images using our method ($\lambda = 0.1, 0.4, 0.6, 0.8$).	36

LIST OF SYMBOLS

X	High-resolution image
Y	Low-resolution image
x	High-resolution image patch
y	Low-resolution image patch
D_h	Dictionary for high-resolution image patches
D_l	Dictionary for low-resolution image patches
D	Downsampling operator in matrix form

CHAPTER 1

INTRODUCTION

1.1 What Is Super-resolution?

In most digital imaging applications, high-resolution images or videos are usually desired for image processing and analysis. Image resolution describes the details contained in an image; the higher the resolution, the more details can be captured. The image resolution is limited by the imaging system, such as image sensors (e.g. CCD) and optics. Constructing imaging chips and optical components to capture very high-resolution images is prohibitively expensive and not practical in most real applications. On the other hand, it is always desirable to enhance the images or video captured already as the legacy from the development of digital imaging.

Another way to address the resolution problem is to use signal processing or machine learning techniques to post-process the captured images. These techniques are specifically referred as super-resolution (SR) reconstruction. Super-resolution image reconstruction has been an active research area since it was originally proposed, because it offers the promise of overcoming some of the inherent resolution limitations of the imaging system and improving the performance of many image processing applications. It is especially helpful in many practical applications:

- Medical imaging: several images limited in resolution quality are captured, and SR techniques can be applied to enhance the resolution.

- Remote sensing: several images of the same area are provided, and an improved resolution image can be sought.
- Surveillance video: frame freeze and zoom regions of interest in videos for human perception and machine recognition.
- Video standard conversion: e.g. from NTSC video to HDTV signals.

1.2 Super-resolution Techniques

Conventional approaches to generating a super-resolution (SR) image require multiple low-resolution images of the same scene, which are aligned with sub-pixel accuracy. The SR task is cast as the inverse problem of recovering the original high-resolution image by fusing the low-resolution images, on the basis of reasonable assumptions or prior knowledge about the observation or generation model from the high-resolution image to the low-resolution images. The fundamental reconstruction constraint for SR is that applying the image generation model to the recovered image should produce the same low-resolution images as observed. However, the SR image reconstruction approach is generally a severely ill-posed problem because of the insufficient number of low-resolution images, ill-conditioned registration, and unknown blurring operators, and because the solution from the reconstruction constraint is not unique. Various regularization methods were proposed to further stabilize the inversion of such an ill-posed problem, such as [1–3]. However, the performance of these reconstruction-based super-resolution algorithms degrades rapidly when the desired magnification factor is large or the available input images are limited. In these cases, the results may be overly smooth, lacking important high-frequency details [4]. Another class of SR approach is based on interpolation [5–7]. While simple interpolation

methods such as bilinear or bicubic interpolation tend to generate overly smooth images with ringing and jagged artifacts, interpolation by exploiting the natural image priors will generally produce more favorable results. Dai et al. [6] represented the local image patches using the background/foreground descriptors and reconstructed the sharp discontinuity between the two. Sun et al. [7] explored the gradient profile prior for local image structures and applied it to super-resolution. Such approaches are effective in preserving the edges in the zoomed image. However, they are limited in modeling the visual complexity of the real images. For natural images with fine textures or smooth shading, these approaches tend to produce watercolor-like artifacts.

A third category of SR approach is based on machine learning techniques, which attempt to capture the co-occurrence prior between low-resolution and high-resolution image patches. Freeman et al. [8] proposed an example-based learning strategy that applies to generic images where the low-resolution to high-resolution prediction is learned via a Markov random field (MRF) solved by belief propagation. Sun et al. [9] extends this approach by using the primal sketch priors to enhance blurred edges, ridges, and corners. Nevertheless, the above methods typically require enormous databases of millions of high-resolution and low-resolution patch pairs to make the databases expressive enough, and are, therefore, computationally intensive. Chang et al. [10] adopt the philosophy of LLE [11] from manifold learning, assuming similarity between the two manifolds in the high-resolution patch space and the low-resolution patch space. The algorithm in [10] maps the local geometry of the low-resolution patch space to the high-resolution patch space, generating high-resolution patch as a linear combination of neighbors. Using this strategy, more patch patterns can be represented using a smaller training database. However, using a fixed number of K neighbors for reconstruction often results in blurring effects, due to over- or under-fitting.

While the approaches mentioned above were proposed for generic image super-resolution, specific image priors can be incorporated when tailed to SR applications for specific domains such as human faces. Baker and Kanade [12] started the pioneering work on face hallucination. However, the gradient pyramid-based prediction does not model the face prior, and the pixels are predicted individually, causing discontinuity and artifacts. C. Liu et al. [13] proposed a two-step statistical approach integrating the global PCA model and a local patch model. Although the algorithm yields good results, it uses the holistic PCA model, which tends to render results similar to the mean face, and the probabilistic local patch model is also complicated and computationally demanding. W. Liu et al. [14] proposed a new approach based on tensor patch and residue compensation. While this algorithm adds more details to the face, it also introduces more artifacts.

1.3 Signal Recovery Based on Sparse Representation

This thesis focuses on the problem of recovering the super-resolution version of a given low-resolution image. Similarly to the aforementioned learning-based methods, we will rely on patches from the input image. However, instead of working directly with the image patch pairs sampled from high- and low-resolution images [15], we learn a compact representation of these patch pairs to capture the co-occurrence prior, significantly improving the speed of the algorithm. Our approach is motivated by recent results in sparse signal representation, which suggest that the linear relationships among high-resolution signals can be accurately recovered from their low-dimensional projections [16, 17]. Although the super-resolution problem is very ill-posed, making precise recovery impossible, the image patch sparse representation demonstrates both effectiveness and robustness in regularizing the inverse problem.

To be more precise, let $\mathbf{D} \in \mathbb{R}^{n \times K}$ be an over-complete dictionary of K bases, and suppose a signal $\mathbf{x} \in \mathbb{R}^n$ can be represented as a sparse linear combination with respect to \mathbf{D} . That is, the signal \mathbf{x} can be written as $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}_0$ where $\boldsymbol{\alpha}_0 \in \mathbb{R}^K$ is a vector with very few ($\ll K$) nonzero entries. In practice, we might only observe a small set of measurements \mathbf{y} of \mathbf{x} :

$$\mathbf{y} \doteq L\mathbf{x} = L\mathbf{D}\boldsymbol{\alpha}_0, \quad (1.1)$$

where $L \in \mathbb{R}^{k \times n}$ with $k < n$ is a projection matrix. In our super-resolution context, \mathbf{x} is a high-resolution image (patch), while \mathbf{y} is its low-resolution counterpart (or features extracted from it). If the dictionary \mathbf{D} is overcomplete, the equation $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$ is underdetermined for the unknown coefficients $\boldsymbol{\alpha}$. The equation $\mathbf{y} = L\mathbf{D}\boldsymbol{\alpha}$ is even more dramatically underdetermined. Nevertheless, under mild conditions, the sparsest solution $\boldsymbol{\alpha}_0$ to this equation will be unique. Furthermore, if \mathbf{D} satisfies an appropriate near-isometry condition, then for a wide variety of matrices L , any sufficiently sparse linear representation of a high-resolution image patch \mathbf{x} in terms of the \mathbf{D} can be recovered (almost) perfectly from the low-resolution image patch [17, 18]. Figure 1.1 shows an example that demonstrates the capabilities of our method derived from this principle. The image of the feline face is blurred and downsampled to half of the original size. And then we zoom the image to the original size using our method. Even for such a complicated texture, sparse representation recovers a visually appealing reconstruction of the original signal.

Recently, sparse representation has been successfully applied to many other related inverse problems in image processing, such as denoising [19] and restoration [20], often improving on the state-of-the-art. For example, in [19], the authors use the K-SVD algorithm [21] to learn an overcomplete dictionary from natural



Figure 1.1: Reconstruction of a feline face with magnification factor 2. Left: result by our method. Right: the original image. There is little noticeable difference.

image patches and successfully apply it to the image denoising problem. In our setting, we do not directly compute the sparse representation of the high-resolution patch. Instead, we will work with two coupled dictionaries, \mathbf{D}_h for high-resolution patches, and \mathbf{D}_ℓ for low-resolution patches. The sparse representation of a low-resolution patch in terms of \mathbf{D}_ℓ will be directly used to recover the corresponding high-resolution patch from \mathbf{D}_h . We obtain a locally consistent solution by allowing patches to overlap and demanding that the reconstructed high-resolution patches agree on the overlapped areas. Unlike the K-SVD algorithm, we try to learn the two overcomplete dictionaries in a probabilistic model similar to [22]. To enforce that the image patch pairs have the same sparse representations with respect to \mathbf{D}_h and \mathbf{D}_ℓ , we learn the two dictionaries simultaneously by concatenating them with normalization. The learned compact dictionaries will be applied to both generic image super-resolution and face hallucination to demonstrate its effectiveness.

Compared to the aforementioned patch-based methods, our algorithm requires only two compact learned dictionaries, instead of a large training patch database. The computation, mainly based on linear programming or convex optimization, is much more efficient and scalable, compared with [8–10]. The online recovery of the

sparse representation uses the low-resolution dictionary, and the high-resolution dictionary is used only to calculate the final high-resolution image. The computed sparse representation adaptively selects the most relevant patch bases in the dictionary to best represent each patch of the given low-resolution image. This leads to superior performance, both qualitatively and quantitatively, and generates sharper edges and clearer textures, compared to methods [10] that use a fixed number of nearest neighbors. In addition, the sparse representation is robust to noise as suggested in [19]; and, thus, our algorithm is more robust to noise in the test image, while other methods cannot perform denoising and super-resolution simultaneously.

1.4 Organization of This Thesis

The remainder of this thesis is organized as follows. Chapter 2 details our formulation and solution to the image super-resolution problem based on sparse representation. Specifically, we study how to apply sparse representation for both generic image super-resolution and face hallucination. In Chapter 3, we discuss how to learn the two dictionaries for the high-resolution and low-resolution image patches. Various experimental results in Chapter 4 demonstrate the efficacy of sparsity as a prior for regularizing image super-resolution.

1.5 Notations

Specifically, \mathbf{X} and \mathbf{Y} denote the high- and low-resolution image, respectively, and \mathbf{x} and \mathbf{y} denote the high- and low-resolution image patch, respectively. We use bold uppercase \mathbf{D} to denote the dictionary for sparse coding; especially, we use \mathbf{D}_h and \mathbf{D}_ℓ to denote the dictionaries for high- and low-resolution image

patches, respectively. Bold lowercase letters denote vectors. Unbold uppercase letters denote regular matrices; i.e., D is used as a downsampling operation in matrix form. Unbold lowercase letters are used as scalars.

CHAPTER 2

SUPER-RESOLUTION VIA SPARSE REPRESENTATION

2.1 Super-resolution Constraints

The single-image super-resolution problem asks: Given a low-resolution image \mathbf{Y} , recover a higher-resolution image \mathbf{X} of the same scene. Two constraints are modeled in this work to solve this ill-posed problem: 1) reconstruction constraint, which requires that the recovered \mathbf{X} should be consistent with the input \mathbf{Y} with respect to the image observation model; and 2) sparsity prior, which assumes that the high-resolution patches can be sparsely represented in an appropriately chosen overcomplete dictionary, and that their sparse representations can be recovered from the low-resolution observation.

2.1.1 Reconstruction constraint

The observed low-resolution image \mathbf{Y} is a blurred and downsampled version of the high-resolution image \mathbf{X} :

$$\mathbf{Y} = SH\mathbf{X}. \quad (2.1)$$

Here, H represents a blurring filter, and S represents the downsampling operator.

Super-resolution remains extremely ill-posed, since for a given low-resolution input \mathbf{Y} , infinitely many high-resolution images \mathbf{X} satisfy the above reconstruction constraint. We further regularize the problem via the following prior on small patches \mathbf{x} of \mathbf{X} .

2.1.2 Sparsity prior

The patches \mathbf{x} of the high-resolution image \mathbf{X} can be represented as a sparse linear combination in a dictionary \mathbf{D}_h trained from high-resolution patches sampled from training images¹:

$$\mathbf{x} \approx \mathbf{D}_h \boldsymbol{\alpha} \quad \text{for some } \boldsymbol{\alpha} \in \mathbb{R}^K \text{ with } \|\boldsymbol{\alpha}\|_0 \ll K. \quad (2.2)$$

The sparse representation $\boldsymbol{\alpha}$ will be recovered by representing patches \mathbf{y} of the input image \mathbf{Y} , with respect to a low-resolution dictionary \mathbf{D}_l co-trained with \mathbf{D}_h . The dictionary training process will be discussed in Chapter 3.

We apply our approach to both generic images and face images. For generic image super-resolution, we divide the problem into two steps. First, as suggested by the sparsity prior Eq. (2.2), we find the sparse representation for each local patch, respecting spatial compatibility between neighbors. Next, using the result from this local sparse representation, we further regularize and refine the entire image using the reconstruction constraint Eq. (2.1). In this strategy, a local model from the sparsity prior is used to recover lost high-frequency for local details. The global model from the reconstruction constraint is then applied to remove possible artifacts from the first step and make the image more consistent and natural. The face images differ from the generic images in that the face images have more regular structure and thus reconstruction constraints in the face subspace can be more effective. For face image super-resolution, we reverse the above two steps to make better use of the global face structure as a regularizer. We first find a suitable subspace for human faces, and apply the reconstruction constraints to recover a medium-resolution image. We then recover the local details using the sparsity prior for image patches.

¹Similar mechanisms – sparse coding with an overcomplete dictionary – are also believed to be employed by the human visual system [23].

The remainder of this section is organized as follows: in Section 2.2, we discuss super-resolution for generic images. We will introduce the local model based on sparse representation and global model based on reconstruction constraints. In Section 2.3 we discuss how to introduce the global face structure into this framework to achieve more accurate and visually appealing super-resolution for face images.

2.2 Generic Image Super-resolution from Sparsity

2.2.1 Local model from sparse representation

Similarly to the patch-based methods mentioned previously, our algorithm tries to infer the high-resolution image patch for each low-resolution image patch from the input. For this local model, we have two dictionaries \mathbf{D}_h and \mathbf{D}_l , which are trained to have the same sparse representations for each high-resolution and low-resolution image patch pair. We subtract the mean pixel value for each patch, so that the dictionary represents image textures rather than absolute intensities. In the recovery process, the mean value for each high-resolution image patch is then predicted by its low-resolution version.

For each input low-resolution patch \mathbf{y} , we find a sparse representation with respect to \mathbf{D}_l . The corresponding high-resolution patch bases \mathbf{D}_h will be combined according to these coefficients to generate the output high-resolution patch \mathbf{x} . The problem of finding the sparsest representation of \mathbf{y} can be formulated as

$$\min \|\boldsymbol{\alpha}\|_0 \quad \text{s.t.} \quad \|F\mathbf{D}_l\boldsymbol{\alpha} - F\mathbf{y}\|_2^2 \leq \epsilon, \quad (2.3)$$

where F is a (linear) feature extraction operator. The main role of F in Eq. (2.3)

is to provide a perceptually meaningful constraint² on how closely the coefficients $\boldsymbol{\alpha}$ must approximate \mathbf{y} . We will discuss the choice of F in Section 3.3.

Although the optimization problem Eq. (2.3) is NP-hard in general, recent results [24, 25] suggest that as long as the desired coefficients $\boldsymbol{\alpha}$ are sufficiently sparse, they can be efficiently recovered by instead minimizing the ℓ^1 -norm,³ as follows:

$$\min \|\boldsymbol{\alpha}\|_1 \quad \text{s.t.} \quad \|F\mathbf{D}_l\boldsymbol{\alpha} - F\mathbf{y}\|_2^2 \leq \epsilon. \quad (2.4)$$

Lagrange multipliers offer an equivalent formulation,

$$\min_{\boldsymbol{\alpha}} \|F\mathbf{D}_l\boldsymbol{\alpha} - F\mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \quad (2.5)$$

where the parameter λ balances sparsity of the solution and fidelity of the approximation to \mathbf{y} . Notice that this is essentially a linear regression regularized with ℓ^1 -norm on the coefficients, known in statistical literature as the Lasso [28].

Solving Eq. (2.5) individually for each local patch does not guarantee the compatibility between adjacent patches. We enforce compatibility between adjacent patches using a one-pass algorithm similar to that of [29].⁴ The patches are processed in raster-scan order in the image, from left to right and top to bottom. We modify Eq. (2.4) so that the super-resolution reconstruction $\mathbf{D}_h\boldsymbol{\alpha}$ of patch \mathbf{y} is constrained to closely agree with the previously computed adjacent

²Traditionally, one would seek the sparsest $\boldsymbol{\alpha}$ s.t. $\|\mathbf{D}_l\boldsymbol{\alpha} - \mathbf{y}\|_2 \leq \epsilon$. For super-resolution, it is more appropriate to replace this 2-norm with a quadratic norm $\|\cdot\|_{FF^T}$ that penalizes visually salient high-frequency errors.

³There are also some recent works showing certain non-convex optimization problems can produce superior sparse solutions to the ℓ^1 convex problem, e.g., [26] and [27].

⁴There are different ways to enforce compatibility. In [10], the values in the overlapped regions are simply averaged, which will result in blurring effects. The greedy one-pass algorithm [29] is shown to work almost as well as the use of a full MRF model [8]. Our algorithm, not based on the MRF model, is essentially the same by trusting partially the previously recovered high-resolution image patches in the overlapped regions.

high-resolution patches. The resulting optimization problem is

$$\begin{aligned} \min \|\boldsymbol{\alpha}\|_1 \quad \text{s.t.} \quad & \|F\mathbf{D}_l\boldsymbol{\alpha} - F\mathbf{y}\|_2^2 \leq \epsilon_1, \\ & \|P\mathbf{D}_h\boldsymbol{\alpha} - \mathbf{w}\|_2^2 \leq \epsilon_2, \end{aligned} \quad (2.6)$$

where the matrix P extracts the region of overlap between the current target patch and previously reconstructed high-resolution image, and \mathbf{w} contains the values of the previously reconstructed high-resolution image on the overlap. The constrained optimization Eq. (2.6) can be similarly reformulated as

$$\min_{\boldsymbol{\alpha}} \|\tilde{\mathbf{D}}\boldsymbol{\alpha} - \tilde{\mathbf{y}}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \quad (2.7)$$

where $\tilde{\mathbf{D}} = \begin{bmatrix} F\mathbf{D}_l \\ \beta P\mathbf{D}_h \end{bmatrix}$ and $\tilde{\mathbf{y}} = \begin{bmatrix} F\mathbf{y} \\ \beta\mathbf{w} \end{bmatrix}$. The parameter β controls the trade-off between matching the low-resolution input and finding a high-resolution patch that is compatible with its neighbors. In all our experiments, we simply set $\beta = 1$. Given the optimal solution $\boldsymbol{\alpha}^*$ to Eq. (2.7), the high-resolution patch can be reconstructed as $\mathbf{x} = \mathbf{D}_h\boldsymbol{\alpha}^*$.

2.2.2 Enforcing global reconstruction constraint

Notice that Eq. (2.4) and Eq. (2.6) do not demand exact equality between the low-resolution patch \mathbf{y} and its reconstruction $\mathbf{D}_l\boldsymbol{\alpha}$. Because of this, and also because of noise, the high-resolution image \mathbf{X}_0 produced by the sparse representation approach of the previous section may not satisfy the reconstruction constraint Eq. (2.1) exactly. We eliminate this discrepancy by projecting \mathbf{X}_0 onto the solution space of $S\mathbf{H}\mathbf{X} = \mathbf{Y}$, computing

$$\mathbf{X}^* = \arg \min_{\mathbf{X}} \|\mathbf{S}\mathbf{H}\mathbf{X} - \mathbf{Y}\|_2^2 + c\|\mathbf{X} - \mathbf{X}_0\|_2^2. \quad (2.8)$$

Algorithm 1 (Super-Resolution via Sparse Representation).

- 1: **Input:** training dictionaries \mathbf{D}_h and \mathbf{D}_l , a low-resolution image \mathbf{Y} .
- 2: **For** each 3×3 patch \mathbf{y} of \mathbf{Y} , taken starting from the upper-left corner with 1 pixel overlap in each direction,
 - Compute the mean pixel value m of patch \mathbf{y} .
 - Solve the optimization problem with $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{y}}$ defined in Eq. (2.7):
 $\min_{\boldsymbol{\alpha}} \|\tilde{\mathbf{D}}\boldsymbol{\alpha} - \tilde{\mathbf{y}}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$.
 - Generate the high-resolution patch $\mathbf{x} = \mathbf{D}_h\boldsymbol{\alpha}^*$. Put the patch $\mathbf{x} + m$ into a high-resolution image \mathbf{X}_0 .
- 3: **End**
- 4: Using gradient descent, find the closest image to \mathbf{X}_0 which satisfies the reconstruction constraint:

$$\mathbf{X}^* = \arg \min_{\mathbf{X}} \|\mathbf{SHX} - \mathbf{Y}\|_2^2 + c\|\mathbf{X} - \mathbf{X}_0\|_2^2.$$

- 5: **Output:** super-resolution image \mathbf{X}^* .
-

The solution to this optimization problem can be efficiently computed using gradient descent. The update equation for this iterative method is

$$\mathbf{X}_{t+1} = \mathbf{X}_t + \nu[H^T S^T(\mathbf{Y} - \mathbf{SHX}_t) + c(\mathbf{X} - \mathbf{X}_0)], \quad (2.9)$$

where \mathbf{X}_t is the estimate of the high-resolution image after the t -th iteration, and ν is the step size of the gradient descent.

We take result \mathbf{X}^* from the above optimization as our final estimate of the high-resolution image. This image is as close as possible to the initial super-resolution \mathbf{X}_0 given by sparsity, while respecting the reconstruction constraint. The entire super-resolution process is summarized as Algorithm 1.

2.2.3 Global optimization interpretation

The simple SR algorithm outlined in the previous two subsections can be viewed as a special case of a more general sparse representation framework for inverse problems in image processing. Related ideas have been profitably applied in image compression, denoising [19], and restoration [20]. In addition to placing our work in a larger context, these connections suggest means of further improving the performance, at the cost of increased computational complexity.

Given sufficient computational resources, one could in principle solve for the coefficients associated with all patches *simultaneously*. Moreover, the entire high-resolution image \mathbf{X} itself can be treated as a variable. Rather than demanding that \mathbf{X} be perfectly reproduced by the sparse coefficients $\boldsymbol{\alpha}$, we can penalize the difference between \mathbf{X} and the high-resolution image given by these coefficients, allowing solutions that are not perfectly sparse but that better satisfy the reconstruction constraints. This leads to a large optimization problem:

$$\begin{aligned} \mathbf{X}^* = \arg \min_{\mathbf{X}, \{\boldsymbol{\alpha}_{ij}\}} & \left\{ \|SH\mathbf{X} - \mathbf{Y}\|_2^2 + \lambda \sum_{i,j} \|\boldsymbol{\alpha}_{ij}\|_0 \right. \\ & \left. + \gamma \sum_{i,j} \|D_h \boldsymbol{\alpha}_{ij} - P_{ij}\mathbf{X}\|_2^2 + \tau \rho(\mathbf{X}) \right\}. \end{aligned} \quad (2.10)$$

Here, $\boldsymbol{\alpha}_{ij}$ denotes the representation coefficients for the $(i, j)_{th}$ patch of \mathbf{X} , and P_{ij} is a projection matrix that selects the $(i, j)_{th}$ patch from \mathbf{X} . A penalty function $\rho(\mathbf{X})$ encodes additional prior knowledge about the high-resolution image. This function may depend on the image category or may take the form of a generic regularization term (e.g., Huber MRF, total variation, bilateral total variation).

Algorithm 1 can be interpreted as a computationally efficient approximation to Eq. (2.10). The sparse representation step recovers the coefficients $\boldsymbol{\alpha}$ by approximately minimizing the sum of the second and third terms of Eq. (2.10). The

sparsity term $\|\boldsymbol{\alpha}_{ij}\|_0$ is relaxed to $\|\boldsymbol{\alpha}_{ij}\|_1$, while the high-resolution fidelity term $\|\mathbf{D}_h\boldsymbol{\alpha}_{ij} - P_{ij}\mathbf{X}\|_2$ is approximated by its low-resolution version $\|F\mathbf{D}_l\boldsymbol{\alpha}_{ij} - F\mathbf{y}_{ij}\|_2$.

Notice, that if the sparse coefficients $\boldsymbol{\alpha}$ are fixed, the third term of Eq. (2.10) essentially penalizes the difference between the super-resolution image \mathbf{X} and the reconstruction given by the coefficients: $\sum_{i,j} \|\mathbf{D}_h\boldsymbol{\alpha}_{ij} - P_{ij}\mathbf{X}\|_2^2 \approx \|\mathbf{X}_0 - \mathbf{X}\|_2^2$. Hence, for small γ , the back-projection step of Algorithm 1 approximately minimizes the sum of the first and third terms of Eq. (2.10).

Algorithm 1 does not, however, incorporate any prior besides sparsity of the representation coefficients, i.e., the term $\rho(\mathbf{X})$ is absent in our approximation. In Chapter 4 we will see that sparsity in a relevant dictionary is a strong enough prior that we can already achieve good super-resolution performance. Nevertheless, in settings where further assumptions on the high-resolution signal are available, these priors can be incorporated into the global reconstruction step of our algorithm.

2.3 Face Super-resolution from Sparsity

Face image resolution enhancement is usually desirable in many surveillance scenarios, where there is always a large distance between the camera and the objects (people) of interest. Unlike the generic image super-resolution discussed earlier, face images are more regular in structure and thus should be easier to handle. Indeed, for face super-resolution, we can deal with even lower-resolution input images. The basic idea is first to use the face prior to zoom the input to a reasonable medium-resolution, and then to employ the local sparsity prior model to recover details. To be precise, the solution is also approached in two steps: 1) global model: use reconstruction constraint to recover a medium high-resolution face image, but the solution is searched only in the face subspace; and 2) local

model: use the local sparse model to recover the image details.

2.3.1 Nonnegative matrix factorization modeling

In face super-resolution, the most frequently used subspace method for modeling the human face is principal component analysis (PCA), which chooses a low-dimensional subspace that captures as much of the variance as possible. However, the PCA bases are holistic, and tend to generate smooth faces similar to the mean. Moreover, because principal component representations allow negative coefficients, the PCA reconstruction is often hard to interpret.

Even though faces are objects with lots of variance, they are made up of several relatively independent parts, such as eyes, eyebrows, noses, mouths, checks, and chins. Nonnegative matrix factorization (NMF) [30] seeks a representation of the given signals as an additive combination of local features. To find such a part-based subspace, NMF is formulated as the following optimization problem:

$$\begin{aligned} \arg \min_{U, V} \|X - UV\|_2^2 \\ \text{s.t. } U \geq 0, V \geq 0, \end{aligned} \tag{2.11}$$

where $X \in \mathbb{R}^{n \times m}$ is the data matrix, $U \in \mathbb{R}^{n \times r}$ is the basis matrix, and $V \in \mathbb{R}^{r \times m}$ is the coefficient matrix. In our context here, X simply consists of a set of pre-aligned high-resolution training face images as its column vectors. The number of the bases r can be chosen as $n * m / (n + m)$, which is smaller than n and m , and means a more compact representation. It can be shown that a locally minimum of Eq. (2.11) can be obtained via the following update rules:

$$\begin{aligned} V_{ij} &\leftarrow V_{ij} \frac{(U^T X)_{ij}}{(U^T U V)_{ij}} \\ U_{ki} &\leftarrow U_{ki} \frac{(X V^T)_{ki}}{(U V V^T)_{ki}}, \end{aligned} \tag{2.12}$$

where $1 \leq i \leq r$, $1 \leq j \leq m$ and $1 \leq k \leq n$. The obtained basis matrix U is often sparse and localized.

2.3.2 Two step face super-resolution

Let \mathbf{X} and \mathbf{Y} denote the high- and low-resolution faces respectively. We obtain \mathbf{Y} from \mathbf{X} by smoothing and downsampling as in Eq. (2.1). We want to recover \mathbf{X} from the observation \mathbf{Y} . In this context, we assume \mathbf{Y} has been pre-aligned to the training database by either manually labeling the feature points or with some automatic face alignment algorithm such as the method used in [13]. We can achieve the optimal solution for \mathbf{X} based on the maximum a posteriori (MAP) criteria:

$$\mathbf{X}^* = \arg \max_{\mathbf{X}} p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}). \quad (2.13)$$

Here, $p(\mathbf{Y}|\mathbf{X})$ models the image observation process, usually with Gaussian noise assumption on the observation \mathbf{Y} , $p(\mathbf{Y}|\mathbf{X}) = 1/Z \exp(-\|SHU\mathbf{c} - \mathbf{Y}\|_2^2/(2 * \sigma^2))$ with Z being a normalization factor, and $p(\mathbf{X})$ is a prior on the underlying high-resolution image \mathbf{X} , typically in the exponential form $p(\mathbf{X}) = \exp(-c\rho(\mathbf{X}))$. Using the rules in Eq. (2.12), we can obtain the basis matrix U , which is composed of sparse bases. Let Ω denote the face subspace spanned by U . Then in the subspace Ω , the super-resolution problem in Eq. (2.13) can be formulated using the reconstruction constraints as

$$\mathbf{c}^* = \arg \min_{\mathbf{c}} \|SHU\mathbf{c} - \mathbf{Y}\|_2^2 + \eta\rho(U\mathbf{c}) \quad \text{s.t.} \quad \mathbf{c} \geq 0, \quad (2.14)$$

where $\rho(U\mathbf{c})$ is a prior term regularizing the high-resolution solution, $\mathbf{c} \in \mathbb{R}^{r \times 1}$ is the coefficient vector in the subspace Ω for estimated the high-resolution face, and η is a parameter used to balance the reconstruction fidelity and the penalty of the

Algorithm 2 (Face Hallucination via Sparse Representation).

- 1: Input: sparse basis matrix U , training dictionaries D_h and D_l , a low-resolution aligned face image Y .
 - 2: Find a smooth high-resolution face \hat{X} from the subspace spanned by U through:
 - Solve the optimization problem in Eq. (2.15):
$$\arg \min_{\mathbf{c}} \|SHU\mathbf{c} - \mathbf{Y}\|_2^2 + \eta\|\Gamma U\mathbf{c}\|_2 \quad s.t. \quad \mathbf{c} \geq 0.$$
 - $\hat{X} = U\mathbf{c}^*$.
 - 3: For each patch \mathbf{y} of \hat{X} , taken starting from the upper-left corner with 1 pixel overlap in each direction,
 - Compute and record the mean pixel value of \mathbf{y} as m .
 - Solve the optimization problem with \tilde{D} and $\tilde{\mathbf{y}}$ defined in Eq. (2.7):
$$\min_{\alpha} \|\tilde{D}\alpha - \tilde{\mathbf{y}}\|_2^2 + \lambda\|\alpha\|_1.$$
 - Generate the high-resolution patch $\mathbf{x} = D_h\alpha^* + m$. Put the patch \mathbf{x} into a high-resolution image X^* .
 - 4: Output: super-resolution face X^* .
-

prior term. In this thesis, we simply use a generic image prior requiring that the solution be smooth. Let Γ denote a matrix performing high-pass filtering. The final formulation for Eq. (2.14) is

$$\mathbf{c}^* = \arg \min_{\mathbf{c}} \|SHU\mathbf{c} - \mathbf{Y}\|_2^2 + \eta\|\Gamma U\mathbf{c}\|_2 \quad s.t. \quad \mathbf{c} \geq 0. \quad (2.15)$$

The medium high-resolution image \hat{X} is approximated by $U\mathbf{c}^*$. The prior term in Eq. (2.15) suppresses the high-frequency components, resulting in over-smoothness in the solution image. We rectify this using the local patch model based on sparse representation mentioned earlier in Section 2.2.1. The complete framework of our algorithm is summarized as Algorithm 2.

CHAPTER 3

LEARNING THE DICTIONARY PAIR

In the previous chapter, we discussed regularizing the super-resolution problem using the sparse prior that each pair of high- and low-resolution image patches has the same sparse representations with respect to the two dictionaries \mathbf{D}_h and \mathbf{D}_l . A straightforward way to obtain two such dictionaries is to sample image patch pairs directly, which preserves the correspondence between the high- and low-resolution patch items [15]. However, such a strategy will result in large dictionaries and, hence, expensive computation. This chapter will focus on learning a more compact dictionary pair for speeding up the computation.

3.1 Single Dictionary Training

Sparse coding is the problem of finding sparse representations of the signals with respect to an over-complete dictionary \mathbf{D} . The dictionary is usually learned from a set of training examples $X = \{x_1, x_2, \dots, x_t\}$. Generally, it is hard to learn a compact dictionary that guarantees that sparse representation of Eq. (2.3) can be recovered from ℓ_1 minimization in Eq. (2.4). Fortunately, many sparse coding algorithms proposed previously suffice for practical applications. In this thesis, we focus on the following formulation:

$$\begin{aligned} \mathbf{D} &= \arg \min_{\mathbf{D}, Z} \|\mathbf{X} - \mathbf{D}Z\|_2^2 + \lambda \|Z\|_1 \\ &\text{s.t. } \|D_i\|_2^2 \leq 1, i = 1, 2, \dots, K, \end{aligned} \tag{3.1}$$

where the ℓ_1 norm $\|Z\|_1$ is to enforce sparsity, and the ℓ_2 norm constraints on the columns of \mathbf{D} remove the scaling ambiguity.¹ This particular formulation has been studied extensively [22, 23, 31]. Equation (3.1) is not convex in both \mathbf{D} and Z , but is convex in one of them with the other fixed. The optimization performs in an alternative manner over Z and \mathbf{D} :

1. Initialize \mathbf{D} with a Gaussian random matrix, with each column unit normalized.
2. Fix \mathbf{D} , update Z by

$$Z = \arg \min_Z \|X - \mathbf{D}Z\|_2^2 + \lambda \|Z\|_1, \quad (3.2)$$

which can be solved efficiently through linear programming.

3. Fix Z , update \mathbf{D} by

$$\begin{aligned} \mathbf{D} &= \arg \min_{\mathbf{D}} \|X - \mathbf{D}Z\|_2^2 \\ &\text{s.t. } \|D_i\|_2^2 \leq 1, i = 1, 2, \dots, K, \end{aligned} \quad (3.3)$$

which is a quadratically constrained quadratic programming that is ready to be solved in many optimization packages.

4. Iterate between 2) and 3) until they converge. In our implementation, we used a Matlab package developed in [22].

¹Note that without the norm constraints, the cost can always be reduced by dividing Z by $c > 1$ and multiplying \mathbf{D} by $c > 1$.

3.2 Joint Dictionary Training

Given the sampled training image patch pairs $P = \{X^h, Y^l\}$, where $X^h = \{x_1, x_2, \dots, x_n\}$ are the set of sampled high-resolution image patches and $Y^l = \{y_1, y_2, \dots, y_n\}$ are the corresponding low-resolution image patches (or features), our goal is to learn dictionaries for high- and low-resolution image patches, so that the sparse representation of the high-resolution patch is the same as the sparse representation of the corresponding low-resolution patch. This is a difficult problem, due to the ill-posed nature of super-resolution. The individual sparse coding problems in the high-resolution and low-resolution patch spaces are

$$\mathbf{D}_h = \arg \min_{\{\mathbf{D}_h, Z\}} \|X^h - \mathbf{D}_h Z\|_2^2 + \lambda \|Z\|_1 \quad (3.4)$$

and

$$\mathbf{D}_l = \arg \min_{\{\mathbf{D}_l, Z\}} \|Y^l - \mathbf{D}_l Z\|_2^2 + \lambda \|Z\|_1 \quad (3.5)$$

respectively. We combine these objectives, forcing the high-resolution and low-resolution representations to share the same codes, instead writing

$$\begin{aligned} \min_{\{\mathbf{D}_h, \mathbf{D}_l, Z\}} & \frac{1}{N} \|X^h - \mathbf{D}_h Z\|_2^2 + \frac{1}{M} \|Y^l - \mathbf{D}_l Z\|_2^2 \\ & + \lambda \left(\frac{1}{N} + \frac{1}{M} \right) \|Z\|_1, \end{aligned} \quad (3.6)$$

where N and M are the dimensions of the high- and low-resolution image patches in vector form. Here, $1/N$ and $1/M$ balance the two cost terms of Eq. (3.4) and Eq. (3.5). Equation (3.6) can be rewritten as

$$\min_{\{\mathbf{D}_h, \mathbf{D}_l, Z\}} \|X_c - \mathbf{D}_c Z\|_2^2 + \lambda \left(\frac{1}{N} + \frac{1}{M} \right) \|Z\|_1, \quad (3.7)$$

or equivalently

$$\min_{\{\mathbf{D}_h, \mathbf{D}_l, Z\}} \|X_c - \mathbf{D}_c Z\|_2^2 + \hat{\lambda} \|Z\|_1, \quad (3.8)$$

where

$$X_c = \begin{bmatrix} \frac{1}{\sqrt{N}} X^h \\ \frac{1}{\sqrt{M}} Y^l \end{bmatrix}, \quad \mathbf{D}_c = \begin{bmatrix} \frac{1}{\sqrt{N}} \mathbf{D}_h \\ \frac{1}{\sqrt{M}} \mathbf{D}_l \end{bmatrix}. \quad (3.9)$$

Thus, we can use the same learning strategy in the single dictionary case for training the two dictionaries for our super-resolution purpose. Note that since we are using features from the low-resolution image patches, \mathbf{D}_h and \mathbf{D}_l are not simply connected by a linear transform; otherwise, the training process of Eq. (3.8) will depend on the high-resolution image patches only (for details, refer to Section 3.3). Figure 3.1 shows the dictionary learned by Eq. (3.8) for generic images.² The learned dictionary demonstrates basic patterns of the image patches, such as orientated edges, instead of raw patch prototypes, due to its compactness.

3.3 Feature Representation for Low-resolution Image Patches

In Eq. (2.3), we use a feature transformation F to ensure that the computed coefficients fit the most relevant part of the low-resolution signal and, hence, have a more accurate prediction for the high-resolution image patch reconstruction. Typically, F is chosen as some kind of high-pass filter. This is reasonable from a perceptual viewpoint, since people are more sensitive to the high-frequency content of the image. The high-frequency components of the low-resolution image are also arguably the most important for predicting the lost high-frequency content

²We omit the dictionary for the low-resolution image patches because we are training on features instead the patches themselves.

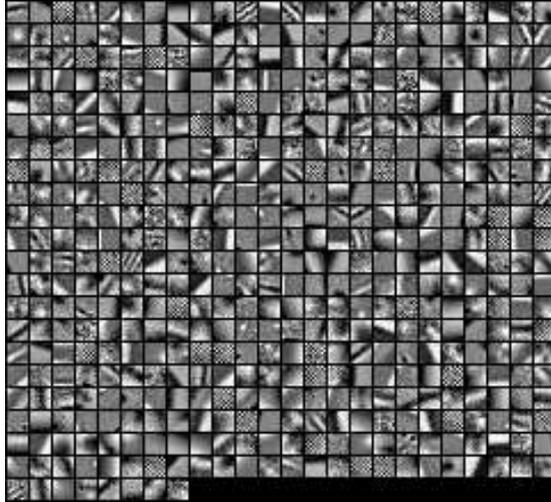


Figure 3.1: The high-resolution image patch dictionary trained by Eq. (3.8) using 100,000 high- and low-resolution image patch pairs sampled from the generic training images. A total of 512 dictionary atoms are learned with each atom of size 9×9 .

in the target high-resolution image.

In the literature, researchers have suggested extracting different features for the low-resolution image patch in order to boost the prediction accuracy. Freeman et al. [8] used a high-pass filter to extract the edge information from the low-resolution input patches as the feature. Sun et al. [9] used a set of Gaussian derivative filters to extract the contours in the low-resolution patches. Chang et al. [10] used the first- and second-order gradients of the patches as the representation. Here, we also use the first- and second-order derivatives as the feature for the low-resolution patch due to their simplicity and effectiveness. The four 1-D filters used to extract the derivatives are

$$\begin{aligned}
 \mathbf{f}_1 &= [-1, 0, 1], & \mathbf{f}_2 &= \mathbf{f}_1^T, \\
 \mathbf{f}_3 &= [1, 0, -2, 0, 1], & \mathbf{f}_4 &= \mathbf{f}_3^T,
 \end{aligned}
 \tag{3.10}$$

where the superscript T means transpose. Applying these four filters yields four

feature vectors for each patch, which are concatenated into one vector as the final representation of the low-resolution patch. In our implementation, the four filters are not applied directly to the sampled low-resolution image patches; instead, we apply the four filters to the training images. Thus, for each low-resolution training image, we get four gradient maps; and we extract four patches from these gradient maps at each location and concatenate them to become the feature vector. Therefore, the feature representation for each low-resolution image patch also encodes its neighboring information, which is beneficial for promoting compatibility among adjacent patches in the final super-resolution image.

In practice, we find that it works better to extract the features from the upsampled version of the low-resolution image instead of the original one. That is, we first upsample the low-resolution image by a factor of two³ using bicubic interpolation, and then extract gradient features from it. Since we know all the zoom ratios, it is easy to track the correspondence between high-resolution image patches and the upsampled low-resolution image patches both for training and testing. Because of the way of extracting features from the low-resolution image patches, the two dictionaries \mathbf{D}_h and \mathbf{D}_l are not simply linearly connected, making the joint learning process in Eq. (3.8) more reasonable.

³We choose two mainly for dimension considerations. For example, if we work on 3×3 patches in the low-resolution image, by upsampling the image by ratio of two, the final feature for the nine-dimensional low-resolution patch will be $6 \times 6 \times 4 = 144$.

CHAPTER 4

EXPERIMENT EVALUATION AND ANALYSIS

In this chapter, we first demonstrate the super-resolution results obtained by applying the above methods on both generic and face images. We then move on to discuss various influential factors for the proposed algorithm, including dictionary size, noise with inputs, and the global reconstruction constraints.

In our experiments, we magnify the input low-resolution image by a factor of three for generic images and four for face images, which is commonplace in the literature of single frame super-resolution. In generic image super-resolution, for the low-resolution images, we always use 3×3 low-resolution patches (upsampled to 6×6), with overlap of 1 pixel between adjacent patches, corresponding to 9×9 patches with overlap of 3 pixels for the high-resolution patches. In face super-resolution, we choose the patch size as 5×5 pixels for both low- and high-resolution face images. For color images, we apply our algorithm to the illuminance channel only, since humans are more sensitive to illuminance changes. We therefore interpolate the color layers (Cb, Cr) using plain bicubic interpolation. We evaluate the results of various methods both visually and qualitatively in root mean square error (RMSE). Even though RMSE is a common criterion in image processing for recovery, it is not quite reliable for rating visual image quality [32], as we will see in the following sections. Note that since we only work on the illuminance channel, the RMSE reported is carried out only on the illuminance channel.

4.1 Single Image Super-resolution

4.1.1 Generic image super-resolution

We apply our methods to generic images such as flowers, human faces, and buildings. The two dictionaries for high- and low-resolution image patches are trained from 100,000 patch pairs sampled from natural images collected from the Internet. We fix the dictionary size as 1024 in all our experiments, which is a balance between computation and image quality. In Section 4.2 we will examine the effects of different dictionary sizes. In the super-resolution algorithm Eq. (2.7), the choice of λ depends on the level of noise in the input image, which we will discuss further in Section 4.3. For generic low-noise images, we always set $\lambda = 0.01$ in all our experiments, which generally yields satisfactory results.

Figure 4.1 compares the outputs of our method with those of the neighborhood embedding method [10]. The neighborhood embedding method is similar to ours in the sense that both methods use the linear combination weights derived from the low-resolution image patch to generate the underlying high-resolution image patch. Unlike our method, the neighborhood embedding method uses fixed k nearest neighbors to find the reconstruction supports and does not include a dictionary training phase. To make fair comparison, we use the 100,000 patch pairs for the neighborhood embedding and try different k 's to get the most visually appealing results. Using a compact dictionary pair, our method is much faster and yet can generate sharper results. As the reconstructed images show in Figure 4.1, there are noticeable differences in the texture of the leaves, the fuzz on the leafstalk, and also the freckles on the face of the girl by comparing the two methods.

In Figure 4.2, we compare our method with several more state-of-the-art methods on an image of the Parthenon used in [6], including back projection [33],



Figure 4.1: The flower and girl image magnified by a factor of three. Left to right: input, bicubic interpolation, neighbor embedding, our method, and the original.

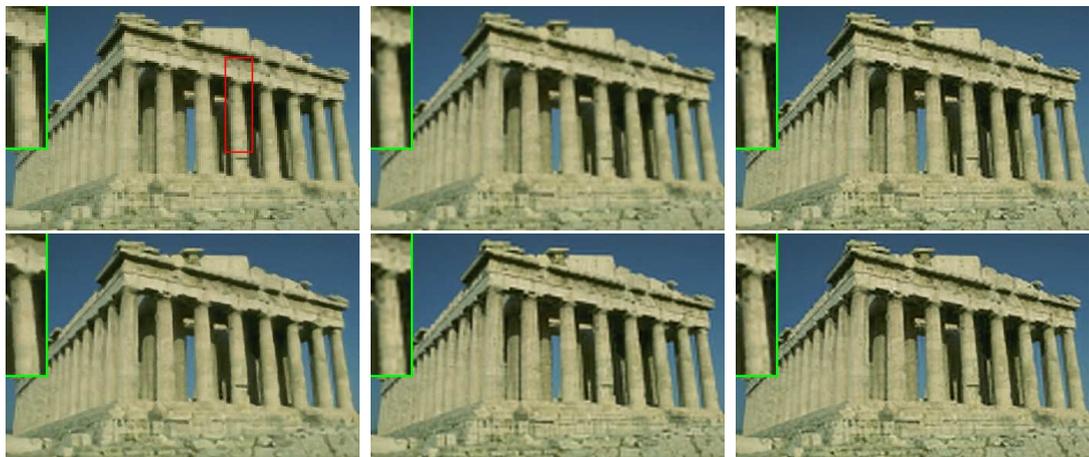


Figure 4.2: Results on an image of the Parthenon with magnification factor three. Top row: low-resolution input, bicubic interpolation, back projection. Bottom row: neighbor embedding, soft edge prior, and our method.

neighbor embedding [10], and the recently proposed method based on a learned soft edge prior [6]. The result from back projection has many jagged effects along the edges. Neighbor embedding generates sharp edges in places, but blurs the texture on the temple's facade. The soft edge prior method gives a decent reconstruction, but introduces undesired smoothing that is not present in our result.

4.1.2 Face super-resolution

In this part, we evaluate our proposed super-resolution algorithm on frontal view human faces. The experiments are conducted on the face database FRGC Ver 1.0 [34]. All these face images were aligned by an automatic alignment algorithm using the eye positions, and then cropped to the size of 100×100 pixels. To obtain the face subspace Ω spanned by W , we select 540 face images as training, including both genders, different races, various ages, and different facial expressions (Figure 4.3). To prepare the coupled dictionaries needed for our sparse representation algorithm, we also sample 100,000 patch pairs from the training images and train the dictionary pair of size 1024. Thirty new face images (from people not in the training set) are chosen as our testing cases, which are blurred and downsampled to the size of 25×25 pixels.

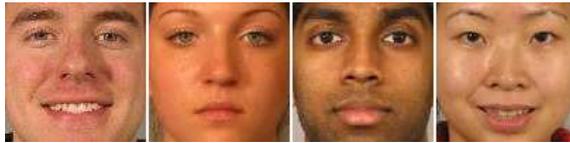


Figure 4.3: Example training faces for the face super-resolution algorithm. The training images include faces of both genders, different ages, different races, and various facial expressions.

As earlier mentioned, face image super-resolution can handle more challenging tasks than generic image super-resolution due to the regular face structure. Indeed, it is not an easy job to zoom the 25×25 low-resolution face image by 4 times using the method for generic image super-resolution. First, the downsampling process loses so much information that it is difficult to predict well a 12×12 high-resolution patch given only a 3×3 image patch. Second, the resolution of the face image is so low that the structures of the face that are useful for super-resolution inference (such as corners and edges) collapses into only several pixels. The two-step approach for face super-resolution, on the other hand, can compen-

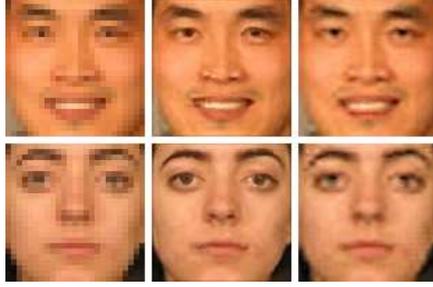


Figure 4.4: The comparison between the two-step face hallucination algorithm with the generic image super-resolution algorithm applied to low-resolution face images. From left to right: input image, super-resolution result using the two-step approach, and super-resolution result using the generic approach.

sate for the lost information in the first step using the redundancy of the face structure by searching the solution in the face subspace regarding the reconstruction constraints. The local model from sparse representation then can be further employed to enhance the edges and textures to achieve sharper results. We also apply the method for generic image directly to the face images, and compare the results with the proposed two-step approach, as shown in Figure 4.4. Since the resolution of the input face image is so low, directly applying the generic approach does not seem to generate a satisfying image.

In our experiments with face images, we also set $\lambda = 0.01$ for sparsity regularization. We compare our algorithm with bicubic interpolation [5] and back-projection [33]. The results are shown in Figure 4.5, which indicate that our method can generate much higher resolution faces. From columns four and five, we can also see that the local patch method based on sparse representation further enhances the edges and textures.

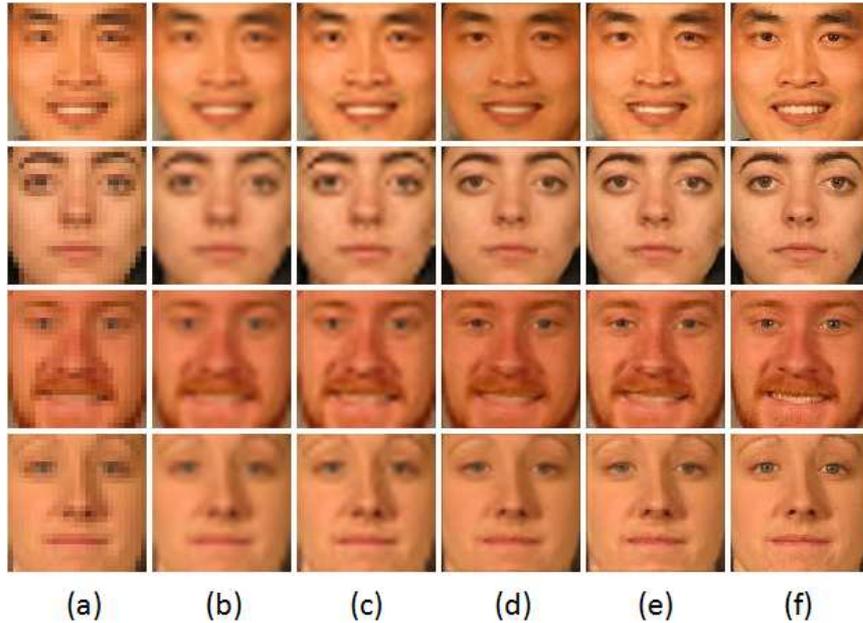


Figure 4.5: Results of our algorithm compared to other methods. From left to right columns: (a) low-resolution input; (b) bicubic interpolation; (c) back-projection; (c) global model via NMF followed by bilateral filtering; (d) global model combined with local model via sparse representation; (f) original.

4.2 Effects of Dictionary Size

The above experimental results show that the sparsity prior for image patches is very effective in regularizing the otherwise ill-posed super-resolution problem. In those results, we fix the dictionary size to be 1024. Intuitively, larger dictionaries should possess more representation power (in the extreme, we can use the sampled patches as the dictionary directly, as in [15]), and thus may yield more accurate approximation, while increasing the computation cost. In this section, we evaluate the effect of dictionary size on generic image super-resolution. From the sampled 100,000 image patch pairs, we train four dictionaries of size 256, 512, 1024, and 2048, and apply them to the same input image. We also use the 100,000 image patches directly as the dictionary for comparison. The results are evaluated both visually and quantitatively in RMSE.

Table 4.1: The RMSEs of the reconstructed images using dictionaries of different sizes, and using the raw image patches directly from which the dictionaries are trained.

Images	bicubic	D256	D512	D1024	D2048	Raw Patches
Girl	5.912	5.606	5.603	5.491	5.473	5.483
Flower	3.530	3.266	3.271	3.212	3.164	3.139
Lena	7.360	6.587	6.572	6.359	6.232	6.029
Statue	9.873	8.826	8.777	8.342	8.237	8.255



Figure 4.6: The effects of dictionary size on the super-resolution reconstruction of Lena. From left to right: dictionary size 256, 512, 1024, and 2048 and the whole sampled patch set.

Figure 4.6 shows the reconstructed results for the Lena image using dictionaries of different sizes. While there are not many visual differences between the results using different dictionary sizes from 256 to 2048 and the whole sampled patch set, we indeed observe the reconstruction artifacts will gradually diminish with larger dictionaries. The visual observation is also supported by the RMSEs of the recovered images. In Table 4.1, we list the RMSEs of the reconstructed images for dictionaries of different sizes. As shown in the table, using larger dictionaries will yield smaller RMSEs, and all of them have smaller RMSEs than those generated by bicubic interpolation. However, the computation is approximately linear to the size of the dictionary; larger dictionaries will result in heavier computation. Figure 4.7 shows the computation time in seconds with “Girl” as the test image. The algorithm is written in Matlab without optimization for speed, and run on

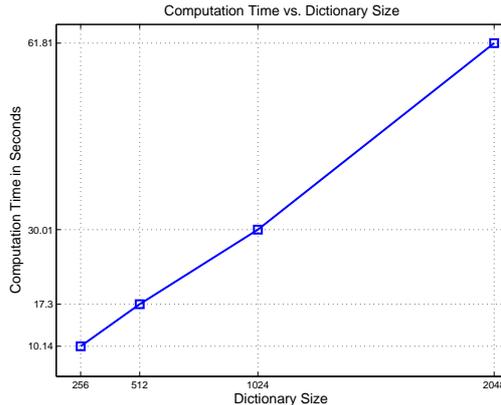


Figure 4.7: The computation time on “Girl” image with dictionaries of different sizes (in seconds).

a laptop of Core duo @ 1.83G with 2G memory. To compare with [15], the computation time is almost an hour, much slower than our current solution with trained compact dictionaries. In practice, one chooses the appropriate dictionary size as a trade-off between reconstruction quality and computation. We find that a dictionary size of 1024 can yield decent outputs, while allowing fast computation.

4.3 Robustness to Noise

Most single-input super-resolution algorithms assume that the input images are clean and free of noise, an assumption that is likely to be violated in real applications. To deal with noisy data, previous algorithms usually divide the recovery process into two disjoint steps: first denoising and then super-resolution. However, the results of such a strategy depend on the specific denoising technique, and any artifacts created during denoising on the low-resolution image will be kept or even magnified in the latter super-resolution process. Here, we demonstrate that by formulating the problem into our sparse representation model, our method is much more robust to noise with input and, thus, can handle super-resolution and denoising simultaneously. Note that in Eq. (2.5) the parameter λ depends on

Table 4.2: The RMSEs of the reconstructed images from different levels of noisy inputs.

Noise Levels / Gaussian σ	0	4	6	8
bicubic	9.873	10.423	11.037	11.772
Neighbor Embedding	9.534	10.734	11.856	13.064
Our method	8.359	9.240	10.454	11.448

the noise level of the input data; the noisier the data, the larger the value of λ should be. Figure 4.8 shows how λ influences the reconstructed results given the same noiseless input image. The larger λ , the smoother the result image texture becomes. This is obvious by formulating Eq. (2.7) into a maximum a posterior (MAP) problem:

$$\boldsymbol{\alpha}^* = \arg \max P(\boldsymbol{\alpha}) \cdot P(\tilde{y}|\boldsymbol{\alpha}, \tilde{\mathbf{D}}), \quad (4.1)$$

where

$$\begin{aligned} P(\boldsymbol{\alpha}) &= \frac{1}{2b} \exp\left(-\frac{\|\boldsymbol{\alpha}\|_1}{b}\right) \\ P(\tilde{y}|\boldsymbol{\alpha}, \tilde{\mathbf{D}}) &= \frac{1}{2\sigma^2} \exp\left(-\frac{1}{2\sigma^2} \|\tilde{\mathbf{D}}\boldsymbol{\alpha} - \tilde{\mathbf{y}}\|_2^2\right), \end{aligned} \quad (4.2)$$

where b is the variance of the Laplacian prior on $\boldsymbol{\alpha}$, and σ^2 is the variance of the noise assumed on the data \tilde{y} . Taking the negative log likelihood in Eq. (4.1), we get the exact optimization problem in Eq. (2.7), with $\lambda = \sigma^2/b$. Suppose the Laplacian variance b is fixed; then if the data becomes noisier (larger σ^2), λ becomes larger. On the other hand, given the input image, the larger the value of λ we set, the more noisy the model will assume the data to be, and thus the model tends to generate smoother results.

To test the robustness of our algorithm to noise, we add different levels of Gaussian noise to the low-resolution input image. The standard deviation of the Gaussian noise ranges from 4 to 10. The regularization parameter λ is empiri-



Figure 4.8: The effects of λ on the recovered image given the input. From left to right, $\lambda = 0.01, 0.05, 0.1, 0.2, 0.3$. The larger λ is, the smoother the result image gets. Note that the results are generated from the local model only.

cally set to be one-tenth of the standard deviation. In Figure 4.9, we show the results of our algorithm applying to the Liberty statue image with different levels of Gaussian noise. For comparison, we also show the results of using bicubic and NE [10]. As expected, the results of bicubic is both noisy and blurred. The number of neighbors chosen decreases as the noise becomes heavier for NE to get better results. As shown, the NE method is good at preserving edges, but fails to distinguish the signal from noise, and therefore generates unwanted noisy results. Our algorithm is capable of performing denoising and super-resolution simultaneously more elegantly. Table 4.2 shows the RMSEs of the reconstructed images from different levels of noisy data. In terms of RMSE, our method outperforms both bicubic interpolation and NE in all cases.

4.4 Effects of Global Constraints

The global reconstruction constraint enforced by Eq. (2.8) is employed to refine the local image patch sparse model, ensuring the recovered high-resolution image is consistent with its low-resolution observation. In our experiments, we observe that the sparsity prior is very effective and contributes the most, while the global constraint in the second step reduces RMSE by removing some minor artifacts, which are hardly seen from the first step. Table 4.3 shows the RMSEs of the results



Figure 4.9: Performance evaluation of our proposed algorithm on noisy data. Noise level (standard deviation of Gaussian noise) from left to right columns: 0, 4, 6 and 8. Top row: input images. Middle row: recovered images using NE ($k = 13, 12, 9, 7$). Bottom row: recovered images using our method ($\lambda = 0.1, 0.4, 0.6, 0.8$).

Table 4.3: The global constraint in the second step further refines the results from local sparse model in the first step and reduces RMSEs.

Methods	Flower	Girl	Parthenon	Lena	Statue
bicubic	3.530	5.912	12.724	7.360	9.873
Local Model	3.365	5.669	12.247	6.775	8.902
Plus Global	3.212	5.491	11.875	6.359	8.237

from the local sparse model only and local model combined with the global model. The RMSEs of bicubic interpolation are again given as references. As shown, the local sparse model can achieve better RMSEs than bicubic interpolation, and the global constraint further reduces the RMSEs of the recovered images. These experiments are carried out with a dictionary size of 1024.

CHAPTER 5

CONCLUSION

In this thesis, we have discussed a new single-image super-resolution algorithm based on image patches with the sparse prior learned from natural image patches as a regularization. The approach is derived from the compressed sensing principle, which states that high-resolution sparse signals can be recovered from their downsampled version by finding the sparsest solution with respect to a properly chosen dictionary. Specifically, the sparse property of image patches is modeled as the sparse prior so as to recover the high-resolution image patches from the low-resolution image patches of the input image. Such a local sparse model is further combined with a global reconstruction model in order to obtain a global optimum. The proposed approach is applied to both generic images and face images. Experimental results demonstrate the effectiveness of using sparsity as a prior for the patch-based super-resolution. In addition, the proposed sparsity regularization is robust to noise compared to the patch-based methods previously proposed.

However, one of the most important questions for future investigation is to determine, in terms of the within-category variation, the size of the dictionary satisfying the sparse representation prior. Tighter connections to the theory of compressed sensing may also yield conditions on the appropriate patch size or feature dimension.

REFERENCES

- [1] R. C. Hardie, K. J. Barnard, and E. A. Armstrong, “Joint map registration and high-resolution image estimation using a sequence of undersampled images,” *IEEE Transactions on Image Processing*, vol. 6, pp. 1621–1633, 1997.
- [2] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, “Fast and robust multiframe super-resolution,” *IEEE Transactions on Image Processing*, vol. 13, pp. 1327–1344, 2004.
- [3] M. E. Tipping and C. M. Bishop, “Bayesian image super-resolution,” in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1303–1310, 2002.
- [4] S. Baker and T. Kanade, “Limits on super-resolution and how to break them,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.
- [5] H. S. Hou and H. C. Andrews, “Cubic spline for image interpolation and digital filtering,” *IEEE Transactions on Signal Processing*, vol. 26, pp. 508–517, 1978.
- [6] S. Dai, M. Han, W. Xu, Y. Wu, and Y. Gong, “Soft edge smoothness prior for alpha channel super resolution,” in *IEEE Conference on Computer Vision and Pattern Classification (CVPR)*, pp. 1–8, 2007.
- [7] J. Sun, Z. Xu, and H. Shum, “Image super-resolution using gradient profile prior,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.
- [8] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, “Learning low-level vision,” *International Journal of Computer Vision*, vol. 40, no. 1, pp. 25–47, 2000.
- [9] J. Sun, N. N. Zheng, H. Tao, and H. Shum, “Image hallucination with primal sketch priors,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 729–736, 2003.

- [10] H. Chang, D.-Y. Yeung, and Y. Xiong, “Super-resolution through neighbor embedding,” in *IEEE Conference on Computer Vision and Pattern Classification (CVPR)*, vol. 1, pp. 275–282, 2004.
- [11] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [12] S. Baker and T. Kanade, “Hallucinating faces,” in *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 83–88, 2000.
- [13] C. Liu, H. Y. Shum, and W. T. Freeman, “Face hallucination: Theory and practice,” *International Journal of Computer Vision*, vol. 75, no. 1, pp. 115–134, 2007.
- [14] W. Liu, D. Lin, and X. Tang, “Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 478–484, 2006.
- [15] J. Yang, J. Wright, T. Huang, and Y. Ma, “Image super-resolution as sparse representation of raw image patches,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.
- [16] E. Candès, “Compressive sensing,” in *Proceedings of the International Congress of Mathematicians*, vol. 3, pp. 1433–1452, 2006.
- [17] D. L. Donoho, “Compressed sensing,” *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [18] H. Rauhut, K. Schnass, and P. Vandergheynst, “Compressed sensing and redundant dictionaries,” *IEEE Transactions on Information Theory*, vol. 54, no. 5, May 2008.
- [19] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Transactions on Image Processing*, vol. 15, pp. 3736–3745, 2006.
- [20] J. Mairal, G. Sapiro, and M. Elad, “Learning multiscale sparse representations for image and video restoration,” *Multiscale Modeling and Simulation*, vol. 7, pp. 214–241, 2008.
- [21] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
- [22] H. Lee, A. Battle, R. Raina, and A. Y. Ng, “Efficient sparse coding algorithms,” in *Advances in Neural Information Processing Systems (NIPS)*, pp. 801–808, 2007.

- [23] B. Olshausen and D. Field, “Sparse coding with an overcomplete basis set: A strategy employed by V1?,” *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.
- [24] D. L. Donoho, “For most large underdetermined systems of linear equations, the minimal ℓ^1 -norm solution is also the sparsest solution,” *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.
- [25] D. L. Donoho, “For most large underdetermined systems of linear equations, the minimal ℓ^1 -norm near-solution approximates the sparsest near-solution,” *Communications on Pure and Applied Mathematics*, vol. 59, no. 7, pp. 907–934, 2006.
- [26] D. Wipf and S. Nagarajan, “A new view of automatic relevance determination,” in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1625–1632, 2008.
- [27] D. Wipf and S. Nagarajan, “Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions,” *Journal of Selected Topics in Signal Processing*, special issue on compressive sensing, (to be published).
- [28] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.
- [29] W. T. Freeman, T. R. Jones, and E. C. Pasztor, “Example-based super-resolution,” *IEEE Computer Graphics and Applications*, vol. 22, pp. 56–65, 2002.
- [30] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.
- [31] J. F. Murray and K. Kreutz-Delgado, “Learning sparse overcomplete codes for images,” *Journal of VLSI Signal Processing*, vol. 45, pp. 97–110, 2007.
- [32] Z. Wang and A. C. Bovik, “Mean squared error: Love it or leave it? –A new look at signal fidelity measures,” *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, Jan. 2009.
- [33] M. Irani and S. Peleg, “Motion analysis for image enhancement: Resolution, occlusion and transparency,” *Journal of Visual Communication and Image Representation*, vol. 4, no. 4, pp. 324–335, 1993.
- [34] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, “Overview of face recognition grand challenge,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 947–954, 2005.