Image Super-resolution via Sparse Representation

Aikaterini Chatzi

MASTER THESIS

— ♦ —

Ioannina, January 2013

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF IOANNINA

ΥΠΕΡΑΝΑΛΥΣΗ ΕΙΚΟΝΩΝ ΜΕΣΩ ΑΡΑΙΩΝ ΑΝΑΠΑΡΑΣΤΑΣΕΩΝ

Η
ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

Υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης
του Τμήματος Πληροφορικής
Εξεταστική Επιτροπή

από την

Αικατερίνη Χατζή

ως μέρος των Υποχρεώσεων

για τη λήψη

του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ-ΕΦΑΡΜΟΓΕΣ

Ιανουάριος 2013

CONTENTS

TABLES

FIGURES

ABSTRACT

' Chatzi, Aikaterini, C. E. MSc, Computer Science Department, University of Ioannina, Greece. January, 2013. Image Super-resolution via Sparse Representation. Thesis Supervisor: Lysimachos Paul Kondi.

Super-resolution is a mathematical term used to describe the process of enhancing the resolution of an imaging system. In the super-resolution process, the details of an image are recovered from several low-resolution images or one single low-resolution image generating high-resolution images of great quality. This technique is of great importance in applications that require zooming in a specific area of an image, such as image processing, medical imaging devices, satellite imaging devices, surveillance cameras, forensic image analysis, visual electronics and document analysis. However, what is that makes super-resolution techniques so popular and important? The answer is simple. The asset of super-resolution is that the transition from low- to high-resolution is achieved by software, using algorithms, rather than expensive hardware. The super-resolution algorithms exceed the limitations introduced by the sensors and the lens of every digital imaging system leading to remarkable high-resolution images.

This thesis focuses on exampled-based super-resolution where the goal is to learn correspondences between low- and high-resolution image patch pairs sampled from a database of low- and high-resolution images (training data), and then apply them to a new low-resolution image (test data) to recover its most likely high-resolution version. Recent results in sparse signal representation suggest that linear relationships among high-resolution signals can be precisely recovered from their low-dimensional projections. This observation has led to a new approach to single-image super-resolution, where sparse representation is used in order to generate the high-resolution outputs. The problem of single image super-resolution based upon sparse representation is examined in the present thesis and a method that reconstructs high-resolution images using sparse representation is presented. This method relies on upsampled low-resolution images to infer the high-resolution output. We propose the direct use of the low-resolution image in order to obtain high-resolution outputs and the results indicate that such a strategy provides appealing results both quantitatively and qualitatively. Furthermore, the process of obtaining the appropriate dictionaries and thus sparse representations is examined thoroughly.

CHAPTER 1. INTRODUCTION

Super-resolution is a mathematical term used to describe the process of enhancing the resolution of an imaging system. This increase of resolution in either image processing or video editing is described by many terms such as "upscale", "upsize", "up-convert" and "uprez". The basic idea of super-resolution is obtaining higher-resolution images using information from several lower-resolution ones, i.e. resolution enhancement. The use of multiple low-resolution images of the same scene in order to generate an upsized image is called multi-frame or multi-image super-resolution. There is also single-frame or single-image super-resolution which uses information drawn from other parts of the low-resolution images or other unrelated images, in order to guess what the high-resolution image should look like. Resolution means pixel density, thus it is a measure of frequency content in an image. High-resolution images offer more information/details than low-resolution ones because they have more pixels in the same area. However, even with today's progress in technology, the hardware for high-resolution images is expensive and in many cases hard to obtain (e.g. cell phones, surveillance cameras). Therefore, the answer to this problem should be software rather than hardware, and this answer is super-resolution, which is important in applications such as medical imaging, satellite imaging and computer vision.

In the classical multi-image super-resolution the subpixel misalignments between several low-resolution images of the same scene are used in order to infer the high-resolution image. This technique is of great importance in applications that require zooming in a specific area of an image such as surveillance cameras and more specifically for forensic image analysis where multiple frames of video of a suspect or a car are available and more details need to be extracted by enhancing the resolution

of these images. Multi-frame super-resolution is effective only when the low-resolution images are slightly different from each other, meaning that they include some kind of motion or different viewing angles. In this way, the entire information about the scene exceeds the information from any single image. If all the images are exactly the same then no extra information can be collected and thus, the output is just an image with less noise than the input images and without any enhancement of the resolution. Fig. 1.1 shows an example of multi-frame super-resolution. A set of similar low-resolution images of the same object (left) is used to generate the high-resolution image (right). One of the low-resolution images is zoomed (center) to be comparable with the high-resolution output.

In the classical single-image super-resolution the goal is to recover the high-resolution version of one given low-resolution image without introducing blur. In exampled-based super-resolution, also known as image hallucination, the goal is to learn correspondences between low- and high-resolution image patch pairs sampled from a database of low- and high-resolution images, and then apply them to a new low-resolution image to recover its most likely high-resolution version. Recent results in sparse signal representation suggest that linear relationships among high-resolution signals can be precisely recovered from their low-dimensional projections. This observation has led to a new approach to single-image super-resolution, where a sparse representation is found for each patch of the low-resolution input and then used to generate the high-resolution output. Fig. 1.2 shows an example of single-image super-resolution.

Face hallucination is a special case of super-resolution where the super-resolution techniques are applied on human face images in order to recover a high-resolution face image from a low-resolution image. Face hallucination is widely used in image enhancement, image compression and face recognition, and became a very popular area of research the past few years because of the advances in technology and the necessity for finding the details of facial features from low-resolution images or videos captured by surveillance cameras. The face hallucination techniques have shown remarkable results at full frontal faces and some degrees off, however at profile images they still perform poorly or badly. Fig. 1.3 shows an example of face

hallucination used for face recognition. Details of the facial characteristics of the individual are recovered (right) by a low-resolution image captured by a surveillance camera (left).

The super-resolution problem is generally treated as an inverse problem of recovering the original high-resolution image by the low-resolution ones, assuming that the low-resolution images are downsampled versions of the high-resolution image and that prior knowledge about the generation model that leads from the high-resolution image to the low-resolution ones exists. Then, the same generation model should lead to the initial low-resolution images after applying it to the recovered image. The drawback is that the reconstruction problem is generally an ill-posed problem because a lot of information is lost during the transition from high-resolution to low-resolution, the number of low-resolution images may be insufficient, the blur operators are ill-conditioned and the reconstruction solution is not unique.

Before the demonstration of image super-resolution by sparse representation, it is important to briefly present some of the main developments on the field of super-resolution. The first who considered the problem of generating a high-resolution image from several downsampled and translationally displaced images were Tsai and Huang in 1984 [1]. They formulated a set of equations in the frequency domain by using the shift property of the Fourier transform and the aliasing relationship between the continuous Fourier transform of the original high-resolution image and the discrete Fourier transform of the observed images, without considering optical blur or noise. They also assumed that the original high-resolution image is band-limited. Noise is taken into account by [2] which improves [1] by including the linear shift invariant blur point spread function. Noise and blur are also taken into account by [3] where a recursive least-squares technique is used in order to solve the same model as [1] in the presence of noise. The advantage of these frequency domain methods is that they are computationally efficient. However, super-resolution uses mostly spatial domain methods (rather than frequency domain ones), because of the advantages they provide, including flexibility in the choice of motion model, motion blur, optical blur, the sampling process and the ease in formulating the constraints (e.g. Markov random fields, projection onto convex sets) [4].

There are super-resolution techniques which are based upon interpolation. The results of some interpolation techniques, such as nearest-neighbor, least-squares plane fitting (LSP), normalized convolution (NC), exact image reconstruction (ER), iterative reconstruction (IT), can be found in [5]. These methods generate more favorable results than simple interpolation methods, such as Bilinear and Bicubic interpolation, which produce overly smooth images with notched artifacts. Another interpolation technique is presented in [6] where the background and foreground descriptors are 'used in order to represent the local image patches and then reconstruct the sharp discontinuity between the two. Images, can also be modeled as probability distribution, because in super-resolution data or parameters which are unknown need to be estimated. [7] uses Maximum a posteriori (MAP) estimation and produces accurate motion estimates assuming rigid-body motion. However, a more important progress is done in [8], where the authors estimate the high-resolution image and the motion parameters simultaneously. Since image super-resolution reconstruction is a severely underdetermined problem, regularization has been widely used to avoid this ill-posed problem [9].

Other methods used in super-resolution are the iterative methods. In [10], recursive least squares, least mean squares and steepest descent are used to approximate the Kalman filter. The performance of these techniques is also analyzed. Back-projection kernel both for image registration and restoration is used in [11], but the authors later modified their method [12] in order to deal with more complicated motion types (e.g. partial occlusion, transparency). Another category of super-resolution methods are the Projection Onto Convex Sets (POCS) where sets, which represent certain characteristics of the image, are determined in order to limit the solution space for the high-resolution reconstruction stage. In this way, the goal is to find the intersection of the convex sets. Such work is [13] where noise is not considered, however, the solution is not unique and depends on the initial guess. These constitute the most significant disadvantages of POCS along with slow convergence. On the other hand, the advantages include simplicity, generality in the choice of the observation model and convenient inclusion of prior knowledge.

Since most quadratic minimization techniques generate overly smooth images, Farsiu et al. [14] use the L_1 norm both for regularization and data fusion in order to achieve better edge preservation. Furthermore, they demonstrate that L_1 norm minimization can be implemented as median estimation, while proposing the total variation method both for deblurring and denoising. In [15] the gradient constraint method which is based on Taylor series expansion is used, to achieve computational effectiveness. Whereas, the authors of [16] present an innovative idea, the subpixel shifts, which are used extensively in multi-image super-resolution. Other works which aim to deal with the computational problems that appear in super-resolution are [17], where the Tikhonov-regularized problem is solved by the conjugate gradient method, and [18] which presents a very significant theoretical result. The authors find the theoretical and practical performance bounds of super-resolution algorithms under various assumptions. The results indicate that, reconstruction-based algorithms are not favorable when the magnification factors are large or the number of the input images is not enough to constrain the solution. Thus, the reconstructed high-resolution image may lack important high-frequency details [19].

This problem led to the development of example-based super-resolution, which was introduced by [20, 21, 22]. As mentioned above, example-based super-resolution learns correspondences between low- and high-resolution image patches. [20] captures this cooccurence (between low- and high-resolution image patches) with the use of a Markov random field solved by belief propagation. [21] introduced the face hallucination problem. All the previously mentioned methods are dealing with generic image super-resolution, whereas [21] deals with human faces. [22] introduces a fast and simple one-pass algorithm which results in resolution independence in image-based representations. [23] proposes a method which is based on the observation that in a natural image, patches tend to redundantly recur within and across different scales of the given low-resolution image. In [24] the authors use the locally linear embedding technique from manifold learning to map the local geometry of the low-resolution patch space to the high-resolution patch space, in order to generate a high-resolution patch as a linear combination of neighbors. Thus, they assume similarity between the two manifolds in the high-resolution patch space and the low-resolution patch space. The disadvantage of this method is that in reconstruction problems a

fixed number of neighbors often leads to blurring results because of overfitting or underfitting.

In this thesis, the problem of super-resolution by sparse representation is demonstrated. Sparse representation is used successfully in many problems of image processing and the results indicate that it provides effectiveness and robustness to noise. Such method is the method of Yang, Wright, Huang and Ma [25] which seeks a sparse representation for each patch of the low-resolution input, and then uses the coefficients of this representation to generate the high-resolution output. This method uses two dictionaries, one for the low- and one for the high-resolution patches, which are trained simultaneously in order to ensure the similarity of the sparse representations between the low- and high-resolution patch pairs. The low-resolution input is upsampled before the training of the dictionaries and during the reconstruction process. A different approach was adopted in the frame of this thesis. The low-resolution image patches are directly used to obtain the two coupled dictionaries and then the original low- resolution patches are used in order to generate the high-resolution image. The method was applied and tested on frontal views of human faces. Furthermore, the training of the two dictionaries is discussed extensively.

In the next chapter, the problem of learning the dictionary pair in order to maintain the correspondence between the low-resolution and high-resolution patches is discussed. Chapter 3 presents the quantitative description of the formulation and reconstruction model. In chapter 4 the results obtained by the image super-resolution based upon sparse representation new approach are demonstrated in comparison with other methods. Finally, in chapter 5 the conclusions that emerged from the study of the image super-resolution problem are discussed and suggestions for future research are made.

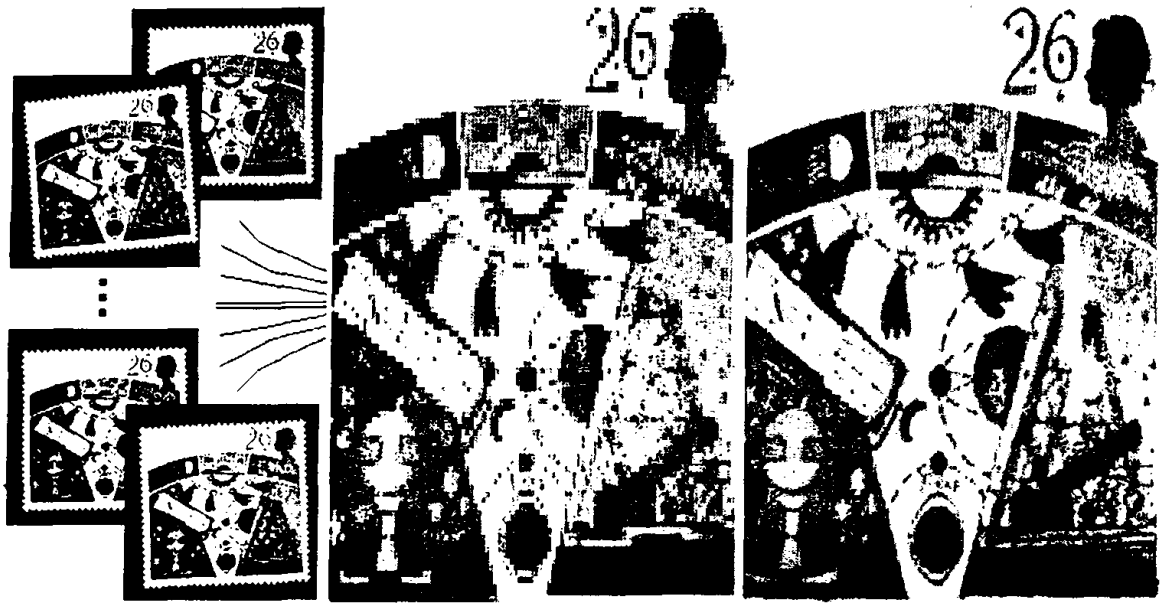


Fig. 1.1: Multi-image super-resolution. Multiple slightly different low-resolution images of the same object (left), one of the low-resolution images zoomed (center) in order to be compared with the resulting high-resolution image (right).

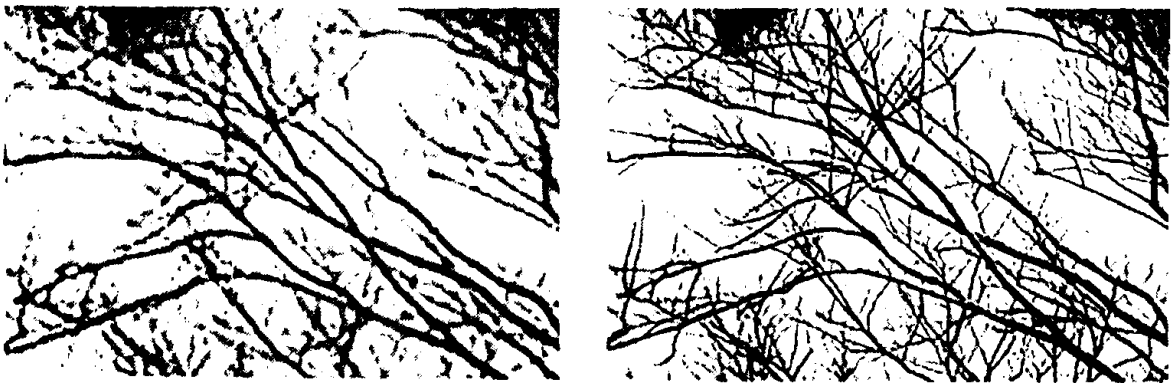


Fig. 1.2: Single-image super-resolution. The low-resolution input (left) and the high-resolution output (right).

Fig. 1.3: Face hallucination. A low-resolution image captured by a surveillance camera (left) and its super-resolved version (right).

CHAPTER 2. DICTIONARIES AND SPARSE REPRESENTATIONS

Sparse representation or sparse approximation or sparse decomposition is the problem of estimating a sparse vector which satisfies a linear system of equations given the observed data and a basis matrix, the so-called dictionary. Sparse representation methods are used in various applications such as image processing, audio processing, medical imaging devices, satellite imaging, visual electronics and document analysis. The sparse representation of signals has gained a lot of interest in recent years because of the observation that time signals can be well-represented by a small number of non-zero coefficients with respect to a suitable dictionary. Therefore, the defining characteristics of the sparse representation problem are an input signal, which is approximated by a linear combination of elementary signals called atoms or codewords, and a preference for sparse linear combinations which is imposed by penalizing non-zero coefficients. A set of atoms composes an overcomplete dictionary which is formed so that the number of atoms exceeds the dimension of the signal, in order to make feasible the representation of any signal by more than one combination of different atoms.

Given a signal and an overcomplete dictionary the goal of the sparse approximation problem is to find the smallest set of atoms from the dictionary to represent the signal. This problem, leads to an additive problem which is obtaining the appropriate

dictionary. In this chapter the construction of dictionaries which is of great importance in image processing and consequently in super-resolution problems will be discussed thoroughly. Furthermore, the dictionary learning technique which was used in the frame of this thesis in order to generate two coupled dictionaries D_l and D_h for the low- and high- resolution patches respectively is also analyzed.

2.1. Sparse Representation: A Closer Look

Let's consider the basic model for sparse representation which is defines as:

$$x = D\alpha \tag{2.1}$$

where $D \in \mathbb{R}^{nxK}$ is an overcomplete dictionary of K atoms (K > n) and $x \in \mathbb{R}^n$ is a signal represented as a sparse linear combination with respect to D. $\alpha \in \mathbb{R}^K$ is the signal which needs to be estimated subject to the constraint that it is sparse, which means that it has very few non-zero entries. It is remarkable that such problems start with observed data in high-dimensional space (n) and find signals which are organized in a lower-dimensional subspace (<<n). The solution to the above problem is the one with the fewest number of non-zero coefficients. Mathematically expressed as:

$$\min_{\alpha} \|\alpha\|_0 \text{ s.t. } x = D\alpha \tag{2.2}$$

or

$$\min_{\alpha} \|\alpha\|_0 \text{ s.t. } \|x - D\alpha\|_2 \leq \epsilon \tag{2.3}$$

where $\|\alpha\|_0$ is the l_0 pseudo-norm which counts the number of non-zero components of α. Although this optimization problem is NP-hard in general, a lot of methods have been proposed for finding approximating solutions. Some of these solve the sparse approximation problem iteratively by processing one coefficient at a time, such as Matching Pursuit (MP) and Orthogonal Matching Pursuit (OMP), and some process all the coefficients simultaneously, such as Basis Pursuit (BP) and Focal

Underdetermined System Solver family of algorithms (FOCUSS).

A first observation, for the previously presented basic model for spare representation, is that any atoms in the dictionary can be picked and secondly, the problem is defined for only a single point x and its noisy observation. In the structured sparsity model groups of atoms are picked, instead of picking atoms individually, and these groups can be overlapping and of varying size. In this case, the objective is to represent x so that it is sparse in the number of groups selected. In the collaborative sparse coding model, more than one observation of the same point is available and the data fitting error is defined as the sum of the l_2-norm for all points.

2.1.1. Matching Pursuit

Matching Pursuit is a greedy algorithm which processes one coefficient at a time in order to find the best sparse representation of a signal with respect to an overcomplete dictionary D. The algorithm, at each iteration, finds a basis vector in the dictionary D that maximizes the correlation with the residual signal or else that has the maximal projection onto the residual signal. Then the residual signal and the coefficients are recomputed by using the existing coefficients to project the residual onto the dictionary D. For a given signal $x \in \mathbb{R}^n$ and a dictionary D of K atoms, $\{d_k\} = \{d_1,d_2,\ldots,d_K\}$, if R_n is the residual signal and α_n the coefficients then the algorithm can be summarized as:

1. Initialize R and the index n:

$$R_1=x,\ n=1.$$

2. Repeat:
   a) find the index of the atom which maximizes the inner product of the dictionary atoms with the signal by:

$$i_n = \arg\max_{w} |\langle R_n, d_w \rangle|$$

b) Update the residual and the coefficients by:

$$\alpha_n = \langle R_n, d_w \rangle;$$
$$R_{n+1} = R_n - \alpha_n d_w;$$
$$n = n+1;$$

3. Until stop condition.

The stop condition can be a predetermined number of iterations, a predetermined number of selected atoms or a constraint on the residual signal, such as $\|R_n\| <$ threshold. The main problem of the Matching Pursuit algorithm is the computational cost and the fact that an atom can be selected multiple times.

2.1.2. Orthogonal Matching Pursuit

Orthogonal Matching Pursuit is also a greedy algorithm which is very similar to the aforementioned Matching Pursuit algorithm. The difference is that OMP deals with the problem of MP by avoiding picking an atom which has already been picked in a previous iteration. This is done by updating, at each iteration, an active set of atoms which have already been picked. The residual signal is recomputed by projecting it onto a linear combination of the atoms in the active set. The algorithm can be summarized as:

1. Initialize R and the index n:

$$R_0 = x, n=1.$$

2. Repeat:
    1. select the index of the next atom which maximizes the inner product of the dictionary atoms with the signal by:

$$i_n = \arg\max_{w} |\langle R_{n-1}, d_w \rangle|$$

    2. Update the current approximation by:

$$x_n = \arg\min_{x_n} \|x - x_n\|_2^2$$

such that $x_n \in \text{span}\{di_1, di_2, \ldots, di_n\}$

3. Update the residual signal by:

$$R_n = x - x_n;$$

3. Until stop condition.

The stop condition can be the same as in Matching Pursuit, a predetermined number of iterations, a predetermined number of selected atoms or a constraint on the residual signal, such as $\|R_n\| <$ threshold. The Orthogonal Matching Pursuit algorithm leads to better results than MP, because each atom is picked only once but the disadvantage is that this requires more computation.

2.1.3. Basis Pursuit

Both Orthogonal Matching Pursuit and Matching Pursuit algorithms solve the l_0-norm version of the sparse representation problem as it is described in (2.2) and (2.3), whereas, Basis Pursuit solves the l_1-norm version of the problem, which means that (2.2) and (2.3) are reformulated as:

$$\min_{\alpha} \|\alpha\|_1 \text{ s.t. } x = D\alpha \tag{2.4}$$

or

$$\min_{\alpha} \|\alpha\|_1 \text{ s.t. } \|x - D\alpha\|_2 \leq \epsilon \tag{2.5}$$

The problem of (2.4) can be solved efficiently through linear programming while the problem of (2.5) is a Quadratically Constrained Quadratic Programming which is ready to be solved in many optimization packages. Furthermore, Basis Pursuit processes all the coefficients simultaneously unlike Orthogonal Matching Pursuit and Matching Pursuit.

2.1.4. Focal Underdetermined System Solver

Focal Underdetermined System Solver (FOCUSS) is an algorithm that solves the sparse representation problem presented in (2.2) and (2.3) by using the l_p-norm instead of the l_0-norm. Thus, the problem of (2.2) and (2.3) can be rewritten as:

$$\min_{\alpha} \|\alpha\|_p \text{ s.t. } x = D\alpha \tag{2.6}$$

or

$$\min_{\alpha} \|\alpha\|_p \text{ s.t. } \|x - D\alpha\|_2 \leq \epsilon \tag{2.7}$$

where $p \leq 1$.

2.2. Dictionaries

In image processing the main problem is to find the sparsest representation of a signal. This sparse representation problem assumes that the dictionary is known; this assumption leads to another problem, which is obtaining the appropriate dictionary. Therefore, given a set of signals which have a sparse representation over an unknown dictionary the goal of the dictionary learning problem is to find the dictionary. In other words, the goal is to obtain the dictionary that generates sparse representations for the training signals.

Although pre-defined dictionaries have been widely used in many applications, such as discrete cosine transform (DCT), discrete Fourier transform (DFT) and wavelets dictionaries (which are used in image compression), the dictionaries which are learned directly from the data have been successfully used in applications where the pre-defined dictionaries are not applicable, and provided a better adaptation of the dictionaries. A great variety of dictionary learning algorithms have been proposed, with most of them having a common characteristic. The optimization process iterates

between two steps, the sparse approximation and the dictionary update. More specifically, starting with an initial dictionary the algorithms find sparse approximations of the set of training signals while keeping the dictionary fixed in step one, and in step two the sparse coefficients are kept fixed while the dictionary is optimized. Some of these algorithms are presented in the following lines.

2.2.1. Probabilistic Methods

Given independently and identically distributed (i.i.d.) data $X = \{x_1, x_2, \ldots, x_n\}$ assumed to be generated by the general model (11), a maximum likelihood estimate D_{ML} of the unknown dictionary D can be determined as [26, 27]:

$$D_{ML} = \arg \max_{\mathbf{D}} P(X;D). \tag{2.8}$$

This optimization problem requires the integration of the hidden, unobservable i.i.d source vectors $A = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ for the computation of P(X;D). The prior $P(\alpha)$, which is assumed to be known, is generally taken to be supergaussian thus, the integration is computationally unreasonable and approximations to this integration are performed, leading to an approximation of P(X;D) which is maximized with respect to X. Then, a better approximation to the integration is made and this process is iterated until the convergence of the estimate of the dictionary D [26]. Therefore, the method of [26] can be summarized as the iteration between the two following steps:

Step 1: calculate the coefficients α_i using a simple gradient descent procedure.

Step 2: update the dictionary D by using:

$$D^{(n+1)} = D^{(n)} - \eta \sum_{i=1}^{N} (D^{(n)}\alpha_i - x_i)\alpha_i^T . \tag{2.9}$$

The authors of [27] handle the integration problem by using a gaussian function in order to approximate the integration. On the other hand, the authors of [26] also propose the maximum a posteriori (MAP) dictionary learning algorithm which maximizes the posterior probability that a given signal can be represented by a

dictionary and the sparse coefficients P(D;X), instead of maximizing the likelihood function P(X;D) [28].

2.2.2. The Method of Optimal Directions (MOD)

The method of optimal directions (MOD) was proposed by [29] for dictionary training. This method defines the errors as:

$$e_i = x_i - D\alpha_i \tag{2.10}$$

assuming that the sparse representation of the data is known. The overall representation mean square error can be calculated by:

$$\|E\|_F^2 = \| [e_1,e_2,\ldots,e_N] \|_F^2 = \| X - DZ \|_F^2 \tag{2.11}$$

where the matrix X includes all the training vectors x_i as columns, the matrix Z includes the representation coefficients α_i and $\|E\|_F$ is the Frobenius norm, which is defined as $\|E\|_F = \sqrt{\sum_i \sum_j E(i,j)^2}$.

The MOD algorithm, at the first step finds a sparse representation using OMP or FOCUSS and then at the second step, assuming that Z is fixed, updates the dictionary D so that the error in (2.11) is minimized. The update of the dictionary D is done by using:

$$D^{(n+1)} = XZ^T \cdot (ZZ^T)^{-1} . \tag{2.12}$$

The MOD method is an improvement of [26] which is more efficient and generates better dictionaries D, due to the fact that MOD assumes that the coefficients are known at each iteration and generates the best possible dictionary, whereas the maximum likelihood method of [26] only gets closer to the best current solution and then calculates the coefficients.

2.2.3. The K-SVD Algorithm

The K-SVD algorithm which is a direct generalization of the K-Means was proposed by [30]. When this algorithm is forced to work with one atom per signal, it trains a dictionary for the Gain-Shape VQ, and when it is forced to have a unit coefficient for this atom it exactly reproduces the K-Means algorithm.

At the first step of the algorithm, the sparse representation step, the best possible dictionary for the sparse representation of the example set X is found by solving the minimization problem:

$$\min_{\mathbf{Z}} (\|X - DZ\|_F^2) \text{ s.t. } \|\alpha_i\|_0 \leq T_0. \tag{2.13}$$

In this step, the dictionary D is fixed while the coefficient matrix Z is updated. Any of the sparse approximation algorithms can be used for the solution of (2.13).

At the second step, the dictionary update step, the K-SVD algorithm updates one column at a time, fixing all columns in D except one, d_k, and finding a new column $\tilde{d}_k$ and new values for its coefficients that best minimize the mean square error as defined in (2.11):

$$\|E\|_F^2 = \| [e_1, e_2, \ldots, e_N] \|_F^2 = \| X - DZ \|_F^2.$$

This is the main difference of the K-SVD algorithm from the previously described methods, which keep Z fixed while finding a better dictionary D. In K-SVD while the columns of D are changing sequentially, the corresponding coefficients are changing as well. In this way, the convergence of the algorithm accelerates.

2.2.4. Recursive Least Squares Dictionary Learning Algorithm (RLS-DLA)

The recursive least squares dictionary learning method was proposed by [31]. The goal is to find a sparse representation of the training set which minimizes the total error as much as possible, that is minimizes the sum of squared errors. Thus, the

minimization problem is the same as in K-SVD and MOD method described earlier with the error given by the equation (2.11):

$$\|E\|_F^2 = \| [e_1,e_2,\ldots,e_N] \|_F^2 = \| X - DZ \|_F^2 .$$

As previously mentioned, the most dictionary learning algorithms iterate between two steps, the sparse approximation step and the update of the dictionary step. In the first step, the algorithm finds Z while keeping D fixed, and in the second step the dictionary D is found while keeping Z fixed. The strategy of RLS-DLA is different.

In the first step, a sparseness constraint is employed in order to compute the sparse coefficients α. Thus, the sparse approximation problem can be formulated as:

$$\alpha = \arg\min_{\alpha} \|x - D\alpha\|_2^2 \quad \text{s.t.} \quad \|\alpha\|_0 \leq s \tag{2.14}$$

where s is the number of the non-zero elements in each column. This optimization problem can be solved by using MP, OMP, BP or FOCUSS.

In the second step, the goal is to update the dictionary D. This is done by minimizing (2.11) and the optimization problem can be written as:

$$D = \arg\min_{D} (\|X - DZ\|_F^2) \tag{2.15}$$

while the Least Squares solution is the same as in the MOD method and is given by the equation (2.12):

$$D^{(n+1)} = XZ^T \cdot (ZZ^T)^{-1} .$$

The authors of the RLS-DLA method assume that the solution for the first i training vectors is known and they try to find the new solution when the next training vector is included. Therefore, the Least Squares solution for the dictionary for these i training vectors can be formulated as:

$$D_i = X_i Z_i^T \cdot (Z_i Z_i^T)^{-1} . \quad (2.16)$$

The RLS-DLA algorithm updates the dictionary continuously as each new training vector is processed, whereas the previously mentioned dictionary learning algorithms update the dictionary after a batch of training vectors has been processed, usually using the whole set of training vectors as one batch. The results indicate that this algorithm has very good convergence properties and the ability to use very large ttaining sets, which results in a general dictionary for the used signal class instead of a specialized dictionary for the particular training set used. Furthermore, the RLS-DLA algorithm is superior to the above mentioned methods both in representation ability of the training set and in the ability of reconstruction of a true underlying dictionary [31].

2.2.5. Simultaneous Codeword Optimization (SimCO)

At the aforementioned K-SVD algorithm the main idea is that in the dictionary update step one column (that is atom) of the dictionary D is updated at a time while the corresponding row of the sparse coefficient matrix Z is also updated. In the Simultaneous codeword optimization (SimCO) method proposed by [32], all the atoms of the dictionary D and the corresponding non-zero coefficients in Z are updated simultaneously.

In the first step of the algorithm, the sparse approximation step, the optimization problem is the same as in the most dictionary learning algorithms, that is the minimization of (2.11):

$$\| X - DZ \|_F^2 .$$

In the second step, the dictionary update step, the authors of the SimCO algorithm in order to update all the atoms of the dictionary D and the corresponding coefficients simultaneously, employ l_2-norm constraints on the columns of D:

$$D = \{D \in \mathbb{R}^{nxK} : \|D_i\|_2 = 1, \forall i \in K\}, \quad (2.17)$$

and the optimization problem can be written as:

$$\min_{\mathbf{D},\mathbf{Z}} \ (\|X - DZ\|_F^2) \quad Z(\Omega) \quad (2.18)$$

where $Z(\Omega)$ are the locations of the non-zero coefficients in Z. What is remarkable about this optimization problem is that the Z that minimizes (2.18) changes as D changes. An update in D is followed by an update of the corresponding optimal Z. The authors also propose a generalization of SimCO in order to update an arbitrary subset of atoms and the corresponding coefficients, instead of updating all the atoms of D simultaneously. The bottom line is that SimCo is different from other dictionary learning methods only as far as the dictionary update step is concerned. When the sparse coefficient matrix Z is fixed, the optimization problem is similar to the MOD method, and when only one column of the dictionary D is selected for update, the optimization is similar to the K-SVD algorithm. However, the main difference is the l_2-norm constraints on the columns of D.

2.2.6. Greedy Adaptive Dictionary Algorithm (GAD)

The greedy adaptive dictionary algorithm (GAD) divides the input data into blocks [33]. Thus, if the input signal is denoted as $x(n) = [x(1), x(2),\ldots,x(N)]$ and the blocks are denoted as $x_m(n) = [x_m(1), x_m(2),\ldots, x_m(K)]$, where K are the atoms of the dictionary, then the blocks construct a new matrix X defined as:

$$X(n) = [x_1(n), x_2(n),\ldots, x_M(n)] \quad (2.19)$$

where the m-th column is represented by the block $x_m(n)$ and the columns are l_1-normalized. The GAD algorithm adaptively learns a dictionary by sequentially extracting the columns of the matrix X.

If $R^j = [r_1^j, r_2^j, \ldots, r_M^j]$ is the residual matrix with r_m^j being a M-dimensional residual column vector, then the initial residual matrix for the algorithm is:

$$R^0 = X. \tag{2.20}$$

The GAD algorithm at each iteration $j \in \{1, 2, \ldots, K\}$ finds the block with the highest l_2-norm, r_m^j, which becomes a dictionary element and iteratively defines a residual matrix. Therefore, the iterative steps of the algorithm can be summarized as:

1. Find the index m of the next block:

$$m = \arg\max_m \| r_m^j \|_2. \tag{2.21}$$

2. Include this block in the dictionary as the j-th dictionary element d^j.

3. Evaluate the coefficients by computing the inner product between the residual vector r_m^j and the dictionary atom d^j of the previous step:

$$\alpha_m^j = \langle r_m^j, d^j \rangle. \tag{2.22}$$

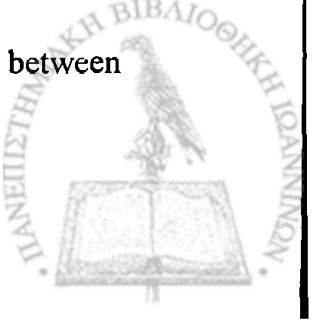
4. Compute the new residual, by removing the component along the chosen atom, for each element m in $r_m^j(n)$:

$$r_m^{j+1} = r_m^j - \alpha_m^j d^j. \tag{2.23}$$

Then the corresponding column of the residual matrix R^j is set to zero because the whole atom is removed. In this way, it is ensured that the transform is orthogonal.

2.2.7. Efficient Sparse Coding Algorithms

The authors of [34] propose efficient sparse coding algorithms which iterate between



the two regular sparse representation steps, sparse approximation step and dictionary update step, and alternatively solve two convex optimization problems. The first is a L_1-regularized least squares problem and the second is an L_2-constrained least squares problem.

In the first step, the sparse approximation step, the optimization is the same as in the most dictionary learning algorithms with an l_1 penalty added over the coefficients Z. Thus, the optimization problem can be written as:

$$\min_{\mathbf{Z}} \|X - DZ\|^2 + \lambda\|Z\|_1 \tag{2.24}$$

where λ is a constant. The authors propose solving this problem by optimizing over each coefficient α of the coefficient matrix Z individually while keeping the dictionary D fixed. The algorithm tries to find the signs of the coefficients by maintaining an active set of potentially nonzero coefficients and their corresponding signs, all the other coefficients must be zero, and systematically searches for the optimal active set and coefficient signs. Each time, a current guess for the active set and the signs is available and the analytical solution α_{new} to the resulting problem is computed. Then, the active set and the signs are updated using an efficient discrete line search between the current solution and α_{new}. The authors named the algorithm which solves the first step of the dictionary learning problem, the feature-sign search algorithm.

In the second step, the dictionary update step, the goal is to update D while keeping the coefficient matrix Z fixed. The optimization problem is the same as it was defined in (2.15):

$$D = \arg\min_{\mathbf{D}} (\|X - DZ\|_F^2)$$

with l_2-norm constraints employed on the columns of the dictionary D. Therefore, the problem is formulated as:

$$D = \arg\min_{\mathbf{D}} (\|X - DZ\|_F^2) \quad \text{s.t.} \quad \|D_i\|^2 \leq c,\ i=1, 2,\ldots, K \tag{2.25}$$

where c is a constant. This quadratically constrained least squares problem can be solved using the Lagrange dual. If $D(\lambda)$ denotes the Lagrange dual with λ representing the dual variables, then the optimization of $D(\lambda)$ can be achieved by using Newton's method or conjugate gradient. After the maximization of the Lagrange dual the optimal dictionary D can be obtained by computing:

$$D^T = (ZZ^T + \Lambda)^{-1}(XZ^T)^T \tag{2.26}$$

where $\Lambda = \text{diag}(\lambda)$.

2.3. Learning the Dictionary

2.3.1. Single Dictionary Training

The dictionary learning process starts with a given set of signals which have a sparse representation over an unknown dictionary and tries to find the dictionary. In other words, the goal of the dictionary learning process is to obtain the dictionary that generates sparse representations for the training signals. The optimization problem which was used for the experiments presented in the next chapter is formulated as [34]:

$$D = \arg\min_{\mathbf{D},\mathbf{Z}} \|X - DZ\|_2^2 + \lambda\|Z\|_1 \quad \text{s.t.} \quad \|D_i\|_2^2 \leq 1,\ i=1,2,\ldots,K \tag{2.27}$$

where the matrix X is the set of training vectors, the matrix D is the dictionary and matrix Z includes the sparse representation coefficients. The l_1-norm, $\|Z\|_1$, enforces the sparsity, whereas the l_2-norm constraints on the columns of D remove the scaling ambiguity. The equation (2.27) is not convex in both D and Z, but it is convex in one of them with the other fixed. This is known as a biconvex optimization problem and can be solved by iterating between the two convex optimization problems contained

in it. Therefore, the dictionary learning process iterates between the two regular steps, the sparse approximation step and the dictionary update step. In the first step, starting with an initial dictionary the algorithm finds sparse approximations of the set of training signals while keeping the dictionary fixed, while in the second step the sparse coefficients are kept fixed while the dictionary is optimized. The algorithm of [34] can be summarized as:

1. Initialize the dictionary D with a Gaussian random matrix, with each column unit normalized.

2. Keep the dictionary D fixed, and update Z by

$$Z = \arg\min_{\mathbf{Z}} \|X - DZ\|_2^2 + \lambda\|Z\|_1 \qquad (2.28)$$

   which is a linear programming problem and can be solved efficiently by existing solvers.

3. Keep the sparse coefficient matrix Z fixed, and update the dictionary D by

$$D = \arg\min_{\mathbf{D}} \|X - DZ\|_2^2 \text{ s.t. } \|D_i\|_2^2 \leq 1,\ i=1,2,\ldots,K \qquad (2.29)$$

   which is a Quadratically Constrained Quadratic Programming that can be solved efficiently by existing solvers. Such is the package developed by [34], and also many other optimization packages.

4. Iterate between 2 and 3 until convergence.

2.3.2. Joint Dictionary Training

The joint dictionary learning strategy is almost the same with the single dictionary learning strategy. As it is mentioned in the previous chapter and as it will be mentioned throughout the next chapter, the goal of the learning-based super-

resolution methods is to learn two coupled dictionaries, one for the low-resolution image patches (D_l) and one for the high-resolution ones (D_h). The two dictionaries are trained simultaneously in order to ensure that the sparse representations between the low-resolution and high-resolution image patch pairs are similar, with respect to D_l and D_h. The learning process starts with a given set of image patch pairs, $P = \{X^h, Y^l\}$, which are sampled from the training images. With $X^h = \{x_1, x_2, \ldots, x_n\}$ the set of the sampled high-resolution image patches is represented and with $Y^l = \{y_1, y_2, \ldots, y_n\}$ the set of the corresponding low-resolution image patches is represented.

The optimization problems for the high-resolution image patches and the low-resolution ones can be respectively written as:

$$D_h = \arg\min_{D_h, Z} \|X^h - D_h Z\|_2^2 + \lambda \|Z\|_1 \tag{2.30}$$

and

$$D_l = \arg\min_{D_l, Z} \|Y^l - D_l Z\|_2^2 + \lambda \|Z\|_1 \,. \tag{2.31}$$

These two objectives are combined, according to [25], forcing the high- and low-resolution representations to share the same codes. Thus, the optimization problem is formulated as:

$$\min_{D_h, D_l, Z} \frac{1}{N} \|X^h - D_h Z\|_2^2 + \frac{1}{M} \|Y^l - D_l Z\|_2^2 + \lambda \left(\frac{1}{N} + \frac{1}{M}\right) \|Z\|_1 \tag{2.32}$$

where N is the dimension of the high-resolution image patches in vector form and M is the dimension of the corresponding low-resolution image patches in vector form. The two cost terms of (2.30) and (2.31) are balanced by the use of 1/N and 1/M. The optimization problem in (2.32) can be written as:

$$\min_{D_h, D_l, Z} \|X_c - D_c Z\|_2^2 + \tilde{\lambda} \|Z\|_1 \tag{2.33}$$

where

$$X_c = \begin{bmatrix} \frac{1}{\sqrt{N}} X^h \\ \frac{1}{\sqrt{M}} Y^l \end{bmatrix}, \quad D_c = \begin{bmatrix} \frac{1}{\sqrt{N}} D_h \\ \frac{1}{\sqrt{M}} D_l \end{bmatrix}, \quad \tilde{\lambda} = \lambda \left(\frac{1}{N} + \frac{1}{M} \right). \tag{2.34}$$

The optimization problem in (2.33) is now the same with the one in (2.27). This means that the two coupled dictionaries can be trained with the single dictionary method described in the previous section. More specifically, the steps of the algorithm are:

1 Initialize the two dictionaries D_h and D_l with a Gaussian random matrix, with each column unit normalized. Combine the two dictionaries to form D_c.

2 Keep the dictionary D_c fixed, and update Z by

$$Z = \arg\min_{\mathbf{Z}} \|X_c - D_c Z\|_2^2 + \tilde{\lambda}\|Z\|_1 \tag{2.35}$$

which is a linear programming problem and can be solved efficiently by existing solvers.

3 Keep the sparse coefficient matrix Z fixed, and update the dictionary D_c by

$$D_c = \arg\min_{\mathbf{D_c}} \|X_c - D_c Z\|_2^2 \text{ s.t. } \|D_i\|_2^2 \le 1, \; i=1,2,\ldots,K \tag{2.36}$$

which is a Quadratically Constrained Quadratic Programming that can be solved efficiently by existing solvers. Such is the package developed by [34], and also many other optimization packages.

4 Iterate between 2 and 3 until convergence.

CHAPTER 3. IMAGE SUPER-RESOLUTION

Super-resolution means resolution enhancement of an imaging system. The sensor and the lens in every digital imaging system result in optical blur and limitations on the highest spatial frequency the given sensor can record. Although it should be easy to "cure" optical blur by simply applying an inverse sharpening, yet, this leads to degradations (caused by the sensor). Other reasons that explain why the reconstruction of a perfect high-resolution image is impossible are: (i) the sensor noise which degrades the image quality and reduces the ability to recover the details which are lost by noise, (ii) uncertainty about the real offsets of the images, which means that the precise camera position and orientation in space are not known during the super-resolution process and thus, need to be estimated by the low-resolution images (introducing errors), and (iii) the diffraction limit because optical systems have fundamental limits on resolution where two close subjects cannot be resolved one from another.

Super-resolution tries to overcome these problems via various methods. The super-resolution methods proposed in literature are of great variety and they can be broadly divided into two categories. The first category includes the classical multi-image super-resolution techniques, which use the subpixel misalignments between several low-resolution images of the same scene in order to infer the high-resolution image. The second category is composed of the learning-based super-resolution techniques.

The learning-based methods learn the correspondences between low- and high-resolution image patch pairs and then apply them to the given low-resolution image to recover its most likely high-resolution version. This section focuses on a learning-based method which tries to recover the high-resolution version of a given low-resolution image by using sparse representation. Furthermore, a global reconstruction constraint is employed in order to ensure that the recovered high-resolution image is consistent with its low-resolution counterpart. The mathematical description of this problem is presented in the following lines.

3.1. Mathematical Description

3.1.1. Basic Model

A great variety of super-resolution methods can be found in literature. However, there is a standard underlying model that relates the high-resolution and the low-resolution image. If Y is the given low resolution image and X the original high-resolution image, then the observation model is:

$$Y = SHX + \eta \tag{3.1}$$

where the matrix S is the downsampling operator, matrix H is the blurring operator, and η is the noise in the generation of Y from X.

3.1.2. Basic Model by Sparse Representation

As mentioned above, a compact representation is learned in order to capture the cooccurrence prior between the low- and high-resolution image patch pairs. Let $D \in \mathbb{R}^{nxK}$ be an overcomplete dictionary of K atoms ($K > n$), and suppose a signal $x \in \mathbb{R}^n$ can be represented as a sparse linear combination with respect to D:

$$x = D\alpha_0 \tag{3.2}$$

where $\alpha_0 \in \mathbb{R}^K$ is a vector with very few ($<< n$) nonzero entries, and signal x is a high-resolution image patch.

If y is the low-resolution counterpart of x then:

$$y = Lx = LD\alpha_0 \tag{3.3}$$

where $L \in \mathbb{R}^{kxn}$ with $k < n$ is a projection matrix.

The dictionary D is overcomplete, as mentioned above, this means that the equations (3.2) and (3.3) are underdetermined for the unknown coefficients α. However, the sparsest solution α_0 to these equations will be unique, under mild condition.

In this thesis, instead of learning a single dictionary D, two coupled dictionaries are learned, one for the low-resolution image patches (D_l) and one for the high-resolution ones (D_h). To ensure that the sparse representations between the low-resolution and high-resolution image patch pairs are similar, with respect to D_l and D_h, the two dictionaries are trained simultaneously. This leads to the idea that, the sparse representation for each low-resolution image patch can be applied with the high-resolution image patch dictionary to generate the corresponding high-resolution image patch [25].

3.2. Description of Image Super-resolution via Sparsity

In this section, the reconstruction method of a high-resolution image which was proposed by Yang, Wright, Huang and Ma [25] is presented. Their approach is a learning-based technique. Such techniques, attempt to capture the relationship between the low-resolution patches and their corresponding high-resolution versions from the training data. This relationship is learned by jointly training two dictionaries, D_l and D_h, so as to have the same sparse representations for each high-resolution and

low-resolution image patch pairs. Therefore, if a low-resolution patch has a representation in D_l, its high-resolution version can be recovered by the same representation in D_h. In practice, there are many high-resolution patches that result in the same low-resolution patch after blurring and downsampling. Thus, the solution space needs to be restricted by regularizing the super-resolution problem. This can be efficiently achieved by the l^1-norm, as long as it is highly likely that any given low-resolution patch has a sparse representation in D_l.

The basic idea proposed by [25] can be summarized in the following equation:

$$x \approx D_h\alpha, \quad \alpha \in \mathbb{R}^K \text{ with } \|\alpha\|_0 << K \tag{3.4}$$

which indicates that the high-resolution patches x of the original image X can be represented as a sparse linear combination in a high-res dictionary D_h trained from high-resolution patches sampled from training images. The sparse representation α is found by representing the low-resolution patches y of the given image Y with respect to a low-resolution dictionary D_l, which is jointly trained with D_h. In simple words, the basic idea is to find a sparse representation for each low-resolution patch y of the input image Y, and then apply the coefficients α of this representation with the high-resolution dictionary D_h to generate the high-resolution patch x of the output X.

3.2.1. The Problem of the Sparsest Representation

The D_l dictionary is formed by low-resolution patches and the D_h dictionary is formed by the corresponding high-resolution patches. As mentioned above, for each low-resolution patch y of the input Y a sparse representation with respect to D_l is found. The corresponding high-resolution patch x of the output image X can be recovered by the same representation in the high-resolution dictionary D_h. The mathematical expression of this problem is:

$$\min \|\alpha\|_0 \text{ s.t. } D_l\alpha \approx y \quad \text{with } \|\alpha\|_0 << K \tag{3.5}$$

This optimization problem (3.5) is general because it contains a rough constraint. This constraint needs to be replaced by a more precise condition which will allow a bounded representation error. Thus, the optimization problem can be rewritten as:

$$\min \|\alpha\|_0 \text{ s.t. } \|D_l\alpha - y\|_2^2 \leq \epsilon \tag{3.6}$$

The optimization problem (3.6) is NP-hard. However, as long as the two dictionaries D_l and D_h are constructed so that it is possible for any given low-resolution patch to have a sparse representation in D_l, the l^1-norm can be used for the minimization:

$$\min \|\alpha\|_1 \text{ s.t. } \|D_l\alpha - y\|_2^2 \leq \epsilon \tag{3.7}$$

Equivalently, the constraint of (3.7) can become a penalty with the use of the Lagrange multipliers and the optimization problem (3.7) can be written as:

$$\min_{\alpha} \|D_l\alpha - y\|_2^2 + \lambda\|\alpha\|_1 \tag{3.8}$$

where λ is a constant that balances the trade-off between fitting the data perfectly and employing a sparse solution. The optimization problem (3.8) indicates that λ depends on how noisy the input data are. The noisier the data, the larger the value of λ should be, and the larger the value of λ, the smoother the result image gets.

The compatibility between adjacent patches cannot be ensured when (3.8) is solved individually for each patch. A lot of different ways have been proposed in order to ensure compatibility between adjacent patches. In [24] the authors simply average the values in the overlapped regions, which results in blurring effects. In [22] a fast and simple one-pass algorithm is used and the results show that it works almost as well as the use of a full Markov random field (MRF) model [20]. For solving (3.7) the idea of overlapping patches is adopted, and the patches are processed in raster-scan order in the image, that is from left to right and top to bottom. The cohesion in the overlapped area is enforced by modifying (3.7) to include one more term in the objective, which leads to the following formulation:

$$\min \|\alpha\|_1 \text{ s.t. } \|D_l\alpha - y\|_2^2 \leq \epsilon_1 \quad \|PD_h\alpha - w\|_2^2 \leq \epsilon_2 \tag{3.9}$$

where the matrix P extracts the overlapping region between the current desired high-resolution patch and the previously reconstructed high-resolution image, and the vector w contains the values of the previously reconstructed high-resolution image in the overlapping region. Equivalently, with the use of the Lagrange multipliers (3.9) can be rewritten as:

$$\min_{\alpha} \|\tilde{D}\alpha - \tilde{y}\|_2^2 + \lambda\|\alpha\|_1 \tag{3.10}$$

where $\tilde{D} = \begin{bmatrix} D_l \\ \beta P D_h \end{bmatrix}$ and $\tilde{y} = \begin{bmatrix} y \\ \beta w \end{bmatrix}$, and the constant β balances the tradeoff between matching the low-resolution input and finding a high-resolution patch that is compatible with its neighbors. In the experiments presented in this thesis β is always equal to 1, $\beta = 1$. Solving the optimization problem (3.10) will produce the optimal solution α^*, which will be used in the reconstruction of the high-resolution patch combined with the high-resolution dictionary D_h, that is:

$$x = D_h\alpha^*. \tag{3.11}$$

3.2.2. The Algorithm

In this section, the algorithm which describes the reconstruction of the high-resolution image of a given low-resolution input by using sparse representation according to the optimization problem described previously is presented. The input data of the problem are the two jointly trained dictionaries D_h and D_l, for the high- and low-resolution image patch pairs respectively and the given low-resolution image Y. The algorithm seeks the sparsest representation for each original low-resolution patch of the input image Y and then uses the coefficients of this representation to generate the high-resolution patch of the output X. The output of the algorithm is the super-resolved

image X.

Algorithm

---

1 Input: the two jointly trained dictionaries D_h and D_l, a low-resolution image Y.

2 For each 3x3 low-resolution patch y of Y, taken starting from the upper-left corner with one pixel overlap in each direction,

   a) Solve the optimization problem (3.10): $\min_\alpha \|\tilde{D}\alpha - \tilde{y}\|_2^2 + \lambda\|\alpha\|_1$.
   b) Generate the high-resolution patch x as described in (3.11): $x = D_h\alpha^*$. Put the reconstructed patch x into a high resolution image X.

3 End

4 Output: Super-resolution image X.

3.3. From Local Optimization to Global Optimization

The problem of super-resolution by sparse representation as described in the previous section, in (3.7) and (3.9), does not enforce exact equality between the low-resolution patch y and its reconstruction $D_l\alpha$. This can be a serious problem because along with noise, the aforementioned algorithm may lead to a high-resolution image X which will contain visible artifacts in the overlapping region; meaning that the super-resolved image X will not satisfy the basic super-resolution model as described in (3.1):

$$Y = SHX + \eta.$$

A global reconstruction constraint has to be employed to fix this incompatibility. Specifically, the authors of [25] propose as a solution the projection of the output X onto the solution space of SHX = Y. Thus, if the output X of the super-resolution algorithm is denoted by X_0 the optimal high-resolution image X* is computed by:

$$X^* = \arg\min_{X} \|SHX - Y\|_2^2 + c\|X - X_0\|_2^2. \quad (3.12)$$

The optimization problem (3.12) can be efficiently solved by the gradient descent method or the back-projection method. The update equations for these two iterative methods are respectively:

$$X_{t+1} = X_t + v\,[H^T S^T (Y - SHX_t) + c(X - X_0)] \quad (3.13)$$

where X_t is the estimate of the high-resolution image after the t-th iteration and v is the step size of the gradient descent method, and

$$X_{t+1} = X_t + ((Y - SHX_t) \uparrow s) * p \quad (3.14)$$

where $\uparrow s$ is an upsampling factor of s and p is a back-projection filter. The final high-resolution image of the super-resolution process is the one generated by the optimization problem (12), which is as close as possible to the super-resolved image X_0 generated by the algorithm described in the previous section and satisfies the basic super-resolution model. Therefore, the algorithm can be rewritten as:

Algorithm

1 Input: the two jointly trained dictionaries D_h and D_l, a low-resolution image Y.

2 For each 3x3 low-resolution patch y of Y, taken starting from the upper-left corner with one pixel overlap in each direction,

   a) Solve the optimization problem (3.10): $\min_\alpha \|\tilde{D}\alpha - \tilde{y}\|_2^2 + \lambda\|\alpha\|_1$.
   b) Generate the high-resolution patch x as described in (3.11): $x = D_h\alpha^*$. Put the reconstructed patch x into a high resolution image X_0.

3 End

4 Find the closest image to X_0 which satisfies the global reconstruction constraint by solving the optimization problem (3.12):

$$X^* = \arg\min_{X} \|SHX - Y\|_2^2 + c\|X - X_0\|_2^2.$$

5 Output: Super-resolution image X^*.

The super-resolution algorithm described in this chapter is a simple approach to a more general sparse representation problem, where the high-resolution image X is enforced to be precisely recovered by the sparse representation coefficients α. It can be improved further to deal with larger images X, meaning that the overall high-resolution image X can be used as a variable. Thus, the difference between X and the high-resolution image recovered by the sparse coefficients α can be penalized in order to yield outputs which are better in terms of respecting the reconstruction constraints but are not perfectly sparse. This is a global approach which can lead to an overall algorithm. This optimization problem can be formulated as:

$$X^* = \arg\min_{X,\alpha_{ij}} \left\{ \|SHX - Y\|_2^2 + \lambda \sum_{ij} \|\alpha_{ij}\|_0 + \gamma \sum_{ij} \|D_h\alpha_{ij} - P_{ij}X\|_2^2 + \tau\, \rho(X) \right\}. \quad (3.15)$$

where α_{ij} are the sparse representation coefficients for the (i,j) patch of the image X, the matrix P_{ij} extracts the (i,j) patch from X and the penalty function $\rho(X)$ adds prior knowledge about the high-resolution image. The first term enforces the proximity between the image Y and its denoised and unknown version X. The two following terms are the image prior and they are included in order to ensure that every patch of the image X has a sparse representation with bounded error.

This large optimization problem has a huge disadvantage which is the cost of the computational complexity. However, there are correspondences between the global approach described in (3.15) and the super-resolution algorithm of this chapter. The sparse representation coefficients α of (3.15) can be found by the minimization of the sum of the second and third term, where the l_0-norm of the second term is replaced by the l_1-norm and the third term is approximated by its low-resolution counterpart $\|D_l\alpha_{ij} - y_{ij}\|_2^2$. Therefore, the aforementioned algorithm can be considered as a

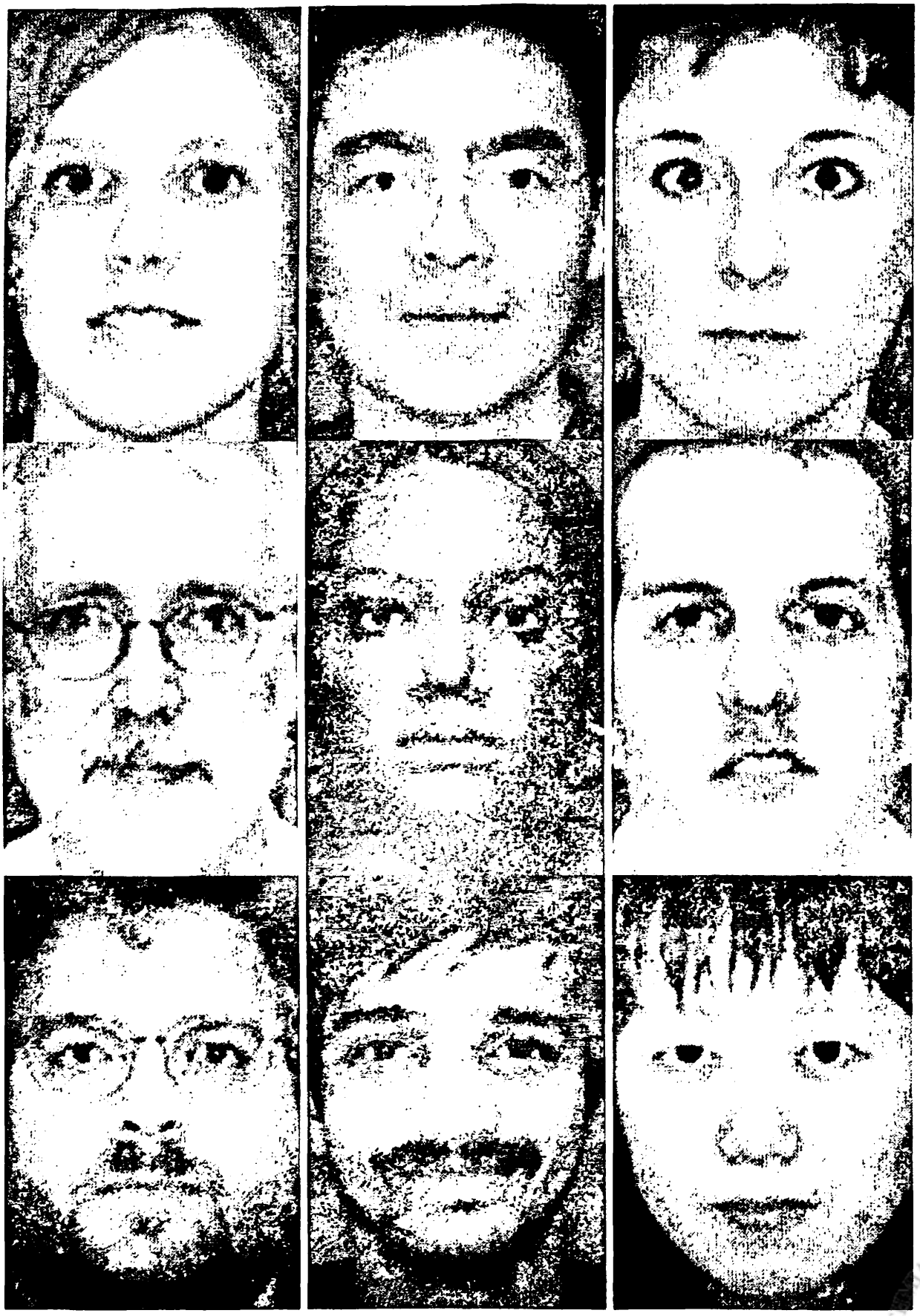
computationally efficient approximation to the large optimization problem (3.15). Furthermore, the third term of (3.15) is related with the $||X_0 - X||$ term of (3.12) because it penalizes the difference between the super-resolution image X and its reconstruction given by the sparse coefficients.

CHAPTER 4. EXPERIMENTAL RESULTS

In this chapter, some of the super-resolved images that were generated by the algorithm described in section 3.2 are presented. The algorithm was applied only on frontal views of human faces. Specifically, for the experiments the CMU MultiPIE database was used. It is important to mention that the low-resolution images which are used in the experiments are synthetic, which means that they were artificially generated from the high-resolution images of the database. The low-resolution images are processed by using 3x3 low-resolution patches with one pixel overlap between neighboring patches in each direction starting from the upper-left corner, while the high-resolution images are processed by using 6x6 high-resolution patches with two pixels overlap in order to preserve the correspondence between the low- and the high-resolution patches.

Fig. 4.1 shows some of the images included in the CMU MultiePIE database. These images which are high-resolution images were used as training images during the joint dictionary learning process in order to provide the high-resolution dictionary D_h for the high-resolution image patches. The same images were used in order to generate the low-resolution images by reducing their size. These downsampled low-resolution images were also used as training images during the dictionary learning phase to generate the low-resolution dictionary D_l for the low-resolution image patches. The specific images which are presented in fig. 4.1 are deliberately chosen, to show that the training faces are of both genders, different races, varying ages,

facial expressions, with accessories or not and specifically for hair or not.



: Some of the images included in the CMU MultiePIE database. Tl
solution images of both genders, different races, varying ages, varyin
ions, with accessories or not and with facial hair or not.

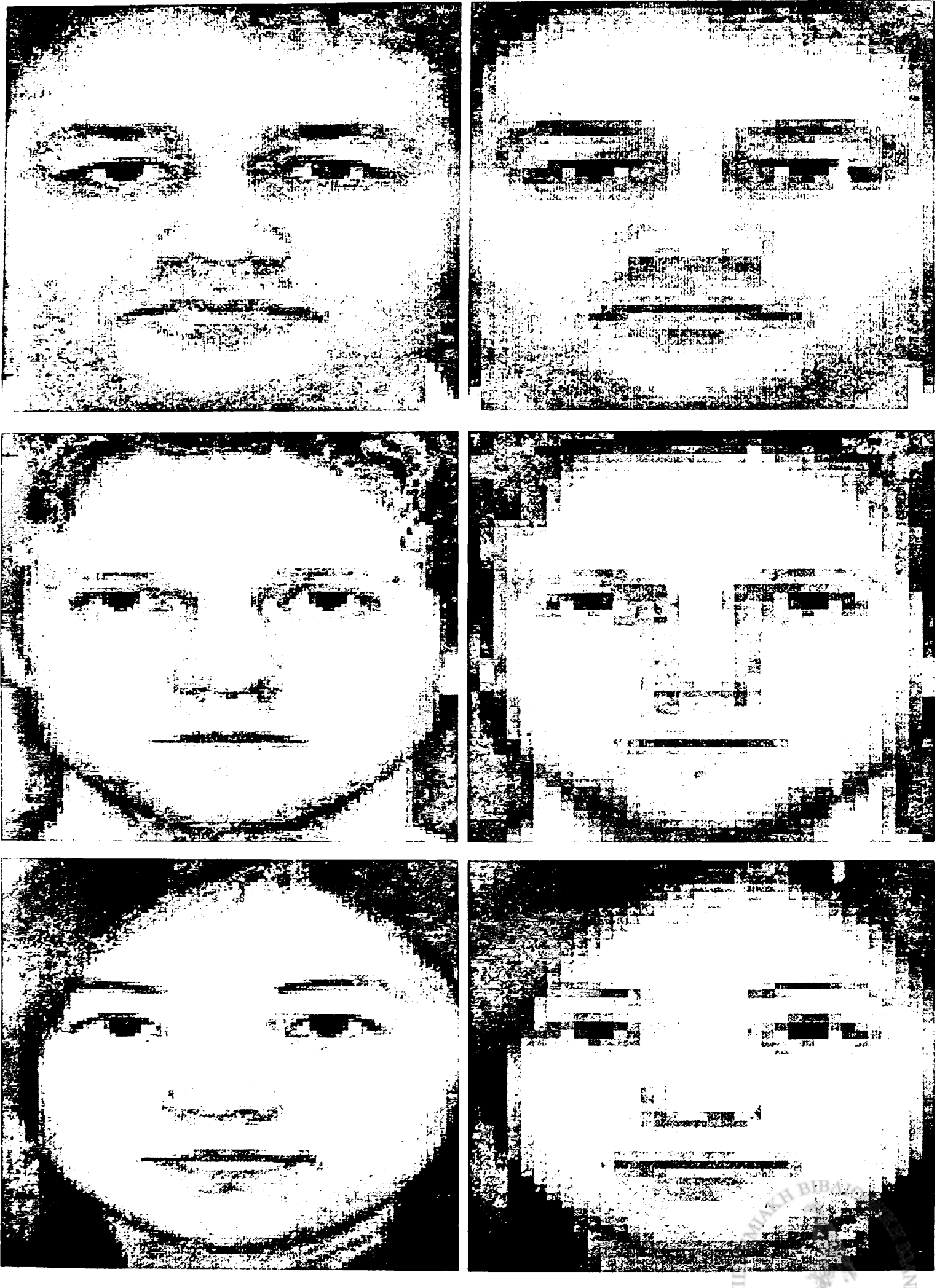


Fig. 4.2: The original high-resolution image (left) and the corresponding low-resolution one (right) which was produced by downsamping the original. A lot of

Fig. 4.2 shows the synthetic low-resolution images and the corresponding high-resolution ones from which they were produced. Since, the CMU MultiePIE database does not include low-resolution images, all the images in the database were downsampled to form the low-resolution images. It is obvious that the downsampling process led to a great loss of the image information making the reconstruction of the original image a challenging job.

In the super-resolution problem (3.10) only one free parameter exists and that is λ, which balances the trade-off between fitting the data perfectly and employing a sparse solution. This optimization problem indicates that λ depends on how noisy the input data are. The noisier the data, the larger the value of λ should be, and the larger the value of λ, the smoother the result image gets. In all the experiments presented in this chapter the value of λ is set to 0.4, $\lambda = 0.4$. The dictionary size is always 1024, both for the dictionary of the low-resolution patches and the dictionary of the high-resolution patches, because larger dictionaries generate better results. The problem is that the computation cost increases while the dictionary gets larger, because obtaining compact dictionaries from the training data, which are the low-resolution patches and the corresponding high-resolution ones, that ensure sparse representation is by itself a biconvex optimization problem. Consequently, this was the problem we were confronted with when we were dealing with the image super-resolution technique described in this thesis.

Furthermore, in this chapter, the presentation of the obtained super-resolved images is followed by a quantitative evaluation of the results. The image quality evaluation methods which are most commonly used in image processing are the Root Mean Square Error (RMSE), the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM). While RMSE and PSNR attempt to quantify the visibility of errors between a distorted image and a reference image by using a variety of known properties of the human eye perception, SSIM which was developed by [35] improves this two methods by relying on the degradation of the structural information, i.e. the image degradation is considered as perceived change in structural information. All these metrics are used for the evaluation of the results.

4.1. Image Quality Evaluation Methods

The image quality evaluation methods can be divided into two categories, the subjective methods and the objective methods. The methods of the first category are based on human judgment and operate without reference to explicit criteria. The methods of the second category are based on comparisons using explicit numerical criteria and several references are possible, such as the ground truth or prior knowledge expressed in terms of statistical parameters and tests. Therefore, the methods that are useful for the evaluation of the results generated by the image super-resolution technique described in chapter 3 are the objective methods. These methods can be further divided into three categories, according to the availability of an original (reference) image with which the reconstructed image is compared. The first category includes the full-reference methods which assume that a complete reference image is available. The second category is composed of the no-reference methods where the reference image is not available and the third category includes the reduced-reference methods where the reference image is only partially available , as a set of extracted features made available as side information to help evaluate the quality of the distorted image. Since, in all the experiments of this thesis the reference image is always available, the full-reference image quality methods will be used for the evaluation of the results. More specifically, RMSE, PSNR and SSIM will be used.

4.1.1. RMSE

The root-mean-square error (RMSE) or root-mean-square deviation (RMSD) is a measure of the differences between values predicted by a model or an estimator and the values actually observed. Thus, it is a measure of the average magnitude of the error. If X(i,j) are the values of the original (reference) image and Y(i,j) are the predicted values of the parameter in question which is the reconstructed image, with i = 1,2,...,N and j = 1,2,...,M then the mathematical definition of the RMSE is:

$$\text{Mean square Error (MSE)} = \frac{1}{N x M} \sum_{i=1}^{N} \sum_{j=1}^{M} (Y(i,j) - X(i,j))^2$$

$$\textbf{RMSE} = \sqrt{\text{MSE}}.$$

RMSE is a squared quantity, since the errors are squared before they are averaged, which means that it is influenced more strongly by large errors than by small errors. The value of RMSE ranges from 0 to infinity and it is ideal when it is small; the smaller the better, with 0 being the perfect score.

4.1.2. PSNR

The peak signal-to-ratio (PSNR) metric is the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Specifically, in image processing the signal is the original image and the noise is the error introduced while reconstructing the image. While RMSE represents the error between the original image and the reconstructed image, PSNR represents the peak of the error, and it is mathematically defined as:

$$\text{PSNR} = 10\log_{10} \frac{R^2}{MSE}$$

where R is the maximum fluctuation in the input image data type. If the input image has a double-precision floating data type then R = 1. If it has an 8-bit unsigned integer data type then R = 255, etc. In the experiments of this chapter it is always 255. The value of PSNR approaches infinity as the value of MSE approaches zero. This indicates that a higher PSNR value provides a higher image quality.

4.1.3. SSIM

The structural similarity index (SSIM) is a metric used to measure the similarity

between two images and it was developed in order to be a more reliable criterion for visual image quality than RMSE and PSNR. SSIM considers the image degradation as perceived change in the structural information thus, models any image distortion as a combination of three factors which are the loss of correlation (s), the luminance distortion (l) and the contrast distortion (c). The SSIM is mathematically defined as:

$$SSIM\,(X,Y) = f(\,l(X,Y),\, c(X,Y),\, s(X,Y))$$

where

$$l(X,Y) = \frac{2\mu_X\mu_Y + C_1}{\mu_X^2 + \mu_Y^2 + C_1}$$

$$c(X,Y) = \frac{2\sigma_X\sigma_Y + C_2}{\sigma_X^2 + \sigma_Y^2 + C_2}$$

$$s(X,Y) = \frac{\sigma_{XY} + C_3}{\sigma_X\sigma_Y + C_3}$$

with μ_X and μ_Y being the average of X and Y respectively, σ_X^2 and σ_Y^2 the variance of X and Y respectively, σ_{XY} the covariance of X and Y, σ_X and σ_Y the standard deviation of X and Y, and C_1, C_2, C_3 being the positive constants which are used to avoid a null denominator. The positive values of the SSIM index are decimals between 0 and 1, where a value 0 indicates that there is no correlation between the two images and a value 1 indicates that the images are identical. In simple words, without the use of mathematics, SSIM method can be described as a comparison of local patterns of pixel intensities that have been normalized for luminance and contrast. However, for the evaluation of the quality of an image it is more useful to use one single measure for the quality of the entire image, instead of using SSIM indexes of local patterns of pixels, thus a mean SSIM index which evaluates the total quality of the whole image is important to be defined. This mean SSIM index is known as MSSIM and is formulated as:

$$MSSIM(X,Y) = \frac{1}{M}\sum_{j=1}^{M} SSIM(x_j,y_j)$$

where M is the number of the local windows of the image, in which the image is divided in order to compute the quality differences between the pixels of the reference image and the corresponding pixels of the reconstructed image, and x_j, y_j are the image contents at the jth local window.

4.1.4. Evaluation of the Evaluation techniques!

The presentation of the full-reference methods RMSE, PSNR and SSIM indicates that the first two are simplest. Therefore, they are mostly used for the quality evaluation in image processing because they contain simple calculations, their physical meanings are clear and they are mathematically convenient for optimization. The disadvantage is that in some cases, one reconstructed image with a low PSNR may appear visually more appealing than another image with higher PSNR due to the way that the human visual perception works. Similarly, a reconstructed image with a high RMSE may appear to be better, that is closer to the reference image, than another one with lower RMSE. Furthermore, various types of degradations applied to the same image may lead to the same value of RMSE. This shows that the RMSE and PSNR metrics are not reliable criteria for visual image quality, although some studies have shown that these two have the best performance in estimating the quality of noisy images. SSIM was developed to improve RMSE and PSNR by taking into account the perceived changes in structural information, that is the strong dependencies between the pixels of an image which carry important information about the structure of the objects in the visual scene. Nevertheless, RMSE and PSNR are still most widely used for the evaluation of reconstructed images in image processing applications, and that is why all these three metrics will be used in the experiments presented in the next section.

Fig. 4.3 shows that the RMSE metric is an insufficient criterion for visual image quality. The images (b) – (f) are different types of distortion of the same reference

image (a). Although, all these images have the same RMSE value (14,5), which means that for the RMSE method these images are of the same quality relative to the reference image, the visual results for some of them are very bad. For the human visual perception, in other words for the human eye, image (b) is of the best quality in comparison with all the other images, while image (c) is of the worst quality. Furthermore, none of these images seems to be visually of the same quality with another, as the value of RMSE implies. This is proven by the MSSIM metric which indicates that the images (b) – (f) are of different quality, with (c) being the best and (d) the worst. Despite these observations, RMSE will be used for the evaluation of the super-resolution results because it is commonly used.

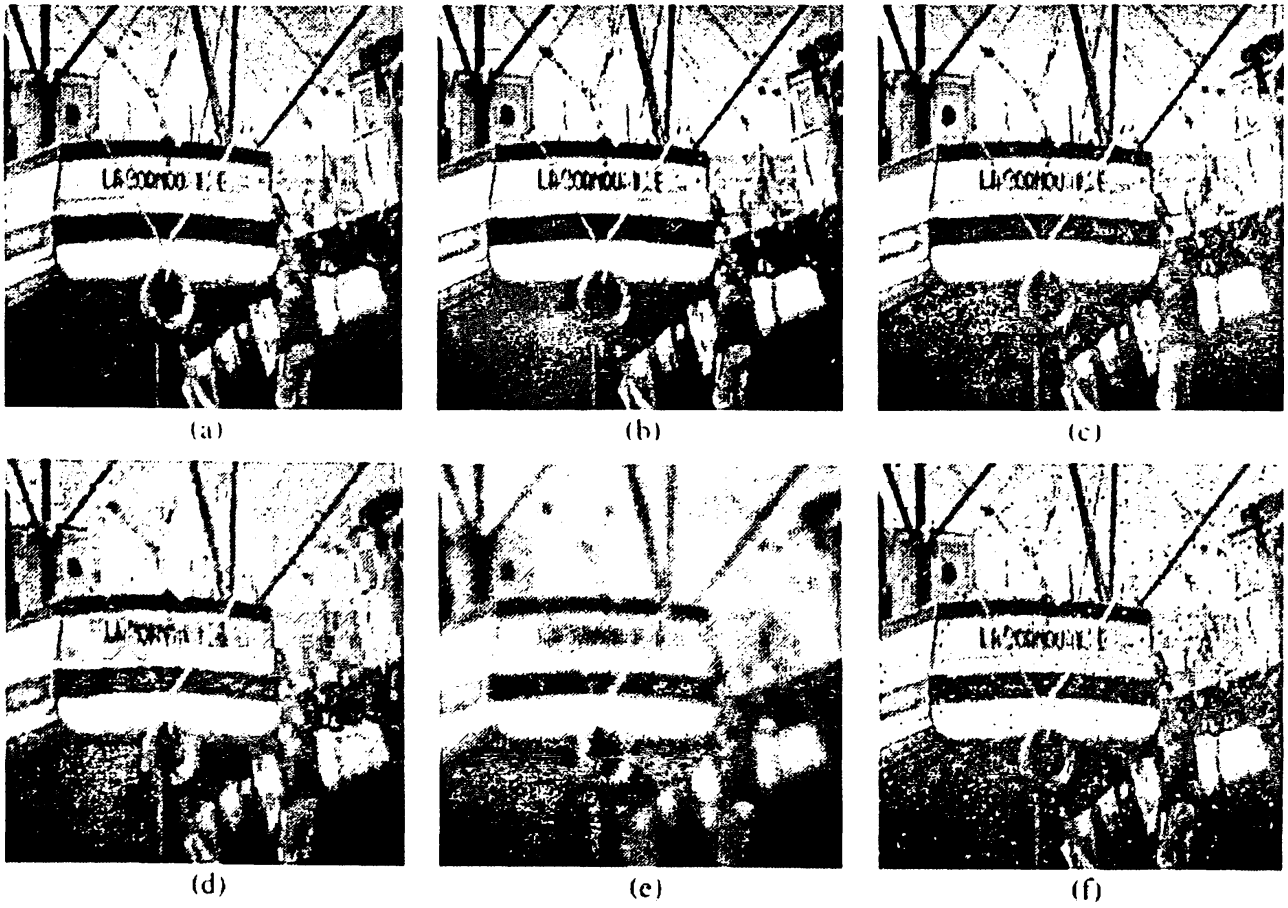


Fig. 4.3: Comparison of different types of distortion of the same reference image, all with RMSE = 14.5. (a) Reference image, (b) Contrast-stretched image, MSSIM = 0.9168, (c) Mean-shifted image, MSSIM = 0,9900, (d) JPEG compressed image, MSSIM = 0.6949, (e) Blurred image, MSSIM = 0.7052 and (f) Salt-pepper impulsive noise contaminated image, MSSIM = 0.7748.

Another example which illustrates that the PSNR metric which is related to the RMSE metric can disagree with the SSIM outcome about the quality of an image when compared with a reference image is shown in fig. 4.4. This figure shows two images that yield the same MSSIM index (0.9480) relative to the same reference image. For the RMSE and the PSNR method, and also for the human eye, these two images are qualitatively different. Both of these metrics conclude that the image on the right is of better quality than the one on the left although the MSSIM index indicates that both of them are of the same quality relative to the reference image on the top. However, a subjective comment would be that the image on the left seems better than the one on the right because it is clearer and smoother, and thus closer to the reference image, but the explicit objective numerical criteria reject such an allegation.

Fig. 4.4: Comparison of two images with the same MSSIM = 0.9480. Reference image (top), image with RMSE = 5.80 and PSNR = 32.86 (left), image with RMSE =5.75 and PSNR = 32.94 (right).

4.2. Results

In this section, the super-resolved images that were generated by the application of the super-resolution algorithm with the use of sparse representation described in chapter 3 are demonstrated. The reconstructed images are presented side by side with the low-resolution input images and the original high-resolution images (which were used for the creation of the low-resolution inputs), to illustrate that the algorithm generates outputs which enhance the resolution of the inputs and which are visually appealing relative to the original high-resolution images. Furthermore, the reconstructed images are compared to the outputs of other methods, such as kernel, nearest-neighbor, bilinear and bicubic interpolation to show that super-resolution via sparse representation yields better results. Once again the comparison is visual by the side by side presentation of the generated images of all the methods, but the comparison is also quantitative to ensure that the subjective factor which is the limitations of the human visual perception will not lead to misunderstandings about the quality of the outputs. Therefore, the three objective image quality evaluation metrics which were described in the previous section, i.e. RMSE, PSNR and MSSIM, will be used. The input images constitute the test examples and they were not used during the training process of the coupled dictionaries. Finally, after training two coupled dictionaries using the method of [25], meaning the use of the upsampled patches during the training and the reconstruction process, we obtained high-resolution images which are compared with the high-resolution images of our approach where the dictionaries are trained using the original low-resolution patches and the sparse representation is also found by the initial low-resolution patches.

Fig. 4.5 shows the results of super-resolution via sparse representation. The low-resolution image which is the input for the algorithm is on the top, and its size is 51x35 pixels. The super-resolved image is the one in the center and it was upsampled by a factor of 2x2, thus its size is 102x70 pixels. Finally, the original high-resolution image is the one at the bottom with 102x70 pixels size. The comparison between the input image and the super-resolved one reveals that the super-resolution process

generates visually appealing results, since all the characteristics of the face are clear with straight and sharp edges, explicitly distinguished at the super-resolved image while the facial information (corners and edges) at the input image are vague and blurred and concentrated only in a couple of pixels. The reconstructed high-resolution image is close to the original high-resolution one something that is proven by the quality evaluation MSSIM index (0.9530), but also by the RMSE (5.15) and PSNR (33.8947) values. In Table 4.1, the values of these three metrics for the reconstructed image are presented. The value of the MSSIM index means that the super-resolved image is 94.8% similar to the original high-resolution image.

Table 4.2 shows the values of the RMSE, PSNR and MSSIM image quality evaluation methods for the reconstructed image shown in fig. 4.6. The MSSIM index value is interpreted as 94.06% similarity between the super-resolved image and the original high-resolution one.

Metric	Value
RMSE	**5.15**
PSNR	**33.8947**
MSSIM	**0.9530**

Table 4.1: RMSE, PSNR and MSSIM values of the reconstructed image in fig. 4.5. The image is 95.3% similar to the reference (original) high-resolution image.

Metric	Value
RMSE	**4.9148**
PSNR	**34.3006**
MSSIM	**0.9406**

Table 4.2: RMSE, PSNR and MSSIM values of the reconstructed image in fig. 4.6. The image is 94.06% similar to the reference (original) high-resolution image.

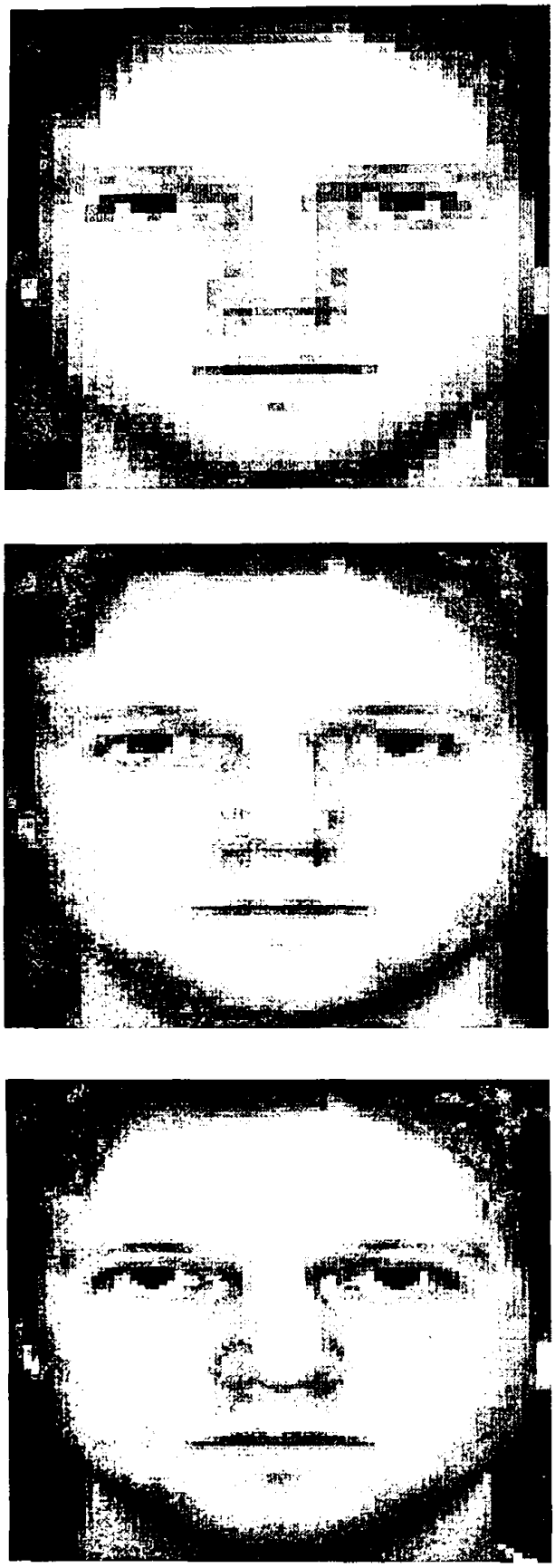


Fig. 4.5: Comparison of the reconstructed image (center) with input (top) and the original high-resolution one (bottom).

Fig. 4.6: Comparison of another reconstructed image (center) wi
input (top) and the original high-resolution one (bottom).

Fig. 4.6 presents another reconstructed image (center) from the test images, meaning that it was not used in the training set during the joint dictionary learning process, compared with the corresponding low-resolution input (top) and the original high-resolution image (bottom). Once again, the information about the corners and edges of the low-resolution image is vague and thus, the super-resolution process becomes challenging. The output, i.e. the super-resolved image, is a very good result where all the facial information has been cleared and smoothed.

Fig.4.7 presents the results of super-resolution via sparse representation compared to other methods, specifically kernel, nearest-neighbor, bilinear and bicubic interpolation. The low-resolution image which is the input for the algorithm is of size 51x35 pixels. All the other images are the high-resolution versions of the input image upsampled by a factor of 2x2, thus their size is 102x70 pixels. The images from left to right are the low-resolution input and its upscaled kernel version on the top, the nearest-neighbor output and the bilinear output in the center and at the bottom the bicubic output and our high-resolution image. The comparison between these images reveals that the kernel and the nearest-neighbor techniques generate bad results with the facial information being as noisy and vague as it was before the magnification, while the corners and edges of the face are still concentrated in a couple of pixels. Bilinear interpolation yields a satisfying result where the corners and edges of the facial information are specific but on the other hand a lot of blur has been added to the image. Bicubic interpolation generates a clearer result which is still blurry, but as the MSSIM index (0.9523) indicates is of good quality relative to the reference image. Finally, our super-resolved output is better and clearer than all the other images without blur, where the characteristics of the face are clear with straight and sharp edges, explicitly distinguished, and the fact that our image is better is proven by the values of the three image quality evaluation methods RMSE (5.15), PSNR (33.8947) and MSSIM (0.9530).

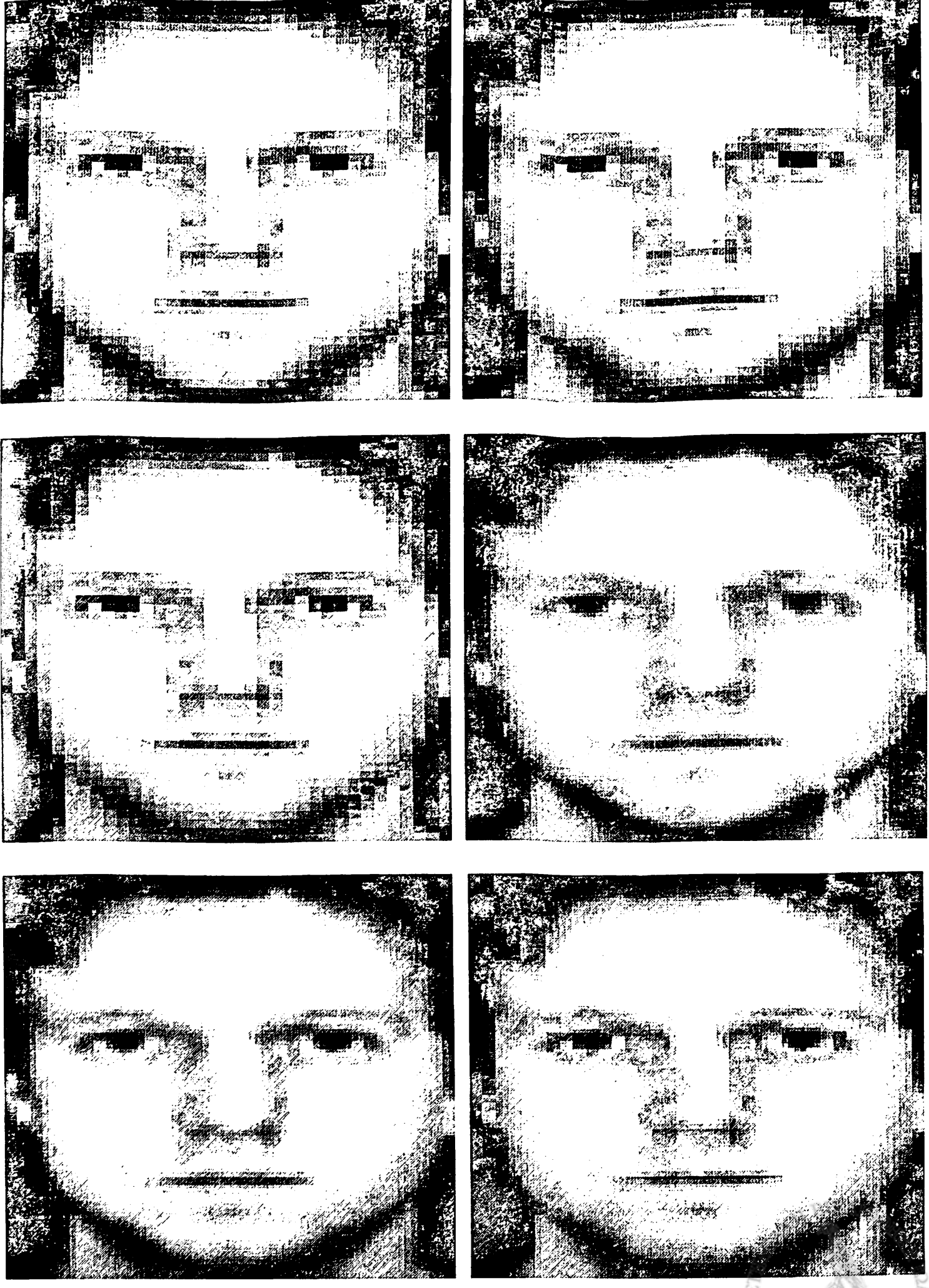


Fig. 4.7: Results of a test image magnified by a factor of 2. Top: low-resolution input, kernel output. Center: nearest-neighbor, bilinear interpolation. Bottom: bicubic interpolation and super-resolution by sparse representation.

In Tables 4.3, 4.4, 4.5, the values of these three metrics for all the reconstructed images are presented. These values show that the subjective conclusions of a human observing the images in fig. 4.7 can be supported by the objective "observers" which are the evaluation techniques. The kernel and nearest-neighbor methods have the worst values (RMSE = 7.3835), (PSNR = 30.7655) and although they seem extremely bad the MSSIM index implies that they are 91.85% similar to the original high-resolution image! The values for the other methods lead to the same conclusions described in the previous paragraph. The bilinear output is better than the one of the kernel and nearest-neighbor methods, bicubic is even better and finally, our super-resolved image is the best according to all the metrics, and specifically the MSSIM index indicates that it is 95.3% similar to the original high-resolution image.

Tables 4.6, 4.7, 4.8 show the values of the RMSE, PSNR and MSSIM image quality evaluation methods for the output images shown in fig. 4.8. Once again, the kernel and nearest-neighbor methods have the worst values for all the three metrics and although their output images seem extremely bad, the MSSIM index indicates that they are 91.05% similar to the original high-resolution image! The values for the other methods lead to the same conclusions with those described for fig. 4.7. Bilinear is better than kernel and nearest-neighbor, bicubic is even better and finally, our super-resolved image is the best according to all the metrics, and specifically the MSSIM index indicates that it is 94.06% similar to the original high-resolution image.

Fig.4.8 presents the results of super-resolution via sparse representation for another test image compared to kernel, nearest-neighbor, bilinear and bicubic interpolation methods. The low-resolution image (top, left) is of size 51x35 pixels. All the other images are the high-resolution versions of the input image upsampled by a factor of 2x2, thus their size is 102x70 pixels. The kernel output (top, right) and the nearest-neighbor output (center, left) are the worst results with noisy and vague corners and edges, concentrated in a couple of pixels. The bilinear output (center, right) is better with specific corners and edges but again annoying blur has been added to the image. The bicubic output (bottom, left) is clearer but still blurry, and according to the MSSIM index 93.98% similar to the original high-resolution image. Finally, our super-resolved output (bottom, right) is better than all the other images without blur,

with straight and sharp edges and 94.06% similarity to the original high-resolution image.

Method	RMSE
Kernel	7.3835
Nearest-neighbor	7.3835
Bilinear	6.1766
Bicubic	5.1967
Our	5.1505

Table 4.3: RMSEs of the images in fig. 4.7. Our super-resolved image has the best value.

Method	PSNR
Kernel	30.7655
Nearest-neighbor	30.7655
Bilinear	32.3158
Bicubic	33.8162
Our	33.8947

Table 4.4: PSNRs of the images in fig. 4.7. Our super-resolved image has the best value.

Method	MSSIM
Kernel	0.9185
Nearest-neighbor	0.9185
Bilinear	0.9429
Bicubic	0.9523
Our	0.9530

Table 4.5: MSSIM values of the images in fig. 4.7. Our super-resolved image has the best value which is interpreted as 95.3% similarity to the original high-resolution image.

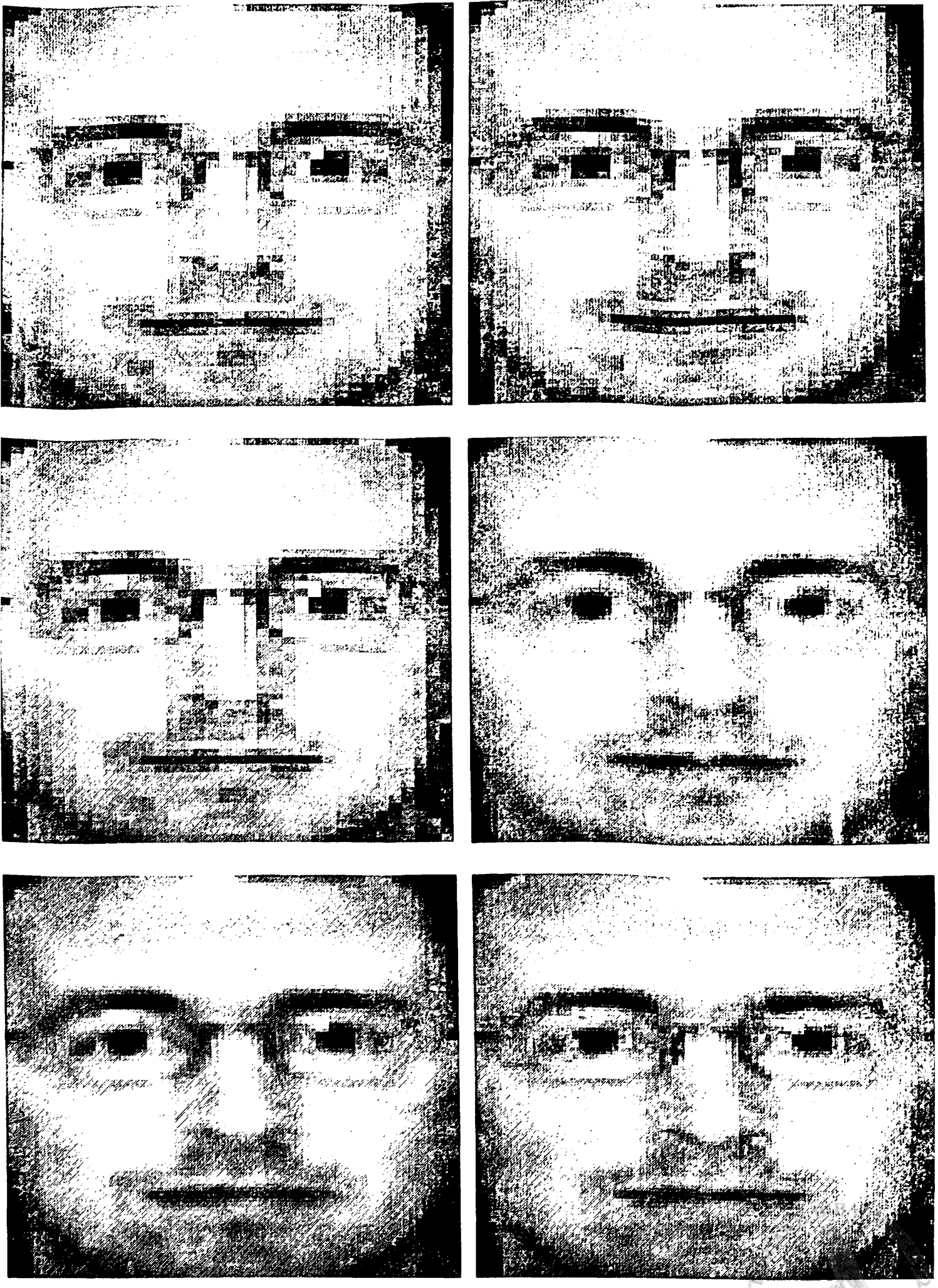


Fig. 4.8: Results of another test image magnified by a factor of 2. Top: low-resolution input, kernel output. Center: nearest-neighbor, bilinear interpolation. Bottom: bicubic interpolation and super-resolution by sparse representation.

Method	RMSE
Kernel	6.8192
Nearest-neighbor	6.8192
Bilinear	5.8228
Bicubic	4.9177
Our	4.9148

Table 4.6: RMSEs of the images in fig. 4.8. Our super-resolved image has the best value.

Method	PSNR
Kernel	31.4561
Nearest-neighbor	31.4561
Bilinear	32.8281
Bicubic	34.2954
Our	34.3006

Table 4.7: PSNRs of the images in fig. 4.8. Our super-resolved image has the best value.

Method	MSSIM
Kernel	0.9105
Nearest-neighbor	0.9105
Bilinear	0.9259
Bicubic	0.9398
Our	0.9406

Table 4.8: MSSIM values of the images in fig. 4.8. Our super-resolved image has the best value which is interpreted as 94.06% similarity to the original high-resolution image.

Fig. 4.9 shows an image generated by using the upsampled version of the low-resolution input image (left) and the one generated directly from the initial low-resolution input (right). The two coupled dictionaries for each approach were trained with the same few patches sampled from the training data (100 patches were chosen to speed up the computations of the learning process). The MSSIM value shows that the images are almost the same relative to the high-resolution image indicating that super-resolved images can be correctly recovered even by the initial information of the low-resolution input.

Fig. 4.9: The super-resolved image given by the upsampled low-resolution input (left), MSSIM = 0,9268, and the super-resolved given by the initial low-resolution input (right), MSSIM = 0,9218.

CHAPTER 5. CONCLUSIONS

The present thesis focuses on the problem of single image super-resolution via sparse representations, meaning the process of obtaining a high-resolution version of a low-resolution image when only one single low-resolution image is known. The sparse representations are used in the recovery process in terms of combined dictionaries which are simultaneously trained from a database of low- and high-resolution image patch pairs sampled from training images. The essence of the dictionary training process is to learn the correspondences between the low- and high-resolution image patch pairs. Once the sparse representation of each patch of the low-resolution image is known, its coefficients can be used to recover the most likely high-resolution version of the output image.

The super-resolution method which was presented by the authors of [25] takes advantage of this observation and applies the sparse representation of a low-resolution image patch with the high-resolution image patch dictionary to yield a high-resolution image patch. The low-resolution input is upsampled before the training of the dictionaries and during the reconstruction process. A different approach was adopted in the frame of this thesis. The initial low-resolution image patches are directly used to obtain the two coupled dictionaries and then the original low-resolution patches are used in order to generate the high-resolution image. This perspective was materialized in the frame of this thesis and tested on synthetic data. Therefore, after obtaining the appropriate low-resolution and its corresponding high-resolution image patch dictionary, the sparse representation of any low-resolution image given as an input can be found with respect to the low-resolution dictionary. Then, the corresponding high-resolution patch bases of the high-resolution dictionary will be combined according to these coefficients to generate the output high-resolution image. The

conclusions drawn from the implementation of the super-resolution by sparse representation new approach method indicate that although the super-resolution problem is severely ill-posed in general, meaning that precise recovery of the high-resolution image is impossible, the results show the effectiveness of the method generating images of high quality.

The problem of obtaining coupled dictionaries, which is of great importance in image processing generally and in super-resolution problems specifically, is also addressed in this thesis. The dictionaries are learned directly from the data and provide sparse representations for the training images. The dictionary learning process is generally divided into two iterative stages, the sparse approximation stage and the dictionary update stage. More specifically, in the first stage the algorithms start with an initial dictionary and try to find sparse approximations of the set of training signals while keeping the dictionary fixed, and in the second stage the sparse coefficients are kept fixed while the dictionary is optimized. Once the dictionaries are known, they are used in the recovery process to generate sparse representations for the test images and furthermore, to contribute in the reconstruction of the output.

Image super-resolution is an active field at the moment because it provides answers and solutions through software instead of expensive hardware to many applications of our days, such as image processing, medical imaging devices, satellite imaging, surveillance cameras, visual electronics and document analysis. Therefore, it is expected to stay at the spotlight for a long time since all the research and work presented so far is promising, indicating that further improvement and enhancement can be achieved in the future.

There are limitations in image super-resolution which should be broken in the future through systematic research. The method presented in the present thesis requires large training data in order to obtain the two coupled sparse dictionaries which are crucial during the reconstruction process. Most of the times the training data used in machine learning techniques are of some hundred thousand so that the compact dictionaries will ensure sparse representation. Due to the computational cost of such a large problem the training examples used for the experiments of the previous chapter were

almost ten times smaller, generating decent high-resolution images. Provided sufficient computational resources more training data could be used in the dictionary learning process and yield even better results than the ones presented. Furthermore, the computational cost is affected by the size of the dictionary. Larger dictionaries which are important for generating more accurate approximation lead to heavier computation. Thus, the computational cost is a problem which needs improvement, for example by developing new or improving the existing dictionary learning solvers.

The method presented has the advantage that the sparse representation coefficients of the low-resolution input are directly used to recover the high-resolution output. However, the disadvantage of this method is that for different increases in resolution, that is different magnification factors, different jointly trained dictionaries need to be constructed! Therefore, improvements and different approaches may result in a more flexible super-resolution method based upon sparse representation.

At the global model described in chapter 3 where the global reconstruction constrain is redefined in order to cope with larger high-resolution images (optimization problem (3.15)), a penalty function which includes prior knowledge about the high-resolution image is included. This type of knowledge was not included in the experiments conducted since the only available prior knowledge was the sparse representation generated from the downsampled signals. Hence, in cases where more information about the high-resolution data is available or can be extracted the performance may be improved.

Furthermore, the single-image technique presented here may result in better outputs if it is combined with multi-image super-resolution in cases where multiple low-resolution images are available. In these cases, the linear relationships among the high-resolution signals, that is the sparsity of the representation coefficients, recovered from the low-resolution signals may be more accurate. Consequently, the obtained dictionaries will be more appropriate.

Since super-resolution by sparse representation was applied on human faces in the frame of this thesis, some further observations and limitations can be mentioned. In

the turbulent days that we live in, face recognition became very popular especially in forensic image analysis in terms of identifying or verifying an individual by an image or a video frame. Although, a great variety of face hallucination techniques are available for generating high-resolution images that portray the details of facial features recovered from low-resolution images and the results they have achieved are remarkable, there is still need for improvement and expansion. Improvement in order to obtain better super-resolved images and expansion in order to achieve remarkable results in cases of profile images instead of full frontal images and some degrees off, cases of poor lighting or with sunglasses or other accessories covering the individual's face. Even though, in our experiments the training images included accessories -not sunglasses- and in the case of males some faces were partially covered by facial hair generating qualitatively and quantitatively appealing results, the performance of the algorithm in the previously mentioned cases (profile, poor lighting, sunglasses or other objects), which are common in real life face hallucination, is not known. Therefore, the presented super-resolution technique can be further tested by applying it at profile images, poor lighting images, accessorized face images and adapting it appropriately in order to generate decent results even in these challenging occasions. Finally, the technique can be further tested and improved to deal with real data and not just synthetic data, artificially generated by the original high-resolution images as the ones used in the frame of this thesis.

BIBLIOGRAFY

[1] R. Y. Tsai and T. S. Huang. "Multiframe image registration and restoration", Advances of Computer Vision and Image Processing, Vol. 1, pp 317-339, 1984.

[2] A. M. Tekalp, M. K. Ozkan and M. I. Sezan. "High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration", Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 3, pp 169–172, March 1992.

[3] S. P. Kim, N. K. Bose, H. M. Valenzuela. "Recursive reconstruction of high resolution image from noisy undersampled multiframes", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 38(6), pp 1013-1027, June 1990.

[4] S. Borman and R. Stevenson. "Spatial resolution enhancement of low-resolution image sequences: a comprehensive review with directions for future research", July 1998.

[5] C. L. L. Hendriks, L. J. van Vliet. "Improving resolution to reduce aliasing in an undersampled image sequence", SPIE, Vol. 3965, pp 214-222, January 2000.

[6] S. Dai, M. Han, W. Xu, Y. Wu, and Y. Gong. "Soft edge smoothness prior for alpha channel super resolution", Proc. IEEE Conf. Computer Vision and Pattern Class., pp 1–8, 2007.

[7] T. R. Tuinstra and R. C. Hardie. "High-resolution image reconstruction from digital video by exploitation of non-global motion", SPIE, Vol. 38(5), pp 806-814, May 1999.

[8] R. C. Hardie, K. J. Barnard, and E. Armstrong. "Joint MAP registration and high-resolution image estimation using a sequence of undersampled images", IEEE Trans. on Image Processing, Vol. 6(12), pp 1612-1633, December 1997.

[9] M. E. Tipping and C. M. Bishop. "Bayesian image super-resolution", Proc. Adv. Neural Inf. Process. Syst. 16, pp 1303–1310, 2003.

[10] M. Elad and A. Feuer. "Super-resolution reconstruction of image sequences", IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), Vol. 21(9), pp 817-834, September 1999.

[11] M. Irani and S. Peleg. "Improving resolution by image registration", CVGIP: Graphical Models and Image Processing, Vol. 53(3), pp 231-239, May 1991.

[12] M. Irani and S. Peleg. "Motion analysis for image enhancement: resolution, occlusion and transparency", Journal of Visual Communications and Image Representation, Vol. 4(4), pp 324-335, December 1993.

[13] H. Stark and P. Oskoui. "High-resolution image recovery from plane-image arrays, using convex projections", Journal of the Optical Society of America A, Vol. 6(11), pp 1715-1726, November 1989.

[14] S. Farsiu, M. D. Robinson, M. Elad and P. Milanfar. "Fast and robust multiframe super resolution", IEEE Trans. on Image Processing, Vol. 13(10), pp 1327-1344, October 2004.

[15] M. S. Alam, J. G. Bognar, R. C. Hardie, B. J. Yasuda. "High resolution image reconstruction using multiple, randomly shifted, low resolution, aliased frames", SPIE, Vol. 3063, pp 102-431, June 1997.

[16] H. Foroosh and J. B. Zerubia. "Extension of phase correlation to subpixel registration", IEEE Trans. on Image Processing, Vol. 11(3), pp 188-200, March 2002.

[17] N. Nguyen, P. Milanfar and G. Golub. "A computationally efficient superresolution image reconstruction algorithm", IEEE Trans. on Image Processing, Vol. 10(4), pp 573-583, April 2001.

[18] Z. Lin and H. Y. Shum. "Fundamental limits of superresolution algorithms under local translation", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 26(1), pp 83-97, January 2004.

[19] S. Baker and T. Kanade. "Limits on super-resolution and how to break them", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24(9), pp 1167–1183, September 2002.

[20] W. T. Freeman, E. C. Pasztor and O. T. Carmichael. "Learning low-level vision", International Journal of Computer Vision, Vol. 40(1), pp 25–47, October 2000.

[21] S. Baker and T. Kanade. "Hallucinating faces", Proc. IEEE Int. Conf. Automatic Face Gesture Recognition, pp 83-88, March 2000.

[22] W. T. Freeman, T. R. Jones and E. C. Pasztor. "Example-based superresolution", IEEE Computer Graphics and Applications, Vol. 22(2), pp 56-65, March-April 2002.

[23] D. Glasner, S. Bagon and M. Irani. "Super-resolution from a single image", IEEE International Conference on Computer Vision, pp 349-356, October 2009.

[24] H. Chang, D. Y. Yeung and Y. Xiong. "Super-resolution through neighbor embedding", Proc. IEEE Conf. Computer Vision and Pattern Class., Vol. 1, pp 275-282, June-July 2004.

[25] J. Yang, J. Wright, T. Huang and Y. Ma. "Image super-resolution via sparse representation", IEEE Transactions on Image Processing, Vol. 19(11), pp 2861-2873, November 2010.

[26] B. A. Olshausen and D. J. Field. "Natural image statistics and efficient coding", Network: Computation in Neural Systems, Vol. 7(2), pp 333-339, May 1996.

[27] M. S. Lewicki and T. J. Sejnowski. "Learning overcomplete representations", Neural Comput., Vol. 12(2), pp 337-365, February 2000.

[28] B. A. Olshausen and B. J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?", Vision Research, Vol. 37(23), pp 3311-3325, December 1997.

[29] K. Engan, S.O. Aase and J.H. Husoy. "Method of optimal directions for frame design", IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 5, pp 2443-2446, March 1999.

[30] 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(11), pp 4311-4322, November 2006.

[31] K. Skretting and K. Engan. "Recursive Least Squares Dictionary Learning Algorithm", IEEE Transactions on Signal Processing, Vol. 58(4), pp 2121-2130, April 2010.

[32] W. Dai, T. Xu and W.Wang. "Simultaneous Codeword Optimization (SimCO) for dictionary update and learning", IEEE Transactions on Signal Processing, Vol. 60(12), pp 6340-6353, December 2012.

[33] M. G. Jafari and M. D. Plumbley. "Speech denoising based on a greedy adaptive dictionary algorithm", EUSIPCO: European Signal Processing Conference, pp 1423-1426, August 2009.

[34] H. Lee, A. Battle, R. Raina, and A. Y. Ng. "Efficient sparse coding algorithms", Proc. Advances in Neural Information Processing Systems, pp. 801–808, 2007.

[35] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli. "Image quality assessment: From error visibility to structural similarity", IEEE Transactions on Image Processing, Vol. 13(4), pp 600-612, April 2004.