ADVANCED SPARSITY TECHNIQUES IN MEDICAL IMAGING AND IMAGE PROCESSING

by

CHEN CHEN

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2015

Copyright © by Chen Chen 2015

All Rights Reserved

To my parents Shizhong and Hongxia, my wife Jingcao, for their endless trust, support, encouragement and love.

ACKNOWLEDGEMENTS

I would like to thank my supervising professor Dr. Junzhou Huang for constantly motivating and encouraging me, and also for his invaluable suggestions during my studies in UTA. He taught me how to conduct research and how to think independently. I learned a lot of tips for scientific experiments and academic writing, which contributes significantly to my publications. None of the work in this thesis would have happened without him. I wish to thank Dr. Chris Ding, Dr. Heng Huang and Dr. Jeff Lei for their interest in my research and for taking time to serve in my dissertation committee.

I would also like to extend my appreciation to Dr. Darin Brezeale for letting me working with him as a teaching assistant. I learned a lot of teaching experience from him. I wish to thank the CSE advisors Dr. Bahram Khalili and Dr. Ramez Elmasri for their support. I want to thank Dr. Hanli Liu and Dr. Fenghua Tian in the Department of Bioengineering. I have been learning a lot from them through the collaborations.

Finally, I would like to express my deep gratitude to my wife and my parents, for their sacrifice, encouragement and patience. Without them, it is impossible for me to achieve such a goal in my career. I also thank several of myfriends who have helped me throughout my career. It is my great pleasure to meet these nice people here.

February 5, 2015

ABSTRACT

ADVANCED SPARSITY TECHNIQUES IN MEDICAL IMAGING AND IMAGE PROCESSING

Chen Chen, M.S.

The University of Texas at Arlington, 2015

Supervising Professor: Junzhou Huang

In the past decades, sparsity techniques has been widely applied in the fields of medical imaging, computer vision, image processing, compressive sensing, machine learning etc., and gained great success. In this work, we propose new models of sparsity techniques, which is an extension to the standard sparsity used in the existing works and in the vein of structure sparsity families. First, we introduce the wavelet tree sparsity in natural images. It shows that the tree sparsity regularization often outperforms the existing standard sparsity based techniques in magnetic resonance imaging. Second, we extend the tree sparsity to forest sparsity on multi-channel data. A new theory is developed for forest sparsity, which is compared with the standard sparsity, tree sparsity and joint sparsity both empirically and theoretically. Motivated by the special datasets in remote sensing, we propose a new sparsity model called dynamic gradient sparsity to improve the fusion results. Moreover, a novel model called deep sparse representation is investigated and successfully used in image registration. Finally, we propose a set of fast reweighted least squares algorithms for different optimization problems based on sparsity regularization.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Problem

We are facing a era of big data. Over 350 million new photos are uploading to Facebook each day [1]. Based on the statistic of Youtobe, "100 hours of video are uploaded to YouTube every minute and over 6 billion hours of video are watched each month, that's almost an hour for every person on Earth"[2]. It is a challenging task to store and transmit such data efficiently. In medical imaging, such as magnetic resonance imaging (MRI) patient has to stay in the machine for more than 30 minutes, for a scan of the majority parts[3]. To obtain a high quality image, it is almost unavoidable to accept some dose of X-rays in computed tomography (CT scans)[4]. If we could find a more efficient way to represent the visual information we needed, the tasks of image processing and medical imaging will become easier and less painful.

1.2 Sparse Representation in Medical Imaging and Image Processing

1.2.1 Compressed Sensing

Actually, most of the information in the natural images is redundant. Our human eyes can quickly obtain key information of an image without looking the details of each pixel. Our interested data pieces are relatively very sparse compared with the whole data. In compressed sensing (CS), the capture of a sparse signal and

[1] http://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9

[2] https://www.youtube.com/yt/press/statistics.html

[3] http://info.shields.com/bid/43435/MRI-CT-and-PET-Scan-Times

[4] http://www.xrayrisk.com

compression are integrated into a single process [18, 19]. Mathematically speaking, if the data is denoted as a vector $x \in \mathbb{R}^N$, we can exactly reconstruct the data with only M linear measurements b instead of N total entries under mild assumptions:

$$b = \Phi x + e \tag{1.1}$$

where $\Phi \in \mathbb{R}^{M \times N}$ is a random projection matrix, with $M < N$. This enables us to avoid the process of acquiring whole data and then compressing it. CS also provides the theoretical support for image recovery from limited number of measurements.

1.2.2 Sparse MRI

In medical imaging (or more precisely, MRI), the scanning of a typical image often costs long time due to both physical and physiological reasons [1]. Local motions e.g. breathing, heart beating during the long time scanning may result in ghosting, smearing, streaking on the reconstructed MR image.

Compressive sensing has received abundant attention in the MRI community since it is first studied in MRI by Lustig et al. [1]. The MR image reconstruction process can be formulated as:

$$\hat{x} = \arg\min_x \{\frac{1}{2}\|Fx - b\|^2 + \alpha\|x\|_{TV} + \beta\|\Phi x\|_1\} \tag{1.2}$$

where x is a MR image to be reconstructed, F is the undersampled Fourier transform, y is the vector of k-space measurements, Φ is the wavelet basis, $||x||_{TV}$ is the total variation defined as $\|x\|_{TV} = \sum_{i=1}^{N} \sqrt{((\nabla_1 x_i)^2 + (\nabla_2 x_i)^2)}$. Here, ∇_1 and ∇_2 denote the forward finite difference operators on the first and second coordinates. α and β are two parameters to be tuned. Such conventional CS-MRI method can reconstruct a MR image accurately from undersampled k-space data by utilizing the sparsity of the image in the wavelet or gradient domain. Therefore the sampling processing (i.e., the scanning) can be significantly reduced.

1.2.3 Applications in image processing and computer vision

In addition to the great success in medical image, sparsity based techniques are also very useful in image processing and computer vision, such as image registration [20] face recognition [21], image super-resolution [22], background substraction [23], photometric stereo [24] etc. The ability of sparse representations to uncover semantic information in image processing and computer vision is based on that, the images (naturally very high dimensional) often lie on or near low-dimensional subspaces, submanifolds, or stratifications [25]. The optimization methods such as ℓ_1 norm minimization can efficiently extract such key structures, and then recover the original image without or with little information lost. Algorithms based on sparse representation can often achieve state-of-the-art performance if the sparsity properly applied [25].

1.3 Motivation

The above mentioned methods have achieved great success in medical imaging and image processing, but they often merely use the sparsity prior to solve the problems. In a lot of practical data, the sparsity patterns are not randomly distributed but follow some special structures. For example, in diffuse optical imaging [26], the activation area of human brain corresponding to a finger tapping task often is sparse among the whole brain. In addition, such activation is often clustered in certain region(s). In dynamic MR images [27], all the images of a cardiac motion have very similar structures along the temporal direction. In background subtraction [28], the foreground objects are often consisted of mutually connected pixels but not randomly distributed ones. Theoretically, it has been shown that better performance can be obtained if we could exploit more prior information about the data [29, 30, 31].

Although both the theories and intuitions encourage advanced methods beyond sparsity techniques for practical applications, the existing structured sparsity based models are still very limited. To bridge this gap, we study several different application and propose advanced data-driven sparsity models to improve the performance. In MRI, we observe the wavelet tree structure of the MR images, which enables us to incorporate the wavelet structure in reconstruction to reduce the sampling rate. This is then extended to the multi-channel images. In remotely sensed images, we find the strong correlations between the panchromatic image and the multispectral image. Thus, the high resolution multispectral image to be recovered can be accurately modeled based on the prior information of the panchromatic image. Unlike the medical images and the remotely sensed images, the natural photos may be captured at significantly different illumination environments and contain partial occlusions and big noise. Therefore it is intuitively to overcome such difficulties separately in a hierarchical architecture. Finally, the existing optimization algorithms are often ℓ_1 norm based minimization. When we need to solve complex sparsity inducing problems, many of the existing algorithms may not work efficiently. This motivate us to develop new efficient algorithms for the challenging optimization problems. The works in this thesis may inspirit more and more advanced work in medical imaging and image processing.

1.4 Organization

The rest of the thesis is organized as follows. Chapter 2 investigates the benefit of tree sparsity in accelerated MRI. We extend the tree sparsity to forest sparsity on multi-channel data in Chapter 3. Theorems of the advantages of forest sparsity are developed. We also show several potential applications of forest sparsity. In Chapter 4, we propose the dynamic gradient sparsity for remotely sensed images. It

shows that the image to be recovered can be more efficiently represented by using the gradient prior information in a reference image. To overcome the difficulties in image registration, we proposed a novel model called deep sparse representation in Chapter 5. It can handle partial occlusions, outliers and intensities distortions in a unified framework. In Chapter 6, a set of fast reweighted least squares algorithms are proposed to solve the convex optimization problems based on sparsity techniques. Finally, we conclude the contributions of this thesis in Chapter 7.

CHAPTER 2

Tree Sparsity in Accelerated Magnetic Resonance Imaging

This chapter investigates the benefits of the tree sparsity in Accelerated Magnetic Resonance Imaging (MRI). In contrast to conventional Compressed Sensing Magnetic Resonance Imaging (CS-MRI) that only relies on the sparsity of MR images in wavelet or gradient domain, we exploit the wavelet tree structure to improve CS-MRI. Simulations and experiments on human brain data demonstrate the significant improvement of the proposed method compared to conventional CS-MRI algorithms. This work was presented under a slightly modification from [88].

2.1 Introduction

Magnetic Resonance Imaging provides a non-invasive manner to aid clinic diagnosis while its limitation is the slow scanning speed. Local motions e.g. breathing, heart beating during the long time scanning may result in ghosting, smearing, streaking on the reconstructed MR image. Parallel MRI (pMRI) [32, 33, 34, 35, 36]and compressed sensing MRI [1] techniques are developed to reduce MR scanning time by undersampling. CS-MRI addresses the issue of recovering images from undersampled k-space data based on compressed sensing (CS) theory [37][18], while the image domain pMRI methods (e.g. SENSE [32]) reconstruct the field of view (FOV) by the aliased images obtained from all coils. This difference makes it possible to combine them in the two-step CS-SENSE scheme to further accelerate MRI scanning [38]. After data is acquired by hardware, the k-space data is first recovered by CS-MRI

methods to aliased images, and then final FOV is unfolded by SENSE from all the aliased images in the first step.

Although both steps are essential to rapid MRI, CS-MRI attracts more attentions recently due to the emerging of CS and sparsity theories and a lot of efficient algorithms (e.g. FISTA [39] and SPGL1 [40]). SparseMRI [1] is the first work to reconstruct MR images from undersampled data based on CS theory, which models MR image reconstruction as a linear combination of least squares fitting, wavelet sparsity and total variation (TV) regularization. The non-smooth terms in their model are smoothed with positive smoothing parameters and then the whole problem is solved by conjugate gradient (CG) method. To improve the reconstruction accuracy and speed, TVCMRI [2] and RecPF [3] use an operator-splitting method and a variable splitting method to solve this problem respectively. FCSA [41] [4] decomposes the original problem into two easy subproblems and separately solves each of them with FISTA [42][39]. They are the state-of-the-art algorithms for CS-MRI. However, such methods may be still limited in clinic MRI due to their reconstruction speed and accuracy. Therefore, algorithms that are both efficient and accurate are quite desirable.

It can be observed that all of the above methods solve the same model as SparseMRI, where the structure of wavelet coefficients are not exploited. Actually, the wavelet coefficients of MR images are not only compressible, but also yield a hierarchical quadtree structure, which is widely applied on image compression [43][44] and signal processing [45]. A typical relationship in the wavelet tree structure is that, if a parent coefficient has a large/small value, its children also tend to be large/small. Recent works on structured sparsity show that the sampling bound could be reduced to $\mathcal{O}(K+\log(N/K))$ by fully exploiting the tree structure instead of $\mathcal{O}(K+K\log(N/K))$ for standard sparsity [29][30], where K represents the non-zero elements of the sparse

data and N is the length of the data. This benefit also can be interpreted as that tree sparsity-inducing penalties encourage tree structure compared with standard sparsity penalties [46][47]. Intuitively, less k-space data is required for the same reconstruction quality, or more accurate reconstruction can be achieved for the same number of k-space samples. Some methods have been proposed to improve compressed sensing imaging by utilizing the tree sparse prior and generally can be divided into three types: greedy algorithms [29][30][48], convex programming [49] and Bayesian learning [50][51][13]. Although these algorithms provide better reconstruction accuracy than those with standard sparsity, they are slow in general due to the intricate tree structure. Apart from this, the sampling matrix is partial Fourier transform in CS-MRI but not random Gaussian matrix that they assumed. Finally, these methods have been validated only on real-valued images while practical MR data is complex-valued. Therefore, no evidence can guarantee the success of these methods in MRI.

In this chapter, we propose a new model to improve conventional CS-MRI [1][2][3][41] [4], where the tree sparsity is combined with standard sparsity and total variation seamlessly. We approximate the tree sparsity as overlapping group sparsity [52]. Due to trade-off between accuracy and computational cost, every coefficient and its parent coefficient are assigned into one group, which force them to be zeros or non-zeros simultaneously. With this configuration, the algorithm will encourage the reconstructed wavelet coefficients to be tree-sparse, but not randomly distributed as by standard sparsity. To solve this overlapping group sparsity problem, an auxiliary variable is introduced to decompose it to three simpler subproblems. Then each of subproblems has a closed form solution or can be solved efficiently by existing techniques. After the data from each coil is reconstructed in the image domain, it is easily to combine pMRI method (e.g. SENSE) in practical applications. Numerical simulations and experiments on human brain MR data demonstrate that the proposed

method always outperforms previous methods on various MR images in terms of both accuracy and computational cost.

2.2 Related Work

2.2.1 CS-MRI

Generally, MR images are sparse in the wavelet domain and the gradient domain, and can be reconstructed with sub-Nyquist-Shannon sampling ratio based on compressed sensing theory. The MRI is first modeled as a CS problem in SparseMRI [1]. Suppose b is the undersampled k-space data, A is the sampling matrix (partial Fourier transform in MRI), then the CS-MRI can be formulated as the linear combination of a least square fitting, total variation and wavelet sparsity regularization:

$$\hat{x} = \arg\min_x \{\frac{1}{2}\|Ax - b\|_2^2 + \alpha\|x\|_{TV} + \beta\|\Phi x\|_1\} \tag{2.1}$$

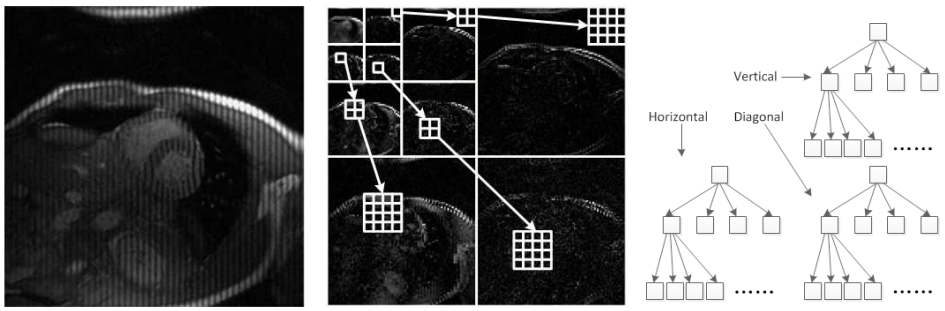
where α and β are two positive parameters, x is the image to be reconstructed and Φ denotes the wavelet transform. $\|x\|_{TV} = \sum_i \sum_j \sqrt{(\nabla_1 x_{ij})^2 + (\nabla_2 x_{ij})^2}$, where ∇_1 and ∇_2 denote the forward finite difference operators on the first and second coordinates. Due to the non-smoothness of ℓ_1 norm and total variation, there is no closed form solution for this problem.

In SparseMRI, the non-smooth terms are transformed to smooth ones by introducing positive smoothing parameters. For example, $||\Phi x||_1 \approx \sqrt{(\Phi x)^T(\Phi x) + \mu}$ where μ is positive and close to zero. Then the approximated problem is solved by classical conjugate gradient (CG) method. Recently, two fast methods TVCMRI [2] and RecPF [3] use an operator-splitting method and a variable splitting method to solve this problem respectively. Both of them have lower time complexity in each iteration, which can substantially reduce the reconstruction time. Accelerated by FISTA [42][39], FCSA [41][4] decomposes the original problem into two subproblems

and then each of them can be solved by existing techniques or has closed form solution. Apart from these, some methods tried to reconstruct compressed MR images by performing ℓ_p-quasinorm ($p < 1$) regularization optimization [53][54][55], which are relatively slow although a little bit of higher compression ratio can be achieved. Overall, all the above methods only improve CS-MRI on the algorithmic level. No structured prior information is utilized other than sparsity.

2.2.2 Theoretical Benefit of Wavelet Tree Structure

The wavelet coefficients for natural data (signals or images) are often approximately sparse, with only a small number of the coefficients have large values and a large fraction of them are approximate zeros. Apart from this, the wavelet coefficients also yield a quadtree structure for a 2D image. The coefficients in the coarsest scale can be seen as the root nodes and the coefficients in the finest scale are the leaf nodes. Each coefficient (non leaf) has four children in the finer scale below it. Figure 2.1 shows the wavelet quadtree structure of an MR image. If this structure can be utilized, the result will be better as more prior information exploited.



(a) (b) (c)

Figure 2.1. Wavelet quadtree structure: a) A cardiac MR image; (b)(c) The corresponding tree structure of the wavelet coefficients.

Based on structured sparsity theories, only $\mathcal{O}(K + \log(N/K))$ measurements is needed to recover tree-sparse data rather than $\mathcal{O}(K + K\log(N/K))$ for standard K-sparse data [30, 31, 29]. Once the location of a non-zero element is fixed, all its ancestors should be non-zeros. Therefore the number of the solution subspaces for $Rx = b$ is significantly limited. The wavelet coefficients also tend to have this good property. If a parent coefficient has a large/small value, its children also tend to be large/small. By exploiting wavelet tree structure, significant improvement can be achieved especially when the data is very compressible ($K << N$).

2.2.3 Algorithmic Benefit of Wavelet Tree Structure

Rao et al. [49] consider the wavelet tree structure as overlapping group lasso regularization [52]:

$$\min_x \{F(x) = \frac{1}{2}\|A\theta - b\|_2^2 + \beta \sum_{g \in \mathcal{G}} ||\theta_g||_2\} \tag{2.2}$$

where θ is the wavelet coefficients. $A = R\Phi^T$ for MR image reconstruction problem, Φ^T is an inverse wavelet transform. β is positive parameter, $\mathcal{G}$ denotes the all parent-child groups and g is one of such groups. When θ is recovered, it can be transferred to the image by an inverse wavelet transform. By geometric interpretation, ℓ_1 norm ball has some singular values at the axes, which only encourages sparseness with no constrains on the selection of axes. The singular values appears on $\ell_{2,1}$ norm ball only when all coordinates in the same group are zeros. Intuitively, overlapping group inducing norm ball encourages overlapping group sparsity [52]. Our work is motivated by Rao's method [49], however, we do not introduce this method to CS-MRI, because the replicating of the sampling matrix in their algorithm is not preferred. Moreover, its solver SpaRSA [56] only achieves the convergence rate $\mathcal{O}(1/k)$ in function value, which is unable to be comparable with the fastest ones with $\mathcal{O}(1/k^2)$.

2.3 Algorithm

To validate the benefit of tree sparsity in accelerated MRI, we propose two algorithms to efficiently solve the constrained and unconstrained tree-based CSMRI problems respectively. The tree structure in MR images is approximated as overlapping groups [52, 49]. The unconstrained problem is solved in FISTA [39] framework and the constrained problem is solved in NESTA [5] framework. Both FISTA and NESTA have the optimal convergence rate for first order methods, that is, $\mathcal{O}(1/k^2)$ in function value where k is the iteration number [57].

2.3.1 Unconstrained Tree-based MRI

Following overlapping group sparsity algorithms [52, 49], The unconstrained MRI problem with tree sparsity can be formulated as:

$$\hat{x} = \min_x \{\frac{1}{2}\|Rx - b\|_2^2 + \beta \sum_{g \in \mathcal{G}} ||\Phi x||_2\} \tag{2.3}$$

where x is the MR image to be reconstructed, R is the partial Fourier transform, b is the measurement vector, Φ denotes the wavelet transform, β is a positive parameter need to be tuned. Here, g denotes one of the groups that encourages tree sparsity (e.g. one node and its parent) and $\mathcal{G}$ denotes the set of all such groups. Due to the nonsmoothness and nonseparability of the overlapping group penalty, it is not easy to solve the problem directly. Instead, we introduce a variable z to constrain the problem:

$$\hat{x} = \arg\min_{x,z} \{\frac{1}{2}\|Rx - b\|_2^2 + \beta \sum_{g \in \mathcal{G}} ||z_g||_2 + \frac{\lambda}{2}||z - G\Phi x||_2^2\} \tag{2.4}$$

where λ is another positive parameter, G is a binary matrix to duplicate the overlapped entries. z is the extended vector of wavelet coefficients x without overlapping.

All terms in our model are convex. For the z subproblem:

$$z_g = \arg\min_{z_g}\{\beta||z_g||_2 + \frac{\lambda}{2}||z_g - (G\Phi x)_g||_2^2\}, \qquad g \in \mathcal{G} \tag{2.5}$$

It has closed form solution by soft thresholding:

$$z_g = \max(||r||_2 - \frac{\beta}{\lambda}, 0)\frac{r}{||r||_2}, \qquad g \in \mathcal{G} \tag{2.6}$$

where $r = (G\Phi x)_g$. We denote this step by $z = shrinkgroup(G\Phi x, \frac{\beta}{\lambda})$ for convenience. For the x-subproblem:

$$x = \arg\min_{x}\{\frac{1}{2}\|Rx - b\|_2^2 + \frac{\lambda}{2}||z - G\Phi x||_2^2\} \tag{2.7}$$

This is a combination of two quadratic terms and has closed form solution: $x = (R^TR + \lambda\Phi^TG^TG\Phi)^{-1}(R^Tb + \Phi^TG^Tz)$. However, the inverse of $R^TR + \lambda\Phi^TG^TG\Phi$ is not easily obtained. In order to validate the benefit of tree structure, we apply FISTA to solve the x subproblem, which can match the convergence rate of FCSA. Let $f(x) = \frac{1}{2}\|Rx - b\|_2^2 + \frac{\lambda}{2}||z - G\Phi x||_2^2$, which is a convex and smooth function with Lipschitz L_f, and $g(x) = 0$. Then our algorithm can be summarized in Algorithm 1, which called FISTA_Tree. Here $\nabla f(r^k) = R^T(Rr^k - b) + \lambda\Phi^TG^T(G\Phi r^k - z)$. R^T and

Algorithm 1 FISTA_Tree

Input: $\rho = 1/L_f$, $r^1 = x^0$, $t^1 = 1, \beta, \lambda, N$
for $k = 1$ **to** N **do**
  $z = shrinkgroup(G\Phi x^{k-1}, \beta/\lambda)$
  $x^k = r^k - \rho\nabla f(r^k)$
  $t^{k+1} = [1 + \sqrt{1 + 4(t^k)^2}]/2$
  $r^{k+1} = x^k + \frac{t^k - 1}{t^{k+1}}(x^k - x^{k-1})$
end for

Φ^T denote the inverse partial Fourier transform and the inverse wavelet transform.

Computational complexity. Note that $G \in \mathbb{R}^{N' \times N}$ is a sparse matrix with each row containing only one nonzero element 1. Therefore, the multiplication by G only cost $\mathcal{O}(N') = \mathcal{O}(N)$ with our group configuration. Suppose x is an image with N pixels. The *shrinkgroup* step can implemented in only $\mathcal{O}(N \log N)$ time and the gradient step also takes $\mathcal{O}(N \log N)$. We can find the total time complexity in each iteration is still $\mathcal{O}(N \log N)$, the same as that of TVCMRI, RecPF and FCSA. This good feature guarantees the proposed algorithm could be comparable with the fastest MRI algorithms in terms of execution speed.

2.3.2 Constrained Tree-based MRI

NESTA [5] solves the constrained problem of standard sparsity:

$$\min_{\theta} ||\theta||_1, \quad s.t. \quad ||b - A\theta||_2 \leq \epsilon \tag{2.8}$$

where θ denotes the set of wavelet coefficients with $\theta = \Phi x$, $A = R\Phi^T$, Φ^T denotes the inverse wavelet transform, ϵ is a small constant. It reaches the optimal convergence rate for first order methods. Similar as the previous subsection, we extend it to solve the tree-based MRI problem:

$$\min_{\theta} ||G\theta||_{2,1}, \quad s.t. \quad ||b - A\theta||_2 \leq \epsilon \tag{2.9}$$

where $||G\theta||_{2,1} = \sum_{g \in \mathcal{G}} ||(G\theta)_g||_2$, and $g, \mathcal{G}$ are the same as those in Algorithm 1. Recall $\ell_{2,1}$ norm also have the form:

$$||G\theta||_{2,1} = \max_{u \in \mathcal{Q}} < u, Gx > \tag{2.10}$$

where the dual feasible set is:

$$\mathcal{Q} = \{u : ||u||_{2,\infty} \leqslant 1\} = \{u : \max_{g \in \mathcal{G}} ||u_g||_2 \leqslant 1\} \tag{2.11}$$

We relax the non-smooth $\ell_{2,1}$ norm to smooth function with:

$$f_\mu(\theta) = \max_{u \in \mathcal{Q}} (< u, G\theta > - \frac{\mu}{2}||u||_2^2) \tag{2.12}$$

where μ is a small fixed number.

Note that $(G\theta)_g = G_g\theta$ where G_g the rows of G correspond to group g. The first order gradient of $f_\mu(\theta)$ with Lipschitz constant L_μ is given by:

$$\bigtriangledown f_\mu(\theta)_g = \begin{cases} \mu^{-1} G_g^T G_g \theta, & ||G_g\theta||_2 < \mu \\ G_g^T G_g \theta / ||G_g\theta||_2, & \texttt{otherwise} \end{cases} \tag{2.13}$$

NESTA assumes the rows of the sampling matrix A are orthogonal, that is, $AA^T = I$ where I denotes the identical matrix. Fortunately, the partial Fourier transform in compressed sensing MRI satisfies this assumption: $AA^T = R\Phi^T\Phi R^T = RR^T = I$, where R^T denotes the inverse operator of R. The whole algorithm based in NESTA [5] framework is given in Algorithm 2.

Computational complexity. As shown in Algorithm 2, the complexity of the proposed algorithm the same as the original NESTA algorithm [5]. It is $6\mathcal{C}+\mathcal{O}(N)$, where $\mathcal{C}$ denotes the complexity of applying A or A^T. In CSMRI, $\mathcal{C} = \mathcal{O}(N \log N)$ if fast Fourier transform (FFT) is applied. Therefore, the total computational complexity is $\mathcal{O}(N \log N)$ for each iteration, the same as that of Algorithm 1.

If we compare the two types of algorithms, the parameters can be manually set in the unconstrained algorithm to determine how sparse the data is. Or the weights between sparseness and the least square fitting can be controlled. However, the constrained algorithm always seeks for the sparsest solution that satisfy the constrain. In the application of MRI, we find that if good parameters can be tuned, the unconstrained algorithm (Algorithm 1) performs better, or vice versa. In contrast, the constrained algorithm (Algorithm 2) has the convenience without tuning the parameter.

Algorithm 2 NESTA_Tree

Input: θ_0, ϵ, $k = 1$, L_μ, μ

while not meet the stopping criterion **do**

1. Compute $\bigtriangledown f_\mu(\theta)$
2. Compute y^k
   $q = \theta^k - L_\mu^{-1} \bigtriangledown f_\mu(\theta)$
   $\lambda_\epsilon = \max(0, \epsilon^{-1}||b - Aq||_2 - L_\mu)$
   $y^k = (I - \frac{\lambda_\epsilon}{\lambda_\epsilon + L_\mu} A^T A)(\frac{\lambda_\epsilon}{L_\mu} A^T b + q)$
3. Compute z^k
   $\alpha^k = 1/2(k+1)$
   $q = x_0 - L_\mu^{-1} \sum_{i \leq k} \bigtriangledown \alpha_i f_\mu(\theta)$
   $\lambda_\epsilon = \max(0, \epsilon^{-1}||b - Aq||_2 - L_\mu)$
   $z^k = (I - \frac{\lambda_\epsilon}{\lambda_\epsilon + L_\mu} A^T A)(\frac{\lambda_\epsilon}{L_\mu} A^T b + q)$
4. Update θ^k
   $\tau^k = 2(k+3)$
   $\theta^k = \tau^k z^k + (1 - \tau^k) y^k$
5. $k = k + 1$

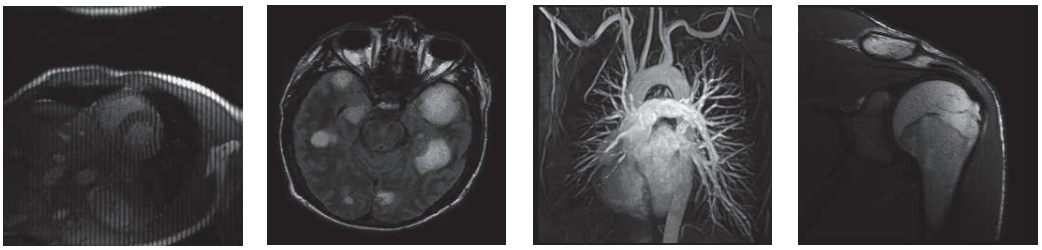
end while

2.4 Experiments

2.4.1 Experiment Setup

We compare the unconstrained algorithm FISTA_Tree with CG [1], TVCMRI [2], RecPF [3], FCSA [4] and compare the constrained algorithm NESTA_Tree with the original NESTA [5] algorithm for CSMRI. For fair comparisons, all code are downloaded from the authors' websites and we carefully follow their experiment setup. We apply all these methods on four real-valued MR images: cardiac, brain, chest and shoulder respectively (shown in Figure 2.2). In addition, a complex valued MR brain image[1] is added to validate the benefit of tree sparsity on complex-valued data. Suppose R is a partial Fourier transform with M rows and N columns. The sampling ratio

[1] `http://www.eecs.berkeley.edu/~mlustig/CS.html`



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

Figure 2.2. MR images: (a) cardiac; (b) brain; (c) chest and (d) shoulder.

is defined as M/N. For simulations with real-valued images, we follow the sampling strategy of previous works [2, 4], which randomly choose more Fourier coefficients from low frequency and less on high frequency. For complex-valued data, the radial sampling mask is used [3], which is more feasible in practical.

In order to study the benefit of tree structure in CSMRI, we remove the TV term in all algorithms. The parameters for real-valued images and complex-valued image are tuned separately. There is no continuation step [5] in the NESTA_Tree algorithm. All experiments are on a desktop with 3.4GHz Intel core i7 3770 CPU. Matlab version is 7.8(2009a). Measurements are added by Gaussian white noise with 0.01 standard deviation. Signal-to-Noise Ratio (SNR) is used for result evaluation:

$$SNR = 10\log_{10}(V_s/V_n) \tag{2.14}$$

where V_n is the Mean Square Error between the original image x_0 and the solution x; V_s denotes the variance of the values in x_0.

2.4.2 Group Configuration for Tree Sparsity

In all previous works, the tree structure are approximated as overlapping groups [49, 58]. In additional, all of them only consider each wavelet coefficient and its parent are assigned into one group. However, the relationship between one coefficient and its grandparent is not exploited. We first conduct an experiment to validate the influence

Table 2.1. Comparisons of SNR (db) on different group sizes for tree sparsity

$\beta\backslash$ group size	1	2	3	4
5×10^{-2}	17.30	**17.94**	16.45	15.33
10^{-2}	16.49	**16.99**	16.95	16.53
5×10^{-3}	16.36	16.62	**16.66**	16.48
10^{-3}	16.21	16.27	**16.29**	16.27

Table 2.2. Comparisons of computational costs (s) on different group sizes for tree sparsity

$\beta\backslash$ group size	1	2	3	4
5×10^{-2}	0.69	0.99	1.11	1.17
10^{-2}	0.70	0.95	1.09	1.15
5×10^{-3}	0.72	0.97	1.07	1.11
10^{-3}	0.70	0.97	1.07	1.11

of the group size to the reconstruction result. Four group sizes are compared: (a) each group one contains one coefficient, which is the same case as standard sparsity; (b) each group contains a coefficient and its parent, which is the same as previous works [49, 58]; (c) each group contains a coefficient, its parent and its grandparent; (d) each group contains 4 coefficients, where the grandparent's parent is also assigned in the same group.

With these group configurations, we test their performance on the FISTA_Tree algorithm, except the standard sparsity case is performed on FISTA. The parameter β determines how strong the tree sparsity assumption is. Table 2.1 and Table 2.2 show the average SNRs and CPU time on the four MR images with various parameter settings. With smaller parameters, the third group configuration performs the

best, while the second group configuration is the best with larger parameters. The computational time increase monotonously as the size of group becomes bigger. Due to the above two reasons, we encourage the use of the second group configuration on CS-MRI.

2.4.3 Visual Comparisons

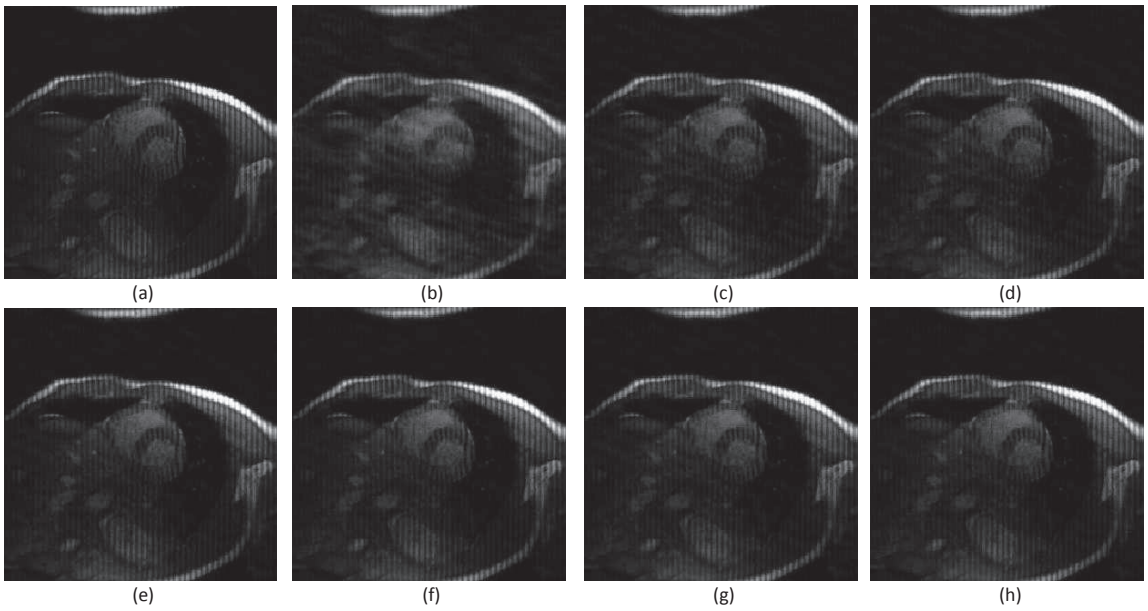


Figure 2.3. Cardiac MR image reconstruction from 20% sampling. (a) The original image. Recovered by : (b) CG [1]; (c) TVCMRI [2]; (d) RecPF [3]; (e) FCSA [4]; (f) FISTA_Tree; (g) NESTA [5]; (h) NESTA_Tree. All algorithms are without total variation regularization. Their SNR are 9.86, 14.70, 15.14, 17.31, 17.93, 16.31 and 16.96 respectively. Their computational time costs are 1.34 s, 1.12 s, 1.25 s, 0.67 s, 0.85 s, 0.88 s and 1.05 s.

We compare proposed tree-based algorithms with the fastest MRI algorithms to validate how much the tree structure can improve existing results. To perform fair comparisons, all methods run 50 iterations except that the CG runs only 8 iterations

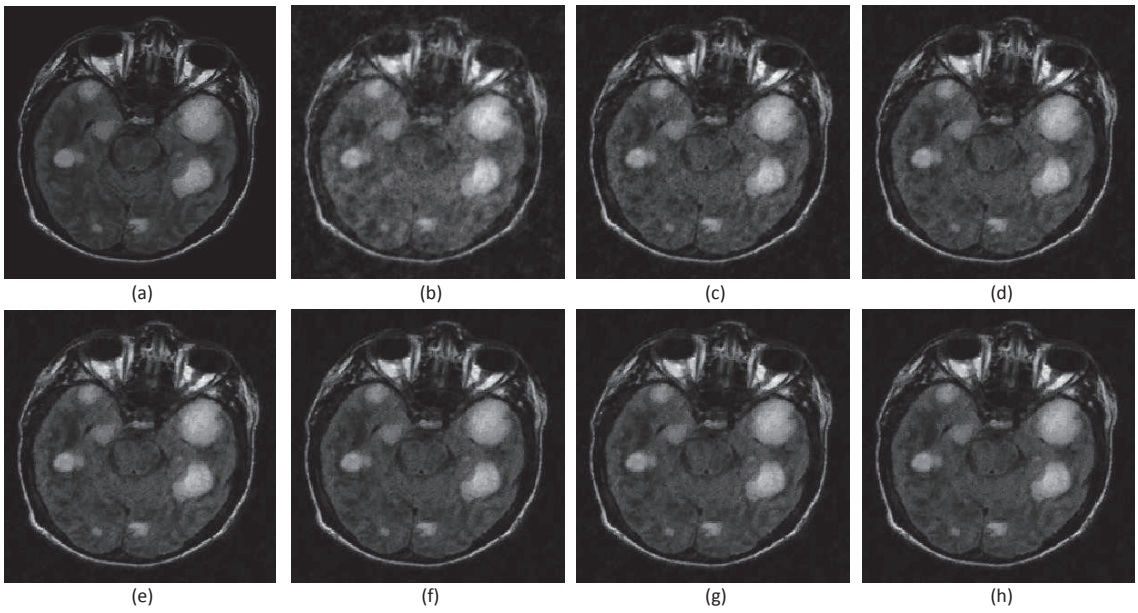


Figure 2.4. Brain MR image reconstruction from 20% sampling. (a) The original image. Recovered by : (b) CG [1]; (c) TVCMRI [2]; (d) RecPF [3]; (e) FCSA [4]; (f) FISTA_Tree; (g) NESTA [5]; (h) NESTA_Tree. All algorithms are without total variation regularization. Their SNR are 10.25, 13.81, 14.22, 15.65, 16.13, 15.05 and 15.52 respectively. Their computational time costs are 1.36 s, 1.11 s, 1.17 s, 0.71 s, 1.02 s, 0.91 s and 1.03 s.

due to its higher computational complexity. Total variation terms are removed in all algorithms, as we only want to validate how much benefit the wavelet tree sparsity can bring compared to standard wavelet sparsity. In this case, FCSA [4] is similar as FISTA [39]. Figure 2.3-2.6 shows the visual results on the four MR images with 20% sampling. It can be found that the proposed unconstrained algorithm FISTA_Tree is always better than CG [1], TVCMRI [2], RecPF [3] and FCSA [4]. These results are consistent with previous observations [4]. Compared the proposed NESTA_Tree with NESTA, our method is still much better. These results are reasonable because no structured prior information has been exploited in previous algorithms other than sparsity, while the tree structure in our algorithms is utilized. Any coefficient that

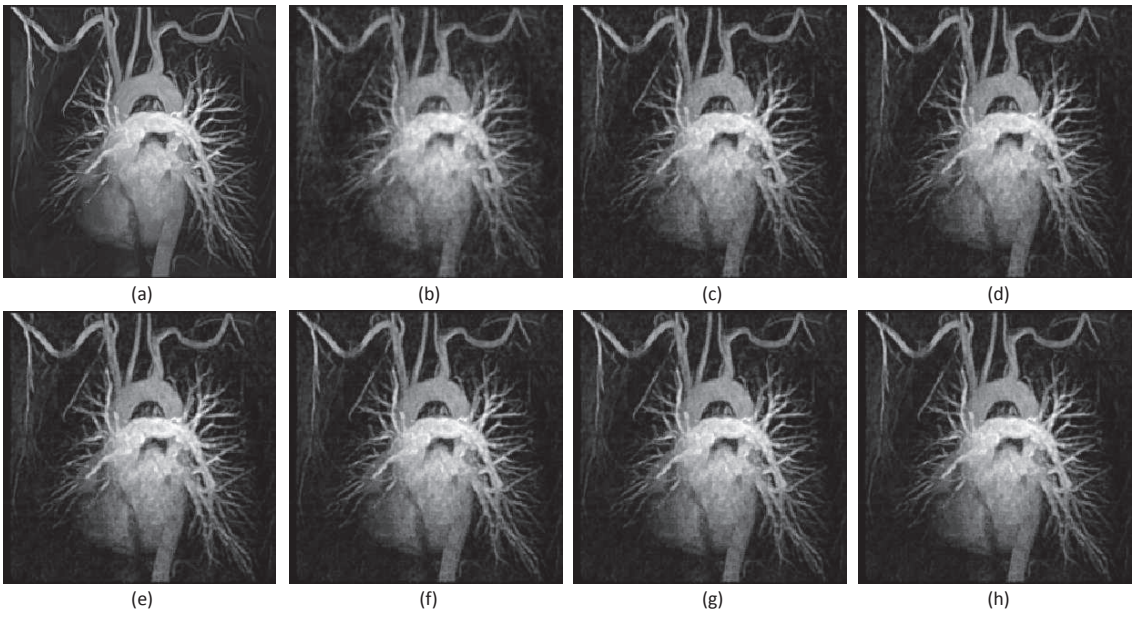


Figure 2.5. Chest MR image reconstruction from 20% sampling. (a) The original image. Recovered by : (b) CG [1]; (c) TVCMRI [2]; (d) RecPF [3]; (e) FCSA [4]; (f) FISTA_Tree; (g) NESTA [5]; (h) NESTA_Tree. All algorithms are without total variation regularization. Their SNR are 11.82, 15.09, 15.36, 15.98, 16.35, 15.91 and 16.30 respectively. Their computational time costs are 1.28 s, 1.12 s, 1.23 s, 0.67 s, 0.96 s, 0.84 s and 1.07 s.

disobeys the tree structure will be penalized in our algorithms, which makes the results closer to the original ones.

2.4.4 SNRs and CPU time

Figure 2.7 gives the performance comparisons between different methods in terms of SNR with 50 iterations. Due to the faster convergence rate of FISTA and NESTA, they always outperforms CG, TVCMRI and RecPF. Moreover, the tree-based algorithms approximated by overlapping group sparsity are always better than those with standard sparsity. Table 2.3 shows all computational costs of different algorithms. CG has the highest computational complexity. TVCMRI and RecPF is much

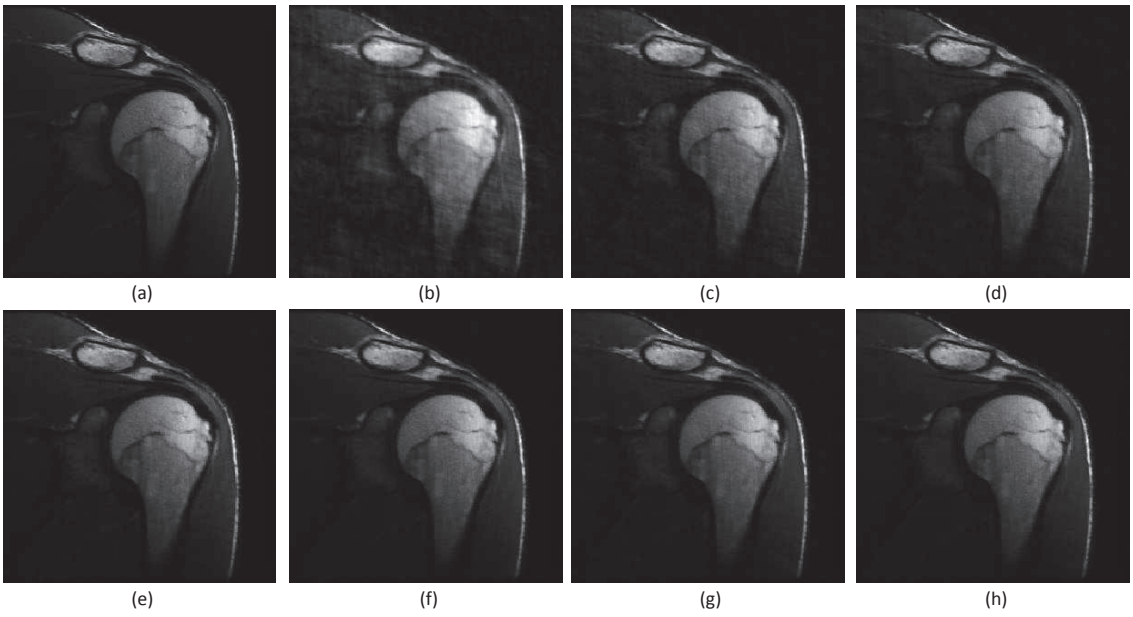


Figure 2.6. Shoulder MR image reconstruction from 20% sampling. (a) The original image. Recovered by : (b) CG [1]; (c) TVCMRI [2]; (d) RecPF [3]; (e) FCSA [4]; (f) FISTA_Tree; (g) NESTA [5]; (h) NESTA_Tree. All algorithms are without total variation regularization. Their SNR are 12.31, 16.80, 17.90, 20.77, 21.04, 20.17 and 20.62 respectively. Their computational time costs are 1.36 s, 1.07 s, 1.25 s, 0.67 s, 0.95 s, 0.82 s and 1.07 s.

faster than CG and slower than FCSA. It is to be expected that tree based algorithms FISTA_Tree and NESTA_Tree are slower than FISTA and NESTA respectively, since the overlapping structure needs more time for computing than non-overlapping structure. However, applying the wavelet transform and the Fourier transform is still the dominant cost, which is the same for all algorithms. As a result, FISTA_Tree and NESTA_Tree are comparable to the corresponding standard sparsity algorithms in term of reconstruction speed, and bring much more improvement on accuracy.

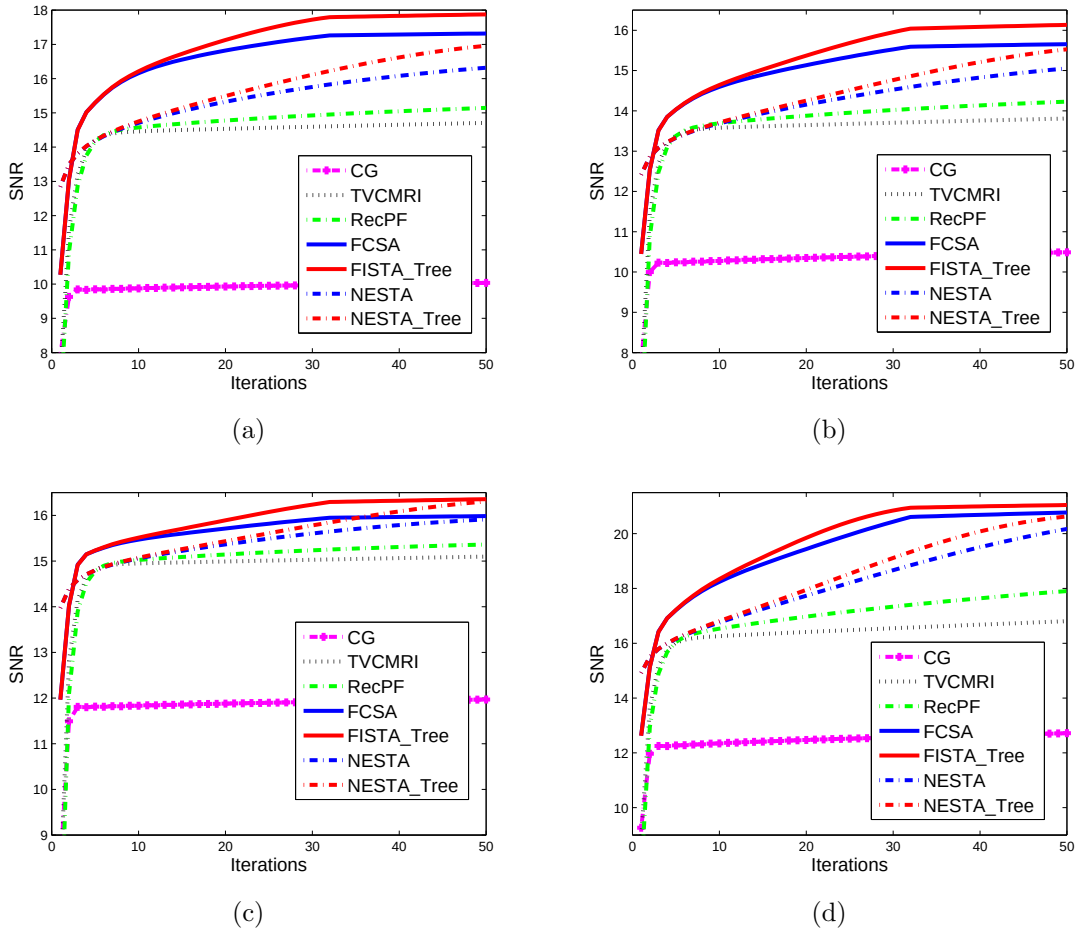


Figure 2.7. Performance comparisons (SNRs) on different MR images: (a) cardiac image; (b) brain image; (c) chest image and (d) shoulder image.

Table 2.3. Comparisons of average computational costs (s) on different MR images with 20% sampling

	Cardiac	Brain	Chest	Shoulder
CG	9.14	8.87	9.21	9.17
TVCMRI	1.12	1.11	1.12	1.07
RecPF	1.25	1.17	1.23	1.25
FCSA	0.67	0.71	0.67	0.67
FISTA_Tree	0.85	1.02	0.96	0.95
NESTA	0.88	0.91	0.84	0.82
NESTA_Tree	1.05	1.03	1.07	1.07

2.4.5 Sampling Ratios

All algorithms are compared under different sampling ratios on the four MR images. Since we have shown that the CG method is far less efficient than other methods, we do not include it in this experiment. To reduce the randomness, we run each experiments 100 times to obtain the average results of each method. The sampling ratio ranges from 17% to 25%. Figure 3.4 shows these results on the four images. We could observe that TVCMRI and RecPF are not comparable to recent algorithms with fast convergence rate. Under the same framework and with similar convergence rate, the tree-based algorithms (i.e. FISTA_Tree and NESTA_Tree) are always better than the corresponding standard sparsity algorithms (i.e. FCSA and NESTA) respectively. These results further demonstrate the benefit of tree sparsity in accelerated MRI.

2.4.6 Complex-valued with Radial Sampling Mask

We have observed the superior performance for tree-based MRI algorithms from numerical simulations. In this subsection, we validate their performance on a complex-valued MR image with 512×512 pixels. The sampling mask is radial mask, which is more feasible than the random sampling mask in practical. With previous results, we only compare the classical method CG [1] and the fastest algorithm FCSA [4] with the proposed tree-based algorithms.

Figure 2.9 presents the visual results reconstructed by difference methods. The image with full sampling is used as reference image. We could observe that tree-based algorithms FISTA_Tree and NESTA_Tree achieve higher SNR than standard sparsity algorithms CG and FCSA. Due to the relative slow convergence rate of CG, it is still not converged after 50 iterations. That is why it has inferior performance to FCSA. This data is scanned with noise. Therefore, we also compare image quality besides

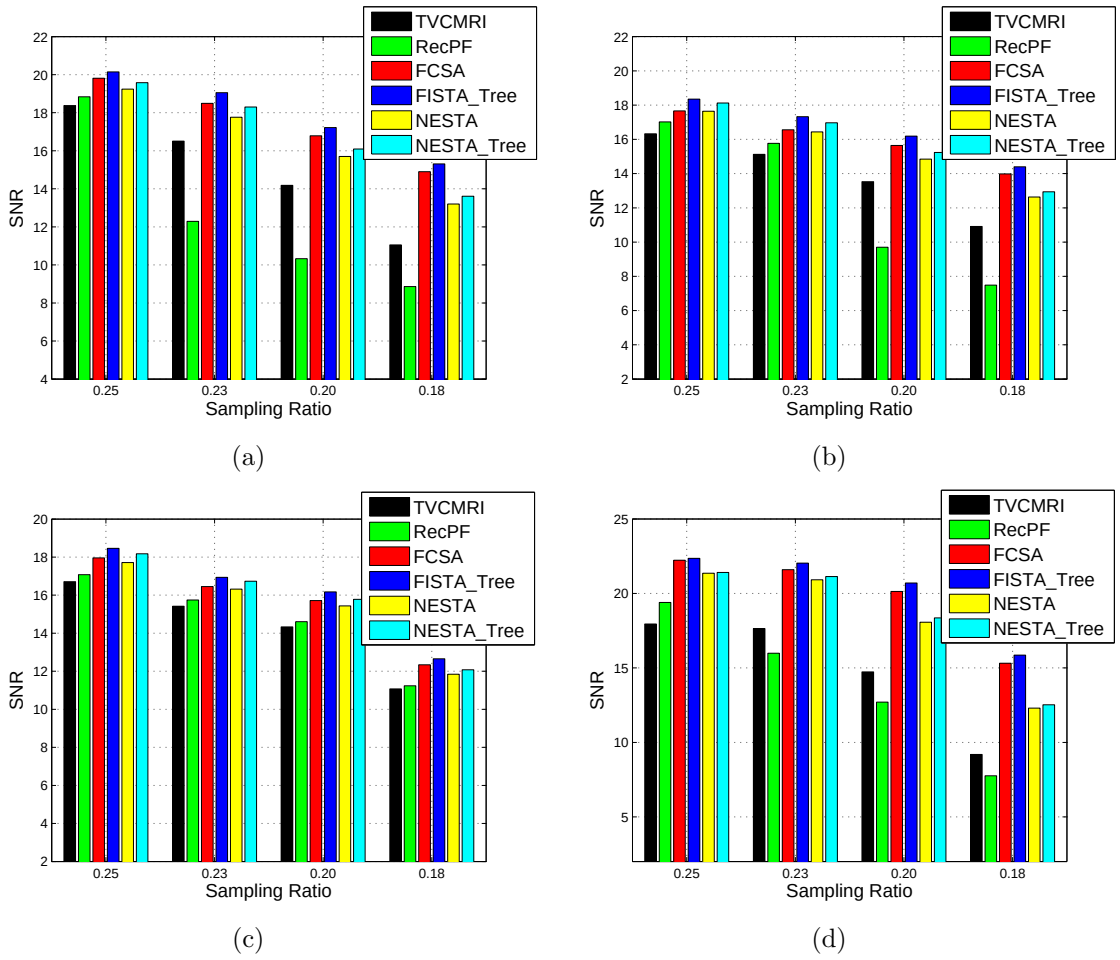


Figure 2.8. Performance comparisons on the four MR images with different sampling ratios: (a) cardiac image; (b) brain image; (c) chest image and (d) shoulder image.

SNR. From the zoomed in areas, image details are lost in the image reconstructed by CG and blurred in that reconstructed by FCSA. However, both tree-based algorithms can preserve significant features on the MR image even with a low sampling ratio.

2.5 Summary

In order to validate the benefit of wavelet tree sparsity in MR image reconstruction, we propose two tree-based algorithms for CS-MRI and compare them with the state-of-the-art algorithms based on standard sparsity. In order to observe the

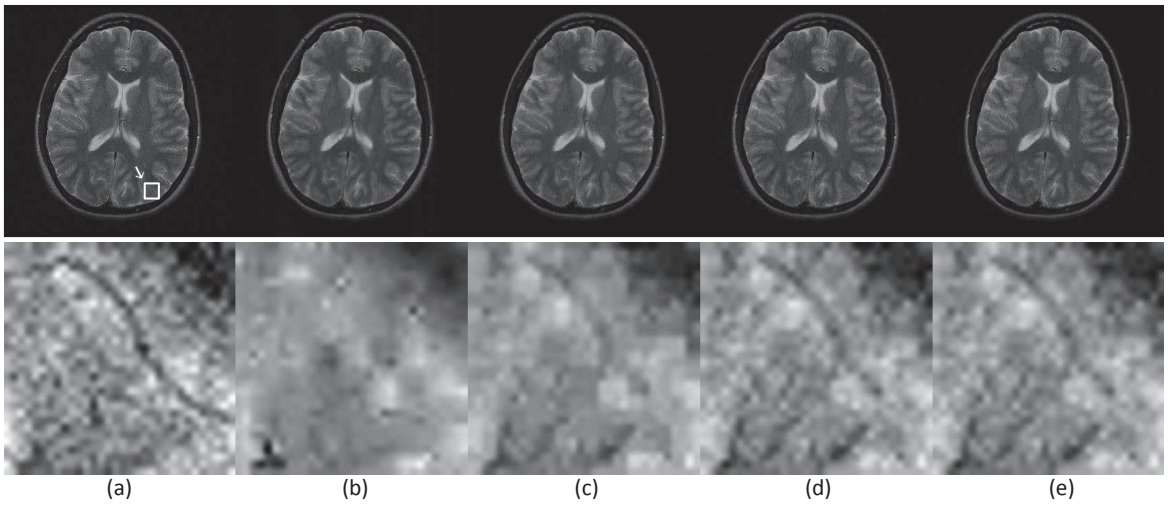


Figure 2.9. Reconstruction on complex-valued MR image with 20% sampling. The first shows the visual results of by different algorithm. The second row shows the zoomed in areas indicated by the white boxes. (a) Inverse FFT with full sampling. (b) CG. (c) FCSA. (d) FISTA_Tree. (e) NESTA_Tree. Their SNRs are 15.31, 16.32, 16.65, 16.68, respectively.

benefit of tree sparsity more clearly, total variation terms are removed in all algorithms. Extensive experiments are conducted to show the practical improvement of the proposed tree-based algorithm on MR images. The results tell that the benefit of the proposed algorithm is far from the conclusion in structured sparsity theory. That is because the tree structure is not as strictly as the structured sparsity theories assumed on practical data. In addition, the approximation in our implementations may be not precise enough. To bridge the gap between the theories and practice, future work will weight the wavelet coefficients on different levels differently, but not treat them equally.

CHAPTER 3

Forest Sparsity for Multi-channel Compressive Sensing

This chapter extends the tree sparsity idea to forest sparsity on multi-channel data. A theory is developed for forest sparsity based on compressed sensing. It shows the theoretical benefits of forest sparsity. This work was presented under a slightly modification from [59].

3.1 Introduction

Sparsity techniques are becoming more and more popular in machine learning, statistics, medical imaging and computer vision as the emerging of compressed sensing. Based on compressed sensing theory [37, 18], a small number of measurements are enough to recover the original data, which is an alternative to Shannon/Nyquist sampling theorem for sparse or compressible data acquisition.

3.1.1 Standard Sparsity and Algorithms

Suppose $A \in \mathbb{R}^{M\times N}$ is the sampling matrix and $b \in \mathbb{R}^M$ is the measurement vector, the problem is to recover the sparse data $x \in \mathbb{R}^N$ by solving the linear system $Ax = b$. Sometimes the data is not sparse but compressible under some base Φ such as wavelet, and the corresponding problem is $A\Phi^{-1}\theta = b$ where θ denotes the set of wavelet coefficients. Although the problem is underdetermined, the data can be perfectly reconstructed if the sampling matrix satisfy restricted isometry property

(RIP) [19] and the number of measurements is larger than $\mathcal{O}(k + k\log(N/k))$ for k-sparse data[1] [60, 61].

To solve the underdetermined problem, we may find the sparsest solution via ℓ_0 norm regularization. However, because the problem is NP-hard [62] and impractical for most applications , ℓ_1 norm regularization methods such as the lasso [63] and basis pursuit (BP) [64] are first used to pursue the sparse solution. It has been proved that the ℓ_1 norm regularization can exactly recover the sparse data for CS inverse problem under mild conditions [65, 37]. Therefore, a lot of efficient algorithms have been proposed for standard sparse recovery. Generally speaking, those algorithms can be classified into three groups: greedy algorithms [66, 67], convex programming [42, 68, 69] and probability based methods [70, 71].

3.1.2 Joint Sparsity and Algorithms

Beyond standard sparsity, the non-zeros components of x often tend to be in some structures. This comes to the concept of *structured sparsity* or model-based compressed sensing [30, 29, 31]. In contrast to standard sparsity that only relies on the sparseness of the data, structured sparsity models exploit both the non-zero values and the corresponding locations. For example, in the multiple measurement vector (MMV) problem, the data is consisted of several vectors that share the same support [2]. This is called *joint sparsity* that widely arise in cognitive radio networks [72], direction-of-arrival estimation in radar [73], multi-channel compressed sensing [74, 75], remote sensing [76] and medical imaging [77, 78]. If the data $X \in \mathbb{R}^{TN\times 1}$ is consist of T k-sparse vectors, the measurement bound could be substantially reduced to $\mathcal{O}(Tk+ k\log(N/q))$ instead of $\mathcal{O}(Tk + Tk\log(N/q))$ for standard sparsity [79, 30, 29, 80].

[1] We mean there are at most k non-zero components in the data.

[2] The set of indices corresponding to the non-zero entries is often called the support

A common way to implement joint sparsity in convex programming is to replace the ℓ_1 norm with $\ell_{2,1}$ norm, which is the summation of ℓ_2 norms of the correlated entries [81, 82]. $\ell_{2,1}$ norm for joint sparsity has been used in many convex solvers and algorithms [83, 78, 40, 16]. In Bayesian sparse learning or approximate message passing [84, 85, 86], data from all channels contribute to the estimation of parameters or hidden variables in the sparse prior model.

3.1.3 Tree Sparsity and Algorithms

Another common structure would be the hierarchical tree structure, which has already been successfully utilized in image compression [43], compressed imaging [50, 13, 49, 14], and machine learning [87]. Most nature signals/images are approximately tree-sparse under the wavelet basis. A typical relationship with *tree sparsity* is that, if a node on the tree is non-zero, all of its ancestors leading to the root should be non-zeros. For multi-channel data $X = [x_1; x_2, ...; x_T]$[3] $\in \mathbb{R}^{NT\times 1}$, $\mathcal{O}(Tk+T\log(N/k))$ measurements are required if each channel x_t is tree-sparse.

Due to the overlapping and intricate structure of tree sparsity, it is much harder to implement. For greedy algorithms, StructOMP [30] and TOMP [48] are developed for exploiting tree structure where the coefficients are updated by only searching the subtree blocks instead of all subspace. In statistical models [50, 13], hierarchical inference is used to model the tree structure, where the value of a node is not independent but relies on the distribution or state of its parent. In convex programming [49, 88], due to the tradeoff between the recovery accuracy and computational complexity, this is often approximated as overlapping group sparsity [52], where each node and its parent are assigned into one group.

[3] In this article, [;] denotes concatenating the data vetically.

3.1.4 Forest Sparsity

Although both joint sparsity and tree sparsity have been widely studied, unfortunately, there is no work that study the benefit of their combinations so far. Actually, in many multi-channel compressed sensing or MMV problems, the data has joint sparsity across different channels and each channel itself is tree-sparse. Note that this differs from C-HiLasso [89], where sparsity is assumed inside the groups. No method has fully exploited both priors and no theory guarantees the performance. In practical applications, researchers and engineers have to choose either joint sparsity algorithms by giving up their intra tree-sparse prior, or tree sparsity algorithms by ignoring their inter correlations.

In this chapter, we propose a new sparsity model called *forest sparsity* to bridge this gap. It is a natural extension of existing structured sparsity models by assuming that the data can be represented by a forest of mutually connected trees. We give the mathematical definition of forest sparsity. Based on compressed sensing theory, we prove that for a forest of T k-sparse trees, only $\mathcal{O}(Tk + \log(N/k))$ measurements are required for successful recovery with high probability. That is much less than the bounds of joint sparsity $\mathcal{O}(Tk + k\log(N/k))$ and tree sparsity $\mathcal{O}(Tk + T\log(N/k))$ on the same data. The theory is further extended to the case on MMV problems, which is ignored in existing structured sparsity theories [30, 29, 31]. Finally, we derive an efficient algorithm to optimize the forest sparsity model. The proposed algorithm is applied on medical imaging applications such as multi-contrast magnetic resonance imaging (MRI), parallel MRI (pMRI), as well as color images, multispectral image reconstruction. Extensive experiments demonstrate the advantages of forest sparsity over the state-of-the-art methods in these applications.

3.2 Background and Related Work

In compressed sensing (CS), the capture of a sparse signal and compression are integrated into a single process [18, 19]. We do not capture sparse data $x \in \mathbb{R}^N$ directly but rather capture $M < N$ linear measurements $b = Ax$ based on a measurement matrix $A \in \mathbb{R}^{M \times N}$. To stably recover the k-sparse data x from M measurements, the measurement matrix A is required to satisfy the Restricted Isometry Property (RIP) [19]. Let Ω_k denote the union k-dimensional subspaces where x lives in.

Definition 1: (k-RIP) *An $M \times N$ matrix A has the k-restricted isometry property with restricted isometry constant $1 > \delta_k > 0$, if for all $x \in \Omega_k$, and*

$$(1 - \delta_k)||x||_2^2 \leq ||Ax||_2^2 \leq (1 + \delta_k)||x||_2^2. \tag{3.1}$$

CS result shows that, if $M = \mathcal{O}(k+k\log(N/k))$, a sub-Gaussian random matrix[4] A can satisfy the RIP with high probability [91, 92].

Recently, structured sparsity theories demonstrate that when there is some structured prior information (e.g. group, tree, graph) in x, the measurement bound could be reduced [29, 30]. Suppose x is in the union of subspaces $\mathcal{A}$, then the k-RIP can be extended to the $\mathcal{A}$-RIP [92]:

Definition 2: ($\mathcal{A}$-RIP) *An $M \times N$ matrix A has the $\mathcal{A}$-restricted isometry property with restricted isometry constant $1 > \delta_{\mathcal{A}} > 0$, if for all $x \in \mathcal{A}$, and*

$$(1 - \delta_{\mathcal{A}})||x||_2^2 \leq ||Ax||_2^2 \leq (1 + \delta_{\mathcal{A}})||x||_2^2. \tag{3.2}$$

$\mathcal{A}$-RIP property has been proved to be sufficient for robust recovery of structured-sparse signals under noisy conditions [29]. The required number of measurements M has been quantified for a sub-Gaussian random matrix A that has the $\mathcal{A}$-RIP [92]:

[4] It includes Gaussian and Bernoulli random matrices etc. [90].

Theorem 1: ($\mathcal{A}$-RIP) *Let $\mathcal{A}$ be the union of L subspaces of k dimension in $\mathbb{R}^N$. For any $t > 0$, let*

$$M \geq \frac{2}{c\delta_{\mathcal{A}_k}}(\ln(2L) + k \ln \frac{12}{\delta_{\mathcal{A}_k}} + t), \tag{3.3}$$

then there exists a constant $c > 0$ and a randomly generated sub-Gaussian matrix $A \in \mathbb{R}^{M \times N}$ satisfies the $\mathcal{A}$-RIP with probability at least $1 - e^{-t}$.

From (3.3), we could intuitively observe that M can be less by reducing the number of subspaces $\mathcal{A}$. It coincides with the intuition that the result will be improved when more priors are utilized. For standard k-sparse data, there is no more constraint to reduce the number of possible subspaces C_N^k. Let $L = C_N^k \approx (eN/k)^k$, the CS result for standard sparsity can be derived from Theorem 1.

Now we consider structured sparse data. Following [29], if a k-sparse data $x \in \mathbb{R}^N$ can form a tree or can be sparsely represented as a tree under one orthogonal sparse basis Φ (e.g. wavelet), and the k non-zero components naturally form a subtree, then it is called tree-sparse data.

Definition 3: *Tree-sparse data in $\mathbb{R}^N$ is defined as*
$\mathcal{T}_k$=$\{x = \Phi^{-1}\theta$*:* $\theta|\Omega^C = 0$*,* $|\Omega| = k$*, where Ω forms a connected subtree.* $\}$.

Here $\Omega \subseteq \{1, 2, ..., N\}$ denotes a subspace of the data as and the support is in Ω. Ω^C denotes the complement of Ω and θ denotes the coefficients under Φ. It implies that, if an entry of θ is in Ω , all its ancestors on the tree must be in Ω.

For tree-sparse data, we say it has the *tree sparsity* property. Most natural signals or images have tree sparsity property, since they can be sparsely represented with the wavelet tree structure. Specially, the wavelet coefficients of a 1D signal form a binary tree and those of a 2D image yield a quadtree. If the union of all subspaces are denoted by Ω_{Tree}, it is obviously that $\Omega_{Tree} \subset \Omega_k$ and the number of subspaces $L_{Tree} < C_N^k$.

Theorem 2: *For tree-sparse data, there exists a sub-Gaussian random matrix* $A \in \mathbb{R}^{M\times N}$ *that has the* $\mathcal{T}_k$*-RIP with probability* $1-e^{-t}$ *if the number of measurements satisfies that:*

$$M \geq \begin{cases} \frac{2}{c_1\delta_{\mathcal{T}_k}}(k + \ln(N/(k+1)) + k\ln(12/\delta_{\mathcal{T}_k}) \\ \quad + \ln 2 + t) \qquad \text{if } k \leq \lfloor \log_2 N \rfloor, \\ \frac{2}{c_1\delta_{\mathcal{T}_k}}(k\ln 4 + \ln(c_2 N/k) + k\ln(12/\delta_{\mathcal{T}_k}) \\ \quad + \ln 2 + t) \qquad \text{if } k > \lfloor \log_2 N \rfloor. \end{cases} \tag{3.4}$$

where c_1 *and* c_2 *denote absolute constants.*

For both case, we have $M = \mathcal{O}(k + \log(N/k))$. Similar conclusion has been drawn in previous articles [30][29].

So far, we have reviewed standard sparsity and tree sparsity on single channel data. For multi-channel data that contains T channels or vectors (i.e. $X = [x_1; x_2; ...; x_T] \in \mathbb{R}^{NT\times 1}$), each of which is standardly k-sparse, the bound for the number of measurement should be $\mathcal{O}(Tk + Tk\log(N/k))$. If each channel is tree-sparse and independently, the measurement bound for a sub-Gaussian random matrix $A \in \mathbb{R}^{TM\times TN}$ is $TM = \mathcal{O}(Tk + T\log(N/k))$.

It is important to note that the T-channel k-sparse data has sparsity Tk but not k. Different from the above independent channels, another case is that all channels of the data may be highly correlated, which corresponds to joint sparse data:

Definition 4: *Joint-sparse data is defined as*

$\mathcal{J}_{T,k}$=$\{X = [x_1; x_2; ...; x_T]$*:* $x_i = \Phi^{-1}\theta_i$*,* $\theta_i|\Omega^C = 0$*,* $|\Omega| = k$*,* $i = 1, 2, ..., T$ $\}$.

Similar as tree-sparse data, joint-sparse data has the *joint sparsity* property. It has to be clarified that joint sparsity does not rely on tree sparsity. The former utilizes the structure across different channels, while the later utilizes the structure

within each channel. Previous works implies that the minimum measurement bound for such joint sparse data is $TM = \mathcal{O}(Tk + k\log(N/k))$ [79, 30, 29].

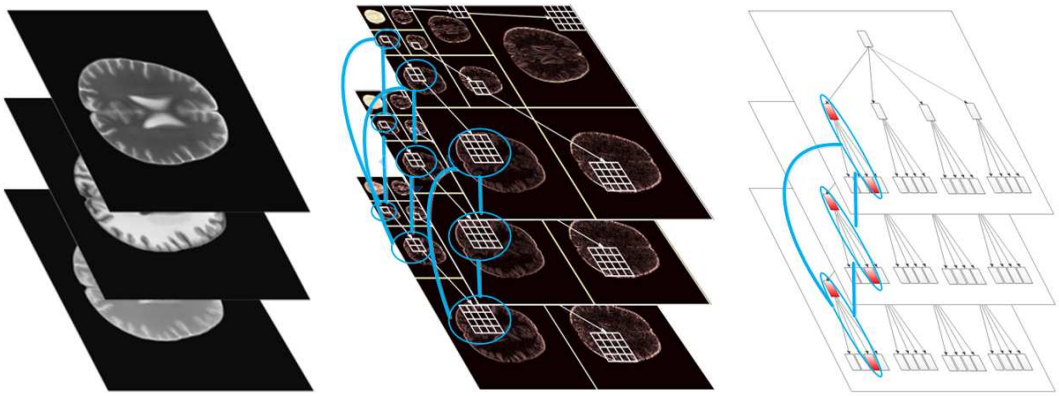
3.3 Forest Sparsity

In practical applications, it happens usually that multi-channel images, such as color images, multispectral images and MR images, have the joint sparsity and tree sparsity simultaneously. It is because: (a) the wavelet coefficients of each channel naturally yield a quadtree; (b) all channels represent the same physical objects (e.g. nature scenes or human organs), and the wavelet coefficients of each channel tend to be large/small simultaneously due to same boundaries of the objects. Therefore, the support of such data is consist of several connected trees and like a forest. Fig. 3.1 shows the forest structure in multi-contrast MR images. We could find that the non-zero coefficients are not random distributed but forms a connected forest. Unfortunately, existing tree-based algorithms can only recover multi-channel data channel-by-channel separately, and it is unknown how to model the tree structure in existing joint sparsity algorithms. In addition, there are no theoretical results in previous works showing how much better the recovery can be improved by fully exploiting the prior information.

Motivated by this limitation, we extend previous works to a more special but widely existed case. For multi-channel data, if it is jointly sparse, and more importantly, the common support of different channels yields a subtree structure, we call this kind of data forest-sparse data:

Definition 5: Forest-sparse data is defined as

$\mathcal{F}_{T,k}$=$\{X = [x_1; x_2; ...; x_T]$*:* $x_i = \Phi^{-1}\theta_i$, $\theta_i|\Omega^C = 0$, $|\Omega| = k$, *where* Ω *forms a connected subtree,* $i = 1, 2, ..., T$ $\}$.



(a) (b) (c)

Figure 3.1. Forest structure on multi-contrast MR images. (a) Three multi-contrast MR images. (b) The wavelet coefficients of the images. Each coefficient tends to be consistent with its parent and children, and the coefficients across different trees at the same position. (c) One joint parent-child group across different trees that used in our algorithm.

Similarly, the forest-sparse data has *forest sparsity* property. This definition implies that if the coefficients at the same position across different channels are non-zeros, all their ancestor coefficients are all non-zeros. Learning with forest sparsity, we search the sparsest solution that follow the forest structure in the CS inverse problem. Any solution that violates the assumption will be penalized. Intuitively, the solution will be more accurate. We obtain our main result in the following theorem:

Theorem 3: *For forest-sparse data, there exists a sub-Gaussian random matrix* $A \in \mathbb{R}^{TM \times TN}$ *that has the* $\mathcal{F}_{T,k}$*-RIP with probability* $1 - e^{-t}$ *if the number of measurements satisfies that:*

$$TM \geq \begin{cases} \frac{2}{c_1 \delta_{\mathcal{F}_{T,k}}} (k + \ln(N/(k+1)) + Tk \ln(12/\delta_{\mathcal{F}_{T,k}}) \\ \qquad + \ln 2 + t) \qquad \text{if } k \leq \lfloor \log_2 N \rfloor, \\ \frac{2}{c_1 \delta_{\mathcal{F}_{T,k}}} (k \ln 4 + \ln((c_2 N)/k) + Tk \ln(12/\delta_{\mathcal{F}_{T,k}}) \\ \qquad + \ln 2 + t) \qquad \text{if } k > \lfloor \log_2 N \rfloor. \end{cases} \tag{3.5}$$

where c_1 *and* c_2 *are absolute constants.*

For both cases, the bound is reduced to $M = \mathcal{O}(Tk + \log(N/k))$. The proofs of Lemma 1 as well as Lemma 2, 4 are included in the appendices. Using the $\mathcal{F}_{T,k}$-RIP, forest-sparse data can be robustly recovered from noisy compressive measurements.

Table 3.1 lists all the measurement bounds for the forest-sparse data with different models. Standard sparsity model only exploits the sparseness while no prior information about the locations of the non-zero elements is involved. It is the classical but worst model for forest-sparse data. These location priors are partially utilized in joint sparsity and tree sparsity models. One of them only studies the correlations across channels, while the other one only learns the intra structure. Our result is significantly better than those of joint sparsity and tree sparsity, and far better than that of standard sparsity, especially when N/k is large. Only the proposed model fully exploits all these structures.

Table 3.1. Measurement bounds for forest-sparse data

Sparse Models	Measurement Bounds
Standard Sparsity	$\mathcal{O}(Tk + Tk\log(N/k))$
Joint Sparsity	$\mathcal{O}(Tk + k\log(N/k))$
Tree Sparsity	$\mathcal{O}(Tk + T\log(N/k))$
Forest Sparsity	$\mathcal{O}(Tk + \log(N/k))$

So far, we have analyzed the result by forest sparsity over previous results. In all these results, the measurement matrix A is assumed to be a dense sub-Gaussian matrix. However, in many practical problems, each data channel $x_t \in \mathbb{R}^N$ is measured by a distinct compressive matrix $A'_t \in \mathbb{R}^{M\times N}$, $t = 1, 2, ..., T$, which are called multiple measurement vectors (MMV) problems or multi-task learning (e.g., [74, 93, 94]). Here and later, we assume that $\{A'_t\}_{t=1}^T$ follow the same distribution but may be

different. Therefore, the matrix A is actually a block-diagonal matrix rather than a dense matrix. The linear system $b = Ax$ can be written as:

$$\begin{bmatrix} b_1 \\ b_2 \\ ... \\ b_T \end{bmatrix} = \begin{bmatrix} A'_1 & & & \\ & A'_2 & & \\ & & ... & \\ & & & A'_T \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_T \end{bmatrix} \tag{3.6}$$

The non-diagonal blocks in A are all zeros. Intuitively, such block-diagonal matrices have no better results than the dense matrices that discussed above, due to the less randomness. Unfortunately, the performance of the random block-diagonal matrices has not been analyzed on structured sparse data before, as all existing structured sparsity theories concentrate on the dense random matrix [30, 29, 31]. In this article, we extend the theoretical result to the block-diagonal matrix in the MMV problems.

Theorem 4: *For forest-sparse data, there exists a block-diagonal matrix A composed by sub-Gaussian random matrices $\{A'_t\}_{t=1}^T$ as in (3.6), that has the $\mathcal{F}_{T,k}$-RIP with probability $1 - e^{-t}$ if the number of measurements satisfies that:*

$$TM \geq \begin{cases} \frac{2T}{c_1 W}(\ln 2 + \ln(N/(k+1)) + Tk\ln(12/\delta_{\mathcal{F}_{T,k}}) \\ \qquad +k+t), \qquad \text{if } k \leq \lfloor \log_2 N \rfloor, \\ \frac{2T}{c_1 W}(\ln 2 + \ln((c_3 N)/k) + Tk\ln(12/\delta_{\mathcal{F}_{T,k}}) \\ \qquad +k\ln 4 + t), \qquad \text{if } k > \lfloor \log_2 N \rfloor. \end{cases} \tag{3.7}$$

where $W = \min(c_2^2\delta^2_{\mathcal{F}_{T,k}}\Gamma_2, c_2\delta_{\mathcal{F}_{T,k}}\Gamma_\infty)$; $\Gamma_2 = \frac{(\sum_{t=1}^T ||x_t||_2^2)^2}{\sum_{t=1}^T ||x_t||_2^4}$ and $\Gamma_\infty = \frac{\sum_{t=1}^T ||x_t||_2^2}{\max_{t=1}^T ||x_t||_2^2}$; c_1, c_2 *and* c_3 *are absolute constants.*

For both cases, the bound can be written as $TM = \mathcal{O}(\frac{T^2k + T\log(N/k)}{\min(\Gamma_2, \Gamma_\infty)})$. In contrast to previous results on dense matrices with i.i.d sub-Gaussian entries, this bound also depends on the energy of the data. It is not hard to find that $1 \leq \Gamma_2 \leq T$ and $1 \leq \Gamma_\infty \leq T$. In the best case, when $||x_1||_2 = ||x_2||_2 = ... = ||x_T||_2$ and $\Gamma_2 = \Gamma_\infty = T$,

the measurement bound is $TM = \mathcal{O}(Tk + \log(N/k)$. It shows a similar performance as the dense sub-Gaussian matrix in Theorem 3. In the worst case, the energy of the data concentrate on one single channel/task, i.e., all $||x_t||_2 = 0$ except a single index $||x_{t'}||_2 \neq 0$. The measurement bound then is $TM = \mathcal{O}(T^2k + T\log(N/k)$, which is even worse than that in Theorem 2 for independent tree sparse channels. Even for the same block-diagonal matrix, the analysis makes clear that its performance may varies significantly depending on the data being measured. In the worst case, their measurement bound can increase T times. However, the increased factor $T/\min(c_2^2\delta^2\Gamma_2, c_2\delta\Gamma_\infty)$ for block-diagonal matrices also applies to standard sparse data, joint sparse data and tree sparse data. For the same measurement matrix and the same data, the advantage of forest sparsity still exists. Due to this reason, we do not evaluate the term $\min(c_2^2\delta^2\Gamma_2, c_2\delta\Gamma_\infty)$ in the experiments, while focus our interest on comparing different sparsity models on the same data.

3.4 Algorithm

In this chapter, the forest structure is approximated as overlapping group sparsity [52] with mixed $\ell_{2,1}$ norm. Although it may not be the best approximation, it is enough to demonstrate the benefit of forest sparsity. To evaluate the forest sparsity model, we need to compare different models via a similar framework. From the definition of forest-sparse data, we could find that a coefficient is large/small, its parent and "neighbors"[5] also tend to be large/small. All parent-child pairs in the same position across different channels are assigned into one group, and the problem becomes overlapping group sparsity regularization. Similar scheme has been used in

[5]Parent denotes the parent node on the same channel while neighbors mean coefficients at the same position on other channels.

approximating tree sparsity [49, 14], where each node and its parent are assigned into one group. We write the approximated problem as:

$$\min_x \frac{1}{2}||Ax - b||_2^2 + \lambda \sum_{g \in \mathcal{G}} ||(\Phi x)_g||_2 \tag{3.8}$$

where g denotes one of the coefficient groups discussed above (an example is demonstrated in Fig.1(c)), $(\cdot)_g$ denotes the coefficients in group g and $\mathcal{G}$ is the set of all groups.

The mixed $\ell_{2,1}$ norm encourages all the components in the same group g to be zeros or non-zeros simultaneously. With our group configuration, it encourages forest sparsity. We present an efficient implementation based on fast iterative shrinkage-thresholding algorithm (FISTA) [42] framework for this problem. This is because FISTA can be easily applied for standard sparsity and joint sparsity, which could make the validation of the benefit of the proposed model more convenient. In addition, the formulation (3.8) can be easily extended to the combination of total variation (TV) via the Fast Composite Splitting Algorithms (FCSA) scheme [4]. Note that other algorithms may be used to solve the forest sparsity problems, e.g. [16, 52, 95], but determining the optimal algorithm for forest sparsity is not the scope of this article.

FISTA [42] is a accelerated version of proximal method which minimizes the object function with the following form:

$$\min\{F(x) = f(x) + g(x)\} \tag{3.9}$$

where $f(x)$ is a convex smooth function with Lipschitz constant L_f and $g(x)$ is a convex but usually nonsmooth function. It comes to the original FISTA when $f(x) = \frac{1}{2}||Ax - b||_2^2$ and $g(x) = \lambda\|\Phi x\|_1$, which is summarized in Algorithm 3, where, A^T denotes the transpose of A.

For the second step, there is closed form solution by soft-thresholding. For joint sparsity problem where $g(x) = \lambda\|\Phi x\|_{2,1}$, the second step also has closed form

Algorithm 3 FISTA [42]

Input: $\rho = 1/L_f$, λ, $n = 1$, $t^1 = 1$ $r^1 = x^0$
while not meet the stopping criterion **do**
  $y = r^n - \rho A^T(Ar^n - b)$
  $x = \arg\min_x\{\frac{1}{2\rho}\|x - y\|^2 + \lambda\|\Phi x\|_1\}$
  $t^{n+1} = 1 + \sqrt{1 + 4(t^n)^2}/2$
  $r^{n+1} = x^n + \frac{t^n - 1}{t^{n+1}}(x^n - x^{n-1})$
  $n = n + 1$
end while

solution. We call the version as FISTA_Joint for joint sparsity. However, for the problem (3.8) with overlapped groups, we can not directly apply FISTA to solve it.

In order to transfer the problem (3.8) to non-overlapping version, we introduce a binary matrix $G \in \mathbb{R}^{D \times TN}$ $(D > TN)$ to duplicate the overlapped coefficients. Each row of G only contains one 1 and all else are 0s. The 1 appears in the i-th column corresponds to the i-th coefficient of Φx. Intuitively, if the coefficient is included in j groups, G will contains j such rows. An auxiliary variable z is used to constrain $G\Phi x$. This scheme is widely utilized in the alternating direction method (ADM) [16]. The alternating formulation becomes:

$$\min_{x,z}\{\frac{1}{2}\|Ax - b\|_2^2 + \lambda\sum_{g\in\mathcal{G}}||z_g||_2 + \frac{\gamma}{2}||z - G\Phi x||_2^2\} \tag{3.10}$$

where γ is another positive parameter. We iteratively solve this alternative formulation by minimizing x and z subproblems respectively. For the z subproblem:

$$\hat{z_g} = \arg\min_{z_g}\{\lambda||z_g||_2 + \frac{\gamma}{2}||z_g - (G\Phi x)_g||_2^2\}, g \in \mathcal{G} \tag{3.11}$$

which has the closed form solution:

$$\hat{z_g} = \max(||(G\Phi x)_g||_2 - \frac{\lambda}{\gamma}, 0)\frac{(G\Phi x)_g}{||(G\Phi x)_g||_2}, g \in \mathcal{G} \tag{3.12}$$

We denote it as a shrinkgroup operation. For the x-subproblem:

$$\hat{x} = \arg\min_x \{\frac{1}{2}\|Ax - b\|_2^2 + \frac{\gamma}{2}||z - G\Phi x||_2^2\} \tag{3.13}$$

The optimal solution is $x = (A^TA + \lambda\Phi^TG^TG\Phi)^{-1}(A^Tb + \lambda\Phi^TG^Tz)$, which contains a large scale inverse problem. Actually, this problem can be efficient solved by various methods. In order to compare with FISTA and FISTA_Joint, we apply FISTA to solve (3.13). This could demonstrate the benefit of forest sparsity more clearly. Let $f(x) = \frac{1}{2}\|Ax - b\|_2^2 + \frac{\lambda}{2}||z - G\Phi x||_2^2$ and $g(x) = 0$. Supposing its Lipschitz constant to be L_f, the whole algorithm is summarized in Algorithm 4.

Algorithm 4 FISTA_Forest

Input: $\rho = 1/L_f$, $r^1 = x^0$, $t^1 = 1, \lambda, \gamma, n = 1$
while not meet the stopping criterion **do**
  $z = shrinkgroup(G\Phi x^{n-1}, \lambda/\gamma)$
  $x^n = r^n - \rho[A^T(Ar^n - b) + \gamma\Phi^TG^T(G\Phi r^n - z)]$
  $t^{n+1} = [1 + \sqrt{1 + 4(t^n)^2}]/2$
  $r^{n+1} = x^n + \frac{t^n - 1}{t^{n+1}}(x^n - x^{n-1})$
  $n = n + 1$
end while

For the first step, we solve (3.11) while $\frac{1}{2}\|Ax - b\|_2^2$ keeps the same. The object function value in (3.10) decreases. For the second step, (3.13) is solved by FISTA iteratively while $\lambda\sum_{g\in\mathcal{G}}||z_g||_2$ keeps the same. Therefore, the object function value in (3.10) decreases in each iteration and the algorithm is convergent. Algorithm 4 is also used to implement tree sparsity by recovering the data channel-by-channel separately. We call it FISTA_Tree.

In some practical applications, the data tends to be forest-sparse but not strictly. We can soften and complement the forest assumption with other penalties, such as joint $\ell_{2,1}$ norm or TV. For example, after combining TV, problem (3.10) becomes:

$$\min_{x,z}\{\frac{1}{2}\|Ax-b\|_2^2+\lambda\sum_{g\in\mathcal{G}}||z_g||_2+\frac{\gamma}{2}||z-G\Phi x||_2^2 \\ +\mu||x||_{TV}\} \tag{3.14}$$

where $||x||_{TV}=\sum_{i=1}^{TN}\sqrt{(\nabla_1 x_i)^2+(\nabla_2 x_i)^2}$; ∇_1 and ∇_2 denote the forward finite difference operators on the first and second coordinates respectively; μ is a positive parameter. Compared with Algorithm 4, we only need to set $g(x)=\mu||x||_{TV}$ and the corresponding subproblem has already been solved [42, 4, 96]. This TV combined algorithm is called FCSA_Forest, which will be used in the experiments. To avoid repetition, it is not listed.

3.5 Applications and Experiments

We conduct experiments on the RGB color image, multi-contrast MR images, MR image of multi-channel coils and the multispectral image to validate the benefit of forest sparsity. All experiments are conducted on a desktop with 3.4GHz Intel core i7 3770 CPU. Matlab version is 7.8(2009a). If the sampling matrix A is M by N, the sampling ratio is defined as M/N. All measurements are mixed with Gaussian white noise of 0.01 standard deviation. Signal-to-Noise Ratio (SNR) is used as the metric for evaluations:

$$SNR=10\log_{10}(V_s/V_n) \tag{3.15}$$

where V_n is the Mean Square Error between the original data x_0 and the reconstructed x; $V_s=var(x_0)$ denotes the power level of the original data where $var(x_0)$ denotes the variance of the values in x_0.

3.5.1 Multi-contrast MRI

Multi-contrast MRI is a popular technique to aid clinical diagnosis. For example T1 weighted MR images could distinguish fat from water, with water appearing darker and fat brighter. In T2 weighted images fat is darker and water is lighter, which is better suited to imaging edema. Although with different intensities, T1/T2 or proton-density weighted MR images are scanned at the same anatomical position. Therefore they are not independent but highly correlated. Multi-contrast MR images are typically forest-sparse under the wavelet basis. Suppose $\{x_t\}_{t=1}^T \in \mathbb{R}^N$ are the multi-contrast images for the same anatomical cross section and $\{b_t\}_{t=1}^T$ are the corresponding undersampled data in Fourier domain, the forest-sparse reconstruction can be formulated as:

$$\hat{x} = \arg\min_x \|\Phi x\|_{\mathcal{F},T} + \lambda \sum_{s=1}^{T} \|R_t x_t - b_t\|^2 \tag{3.16}$$

where x is the vertorized data of $[x_1, ..., x_T]$ and R_t is the measurement matrix for the image x_t. This is an extension of conventional CS-MRI [1]. Fig. 3.1 shows an example of the forest structure in multi-contrast MR images.

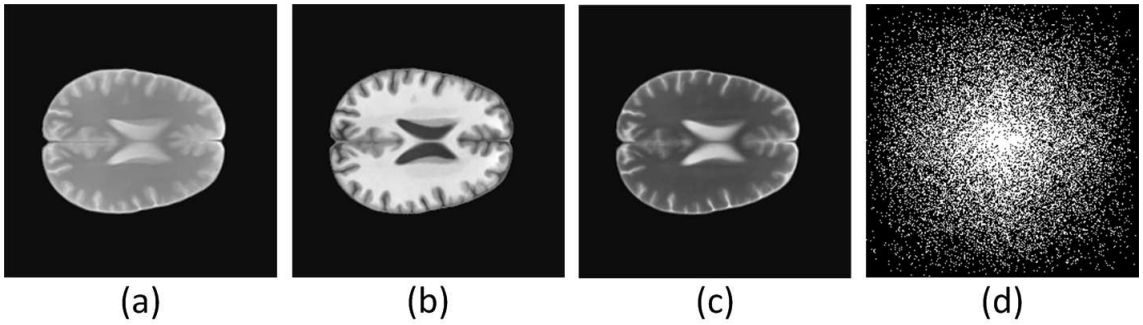


Figure 3.2. (a)-(c): the original multi-constrast images. (d): the sampling mask.

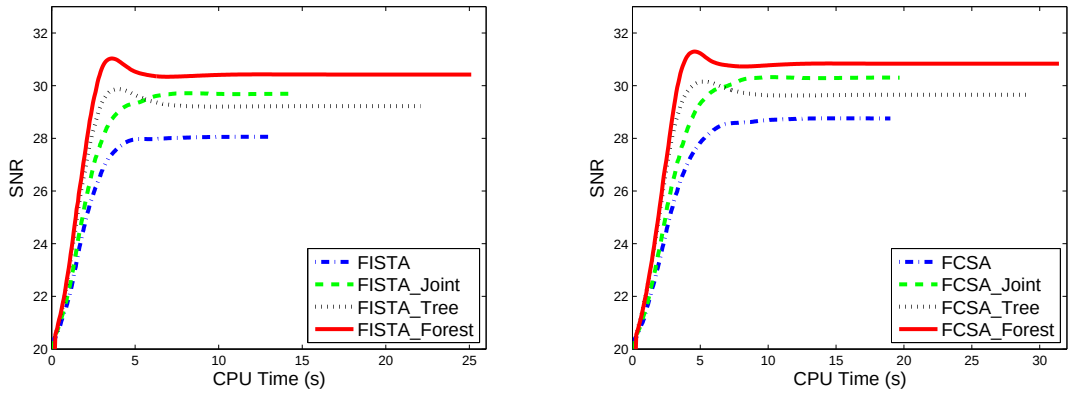
The data is extracted from the SRI24 Multi-Channel Brain Atlas Dataset [97]. In the Fourier domain, we randomly obtain more samples in low frequencies and less

samples in higher frequencies. This sampling scheme has been widely used for CS-MRI [1, 2, 4]. Fig. 3.2 shows the original multi-contrast MR images and the sampling mask.

We compare four algorithms on this dataset: FISTA, FISTA_Joint, FISTA_Tree and FISTA_Forest. The parameter λ is set 0.035, and γ is set to 0.5λ. We run each algorithm 400 iterations. Fig. 3.3 (a) demonstrates the performance comparisons among different algorithms. From the figure, we could observe that modeling with forest sparsity achieves the highest SNR after convergence. Although the algorithm for forest sparsity takes more time due to the overlapping structure, it always outperforms all others in terms of accuracy.

In addition, as total variation is very popular in CS-MRI [1, 4, 78], we compare our FCSA_Forest algorithm with FCSA [4] (TV is combined in FISTA), FCSA_Joint [78] (TV is combined in FISTA_Joint) and FCSA_Tree. The parameter μ for TV is set 0.001, the same as that in previous works [2, 4]. Fig. 3.3 (b) demonstrates the performance comparison including TV regularization. Compared with Fig. 3.3 (a), all algorithms improve at different degrees. However, the ranking does not change, which validates the superiority of forest sparsity. As FCSA has been proved to be better than other algorithms for general compressed sensing MRI (CS-MRI) [1, 2, 3] and FCSA_Joint [78] better than [77, 98] in multi-contrast MRI, the proposed method further improves CS-MRI and make it more feasible than before.

In order to validate the benefit of forest sparsity in terms of measurement number, we conduct an experiment to reconstruct multi-contrast MR images from different sampling ratios. Fig. 3.4 demonstrates the final results of four algorithms with sampling ratio from 16% to 26%. With more sampling, all algorithms have better performance. However, The forest sparsity algorithm always achieves the best reconstruction. For the same reconstruction accuracy, the FISTA_Forest algorithm only



(a) (b)

Figure 3.3. Performance comparisons among different algorithms. (a) Multi-constrast MR images reconstruction with 20% sampling. Their final SNRs are 28.05, 29.69, 29.22 and 30.42 respectively. The time costs are 13.11s, 14.43s, 22.08s and 25.11s respectively. (b) Multi-constrast MR images reconstruction with 20% sampling by both wavelet sparsity and TV regularization. Their final SNRs are 28.75, 30.30, 29.65 and 30.83 respectively. The time costs are 19.00 s, 19.68 s, 29.11 s and 31.41 s, respectively.

requires about 16% measurements to achieve SNR 28, which is approximately 2%, 3%, 5% less than that of FISTA_Joint, FISTA_Tree and FISTA respectively. More results of forest sparsity on multi-contrast MRI can be found in [99].

3.5.2 parallel MRI

To improve the scanning speed of MRI, an efficient and feasible way is to acquire the data in parallel with multi-channel coils. The scanning time depends on the number of measurements in Fourier domain, and it will be significantly reduced when each coil only acquires a small fraction of the whole measurements. The bottleneck is how to reconstruct the original MR image efficiently and precisely. This issue is called pMRI in literature. Sparsity techniques have been used to improve the classical method SENSE [32]. However, when the coil sensitivity can not be estimated

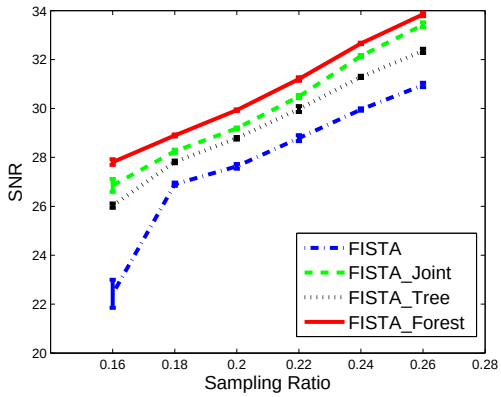
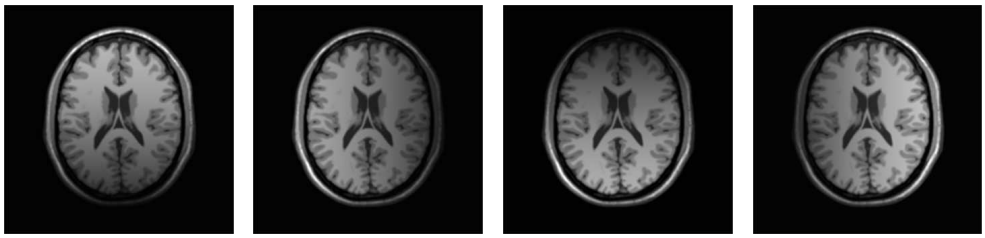


Figure 3.4. Reconstruction performance with different sampling ratios.

precisely, the final image would contain visual artifacts. Unlike previous CS-SENSE [38] which reconstructs the images of multi-coils individually, calibrationless parallel MRI [100, 101] recovers the aliased images of all coils jointly by assuming the data is jointly sparse.

Let T equal to the number of coils and b_t be the measurement vector from coil t. It is therefore the same CS problem as (3.16). The final result of CaLM-MRI is obtained by a sum of square (SoS) approach without coil sensitivity and SENSE encoding. It shows comparable results with those methods which need precise coil configuration. As shown in Fig. 3.5, the appearances of different images obtained from multi-coils are very similar. This method can be improved with forest sparsity, since the images follow the forest sparsity assumption.

There are two steps for compressed sensing pMRI reconstruction in CaLM-MRI [100]: 1) the aliased images are recovered from the undersampled Fourier signals of different coil channels by CS methods; 2) The final image for clinical diagnosis is synthesized by the recovered aliased images using the sum-of-square (SoS) approach. As discussed above, these aliased images should be forest-sparse under the wavelet basis. We compare our algorithm with FISTA_Joint and SPGL1 [40] which solves the



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

Figure 3.5. The aliased MR images of multi-coils. Due to the different locations of the coils, they have different sensitivities to the same image.

joint $\ell_{2,1}$ norm problem in CaLM-MRI. For the second step, all methods use the SoS approach from the aliased images that they recovered. All algorithms run enough time until it has converged.

Table 3.2. Comparisons of SNRs (dB) on different sampling ratios with 4 coils

	sampling ratios	25%	20%	17%	15%
SNR of Aliased Images	SPGL1	26.72	24.59	23.08	22.31
	FISTA_Joint	26.95	24.73	23.06	22.21
	FISTA_Forest	**27.47**	**25.22**	**23.37**	**22.59**
SNR of Final Image	SPGL1	20.64	20.35	19.12	18.64
	FISTA_Joint	20.79	20.41	19.75	18.49
	FISTA_Forest	**22.62**	**22.29**	**21.03**	**20.47**

Table 3.3. Comparisons of SNRs (dB) on different number of coils with 20% sampling ratio

	number of coils	2	4	6	8
SNR of Aliased Images	SPGL1	23.33	24.61	24.74	25.16
	FISTA_Joint	23.41	24.71	24.89	25.23
	FISTA_Forest	**24.25**	**25.12**	**25.29**	**25.52**
SNR of Final Image	SPGL1	21.76	18.95	21.05	21.32
	FISTA_Joint	21.90	18.94	21.15	21.87
	FISTA_Forest	**22.44**	**22.22**	**22.52**	**22.52**

Table 3.2 and Table 3.3 show all the comprehensive comparisons among these algorithms. For the same algorithm, more measurements or more number of coils tend to increase the SNRs of aliased images, although it does not result in linear improvement for the final image reconstruction. Another observation is that FISTA_Joint and SPGL1 have similar performance in terms of SNR on this data. This is because both of them solve the same joint sparsity problem, even with different schemes. Upgrading the model to forest sparsity, significant improvement can be gained. Finally, it is unknown how to combine TV in SPGL1. However, both FISTA_Joint and FISTA_Forest can easily combine TV, which can further enhance the results [78].

3.5.3 Color Image Reconstruction

Color images captured by optical camera can be represented as combinations of red, green, blue three colors. Different colors synthesized by these three colors seems realistic to human eyes. By observing the color channels are highly correlated, joint sparsity prior is utilized in recent recovery [75]. Modeling with $\ell_{2,1}$ norm regularization can gain additional SNR to standard ℓ_1 norm regularization. Further more, each color channel tends to be wavelet tree-sparse. If we model the problem with forest sparsity, this result would be reasonably better.

For color images, we compare our algorithm with FISTA, FISTA_Joint and FISTA_Tree. Fig. 3.6 shows the visual results recovered by different sparse penalties. Only after 50 iterations, the image recovered by our algorithm is very close to the original one with the fewest artifacts (shown in the zoomed region of interest).

3.5.4 Multispectral Image Reconstruction

Different from common color images, a multispectral or hyperspectral image is consisted of much more bands, which provides both spatial and spectral repre-

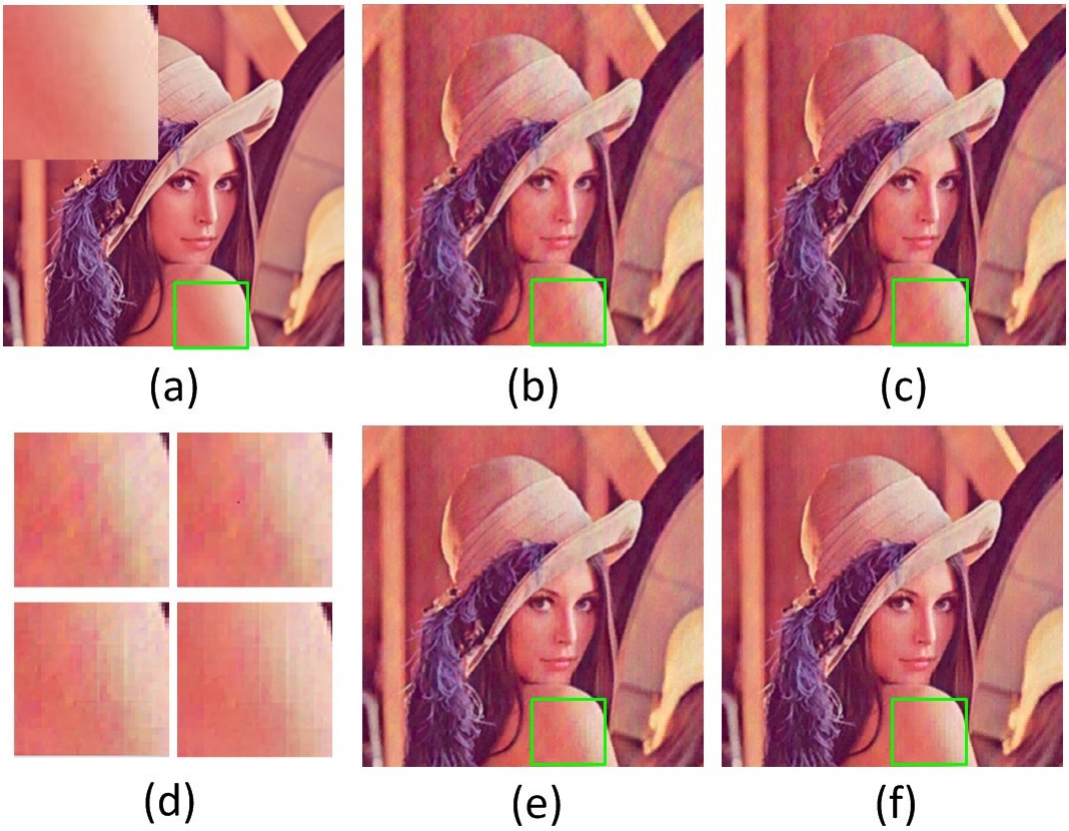


Figure 3.6. Visual comparisons on the lena image reconstruction after 50 iterations with about 20% sampling. (a) the original image and the patch detail; (b) recovered by FISTA; (c) recovered by FISTA_Joint; (d) the patch details for each recovered image; (e) recovered by FISTA_Tree; (f) recovered by FISTA_Forest. Their SNRs are 16.65, 17.41, 17.66 and 18.92, respectively.

sentations of scenes. It is widely utilized on remote sensing with applications to agriculture, environment detection etc. However, the collection of large amount of data costs both huge imaging time and storage space. By compressed sensing data acquisition, the cost of imaging for remote sensing data could be significantly reduced [102]. Like RGB images, the bands of multispectral image should represent the same scene. Each band has tree sparsity property. Therefore, they follow the forest sparsity

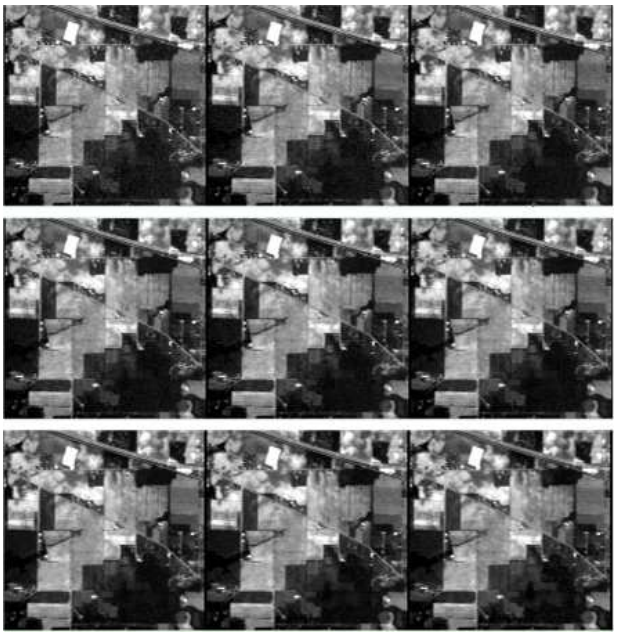


Figure 3.7. The original multispectral image: band 6 to band 14 .

assumption. Fig. 3.7 shows bands 6 to 14 of a multispectral image of 1992 AVIRIS Indian Pine Test Site 3 [6].

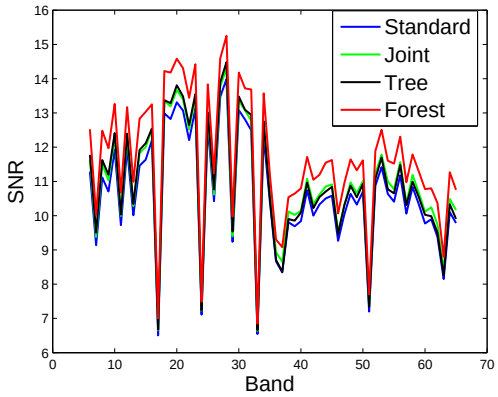


Figure 3.8. Multispectral image reconstruction results by different sparse models with about 20% sampling.

[6]The data is downloaded from `https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html`

For multispectral image, we test a dataset of 1992 AVIRIS image Indian Pine Test Site 3 (examples shown in Fig. 3.7). It is a 2×2 mile portion of Northwest Tippecanoe County of Indiana. There are total 220 bands. Each band is recovered separately for standard sparsity and tree sparsity, while every 3 bands are reconstructed simultaneously by joint-sparse model and forest-sparse model. Each image is cropped to 128×128 for convenience. The number of wavelet decomposition levels is set to 3. The SNRs of all recovered images for band 6 to 66 are shown in Fig. 3.8. One could observe that modeling with forest sparsity always achieves the highest SNRs, which validates the benefit of forest sparsity.

3.6 Summery

In this chapter, we have proposed a novel model *forest sparsity* for sparse learning and compressed sensing. This model enriches the family of structured sparsity and can be widely applied on numerous fields of sparse regularization problems. The benefit of the proposed model has been theoretically proved and empirically validated in practical applications. Under compressed sensing assumptions, significant reduction of measurements is achieved with forest sparsity compared with standard sparsity, joint sparsity or independent tree sparsity. A fast algorithm is developed to efficiently solve the forest sparsity problem. While applying it on practical applications such as multi-contrast MRI, pMRI, multispectral image and color image reconstruction, extensive experiments demonstrate the superiority of forest sparsity over standard sparsity, joint sparsity and tree sparsity in terms of both accuracy and computational complexity.

3.7 Appendix: Proofs of Theorems

3.7.1 Proof of Theorem 2

The proof is conducted on the binary tree case for convenience. The bound for quadtree can be easily extended.

First, we need to figure out the number of subtrees (size k) of a binary tree (size N). Note that the root of the subtrees should be the binary tree's root.

Case 1: when $k \leq \lfloor \log_2 N \rfloor$, the number of subtrees of size k is just the Catalan number:

$$L_{\mathcal{T}} = \frac{1}{k+1}\binom{2k}{k} \leq \frac{(2e)^k}{k+1} \leq \frac{e^k N}{k+1}. \tag{3.17}$$

Case 2: when $k > \lfloor \log_2 N \rfloor$, the number of subtrees of size k should follow [29]:

$$\begin{aligned} L_{\mathcal{T}} &\leq \frac{4^k}{k}\left(\frac{6}{\sqrt{\pi k}} \ln \frac{\log_2 N}{\lfloor \log_2 k \rfloor}) + \frac{128}{e^2 \lfloor \log_2 k \rfloor}\right) \\ &\leq \frac{4^k}{k}\left(\frac{c_1 \log_2 N}{\lfloor \log_2 k \rfloor}) + \frac{c2}{\lfloor \log_2 k \rfloor}\right) \\ &\leq \frac{4^k}{k}\frac{c_1 \log_2(c_3 N)}{\log_2 k} \\ &\leq \frac{4^k (c_4 N)}{k}. \end{aligned} \tag{3.18}$$

where c_1, c_2, c_3, c_4 are some constants. Therefore we have:

$$L_{\mathcal{T}} \leq \begin{cases} \frac{e^k N}{k+1} & \text{if } k \leq \lfloor \log_2 N \rfloor, \\ \frac{4^k (c_4 N)}{k} & \text{if } k > \lfloor \log_2 N \rfloor. \end{cases} \tag{3.19}$$

According to Theorem 1:

$$M \geq \frac{2}{c\delta}(\ln(2L) + k \ln \frac{12}{\delta} + t). \tag{3.20}$$

With (3.20), the number of measurements should satisfy:

$$M \geq \begin{cases} \frac{2}{c\delta_{\mathcal{T}_k}}(k + \ln(N/(k+1)) + k\ln(12/\delta_{\mathcal{T}_k}) \\ \quad + \ln 2 + t) \qquad \text{if } k \leq \lfloor \log_2 N \rfloor, \\ \frac{2}{c\delta_{\mathcal{T}_k}}(k\ln 4 + \ln(c_4 N/k) + k\ln(12/\delta_{\mathcal{T}_k}) \\ \quad + \ln 2 + t) \qquad \text{if } k > \lfloor \log_2 N \rfloor. \end{cases} \tag{3.21}$$

For both cases, we have $M = \mathcal{O}(k + \log(N/k))$ as the minimum number of measurements. Similar bound also has been proved in previous papers [29] [30].

3.7.2 Proof of Theorem 3

If the data is forest-sparse, the support set of different trees are dependent. It means if the support set for one tree is fixed, then all support sets for other trees are fixed. Accordingly, the number of combinations is still $L_{\mathcal{T}}$. Note that the sparsity number is Tk as there are T trees. Therefore,

$$TM \geq \begin{cases} \frac{2}{c\delta_{\mathcal{F}_{T,k}}}(k + \ln(N/(k+1)) + Tk\ln(12/\delta_{\mathcal{F}_{T,k}}) \\ \quad + \ln 2 + t) \qquad \text{if } k \leq \lfloor \log_2 N \rfloor, \\ \frac{2}{c\delta_{\mathcal{F}_{T,k}}}(k\ln 4 + \ln((c_4 N)/k) + Tk\ln(12/\delta_{\mathcal{F}_{T,k}}) \\ \quad + \ln 2 + t) \qquad \text{if } k > \lfloor \log_2 N \rfloor. \end{cases} \tag{3.22}$$

For both cases, the bound is reduced to $TM = \mathcal{O}(Tk + \log(N/k))$.

3.7.3 Proof of Theorem 4

We first derive the sufficient condition that guarantees the RIP for block-diagonal matrices.

Theorem 5. *Let a matrix $A \in \mathbb{R}^{TM\times TN}$ be composed by sub-Gaussian random matrices $\{A'_t \in \mathbb{R}^{M\times N}\}_{t=1}^{T}$ as in (3.6). For any fixed subset $S \subset \{1, 2, ..., TN\}$ with $|S| = TK$ and $0 < \delta < 1$, we have with probability exceeding $1-2(12/\delta)^{TK}e^{-c_1\frac{M}{2}\min(c_2^2\delta^2\Gamma_2, c_2\delta\Gamma_\infty)}$:*

$$(1-\delta)||X||_2 \leq ||A_S X||_2 \leq (1+\delta)||X||_2, \tag{3.23}$$

for all $X = [x_1; x_2; ...; x_T] \in \mathbb{R}^{Tk\times 1}$. c_1 and c_2 are absolute constants, $\Gamma_2 = \frac{(\sum_{t=1}^{T}||x_t||_2^2)^2}{\sum_{t=1}^{T}||x_t||_2^4}$ and $\Gamma_\infty = \frac{\sum_{t=1}^{T}||x_t||_2^2}{\max_{t=1}^{T}||x_t||_2^2}$.

Proof. Let's denote $\bar{X} = X/||X||_2$ and we have $||\bar{X}||_2 = 1$. We choose a finite set of points $Q = \{q_i\}$, such that $q_i \in \mathbb{R}^{Tk\times 1}$ and $||q_i||_2 = 1$ for all i. We have $\min_i ||\bar{X} - q_i||^2 \leq \epsilon_1$ and covering number satisfies $|Q| \leq (1 + 2/\epsilon_1)^{TK}$ for any $\epsilon_1 > 0$ (see Chap 13 of [103]).

As the block-diagonal matrix A is composed by sub-Gaussian random matrices, we have for each i and any $\epsilon_2 > 0$:

$$\begin{aligned}&\mathrm{P}(|||Aq_i||_2^2 - ||q_i||_2^2| \geq \epsilon_2||q_i||_2^2) \\ &\qquad\qquad \leq 2e^{-c_1\frac{M}{2}\min(c_2^2\epsilon_2^2\Gamma_2, c_2\epsilon_2\Gamma_\infty)},\end{aligned} \tag{3.24}$$

with Γ_2 and Γ_∞ defined above. This probability is indicated in Theorem III.1 of [104].

Taking union bound, we obtain with probability exceeding $1 - 2(1 + 2/\epsilon_1)^{TK}e^{-c_1\frac{M}{2}\min(c_2^2\epsilon_2^2\Gamma_2, c_2\epsilon_2\Gamma_\infty)}$:

$$(1-\epsilon_2) \leq ||A_S q_i||_2^2 \leq (1+\epsilon_2), \text{ for all } q_i \in Q, \tag{3.25}$$

which gives

$$(1-\epsilon_2) \leq ||A_S q_i||_2 \leq (1+\epsilon_2), \text{ for all } q_i \in Q. \tag{3.26}$$

Now we define ρ as the smallest nonnegative number such that

$$||A_S\bar{X}||_2 \leq (1+\rho), \tag{3.27}$$

for all $\bar{X} \in \mathbb{R}^{Tk\times 1}$ and $||\bar{X}||_2 = 1$. We have

$$\begin{aligned} ||A_S\bar{X}||_2 &\leq ||A_S q_i||_2 + ||A_s(\bar{X} - q_i)||_2 \\ &\leq ||A_S q_i||_2 + ||A_s(\bar{X} - q_i)||_2 \\ &\leq (1+\epsilon_2) + (1+\rho)\epsilon_1. \end{aligned} \tag{3.28}$$

As ρ as the smallest nonnegative number for (3.27), we have:

$$1 + \rho \leq (1+\epsilon_2) + (1+\rho)\epsilon_1, \tag{3.29}$$

and

$$\rho \leq (\epsilon_1 + \epsilon_2)/(1-\epsilon_1). \tag{3.30}$$

Note the above result holds for any ϵ_1 and ϵ_2. We choose $\epsilon_1 = \delta/4$ and $\epsilon_2 = \delta/2$. Since $0 < \delta < 1$, it is easy to see that $\rho \leq \delta$, which proves

$$||A_S\bar{X}||_2 \leq (1+\delta). \tag{3.31}$$

Similar, $(1-\delta) \leq ||A_S\bar{X}||_2$ can be proved using the same way. Finally, we obtain with probability exceeding $1 - 2(12/\delta)^{TK}e^{-c_1\frac{M}{2}\min(c_2^2\delta^2\Gamma_2, c_2\delta\Gamma_\infty)}$:

$$(1-\delta) \leq \frac{||A_S X||_2}{||X||_2} \leq (1+\delta), \tag{3.32}$$

which completes the proof as $1 + 2/\epsilon_1 = (\delta+8)/\delta \leq 12/\delta$.

Based on this theorem, we know that any Tk-sparse $X \in \mathbb{R}^{TN\times 1}$ satisfies

$$(1-\delta)||X||_2 \leq ||AX||_2 \leq (1+\delta)||X||_2, \tag{3.33}$$

with probability exceeding $1 - 2(12/\delta)^{TK}e^{-c_1\frac{M}{2}\min(c_2^2\delta^2\Gamma_2, c_2\delta\Gamma_\infty)}$.

Suppose there are L combinations of such set S, from appendix A and B we know that

$$L \leq \begin{cases} \frac{e^k N}{k+1} & \text{if } k \leq \lfloor \log_2 N \rfloor, \\ \frac{4^k(c_4 N)}{k} & \text{if } k > \lfloor \log_2 N \rfloor. \end{cases} \tag{3.34}$$

for forest sparse data.

By taking the union bound, we known that (3.23) fails with probability less than

$$2L(12/\delta_{\mathcal{F}_{T,k}})^{TK} e^{-c_1 \frac{M}{2} \min(c_2^2 \delta^2_{\mathcal{F}_{T,k}} \Gamma_2, c_2 \delta_{\mathcal{F}_{T,k}} \Gamma_\infty)} \leq e^{-t}, \tag{3.35}$$

which gives

$$TM \geq \begin{cases} \frac{2T(\ln 2 + k + \ln(N/(k+1)) + Tk \ln(12/\delta_{\mathcal{F}_{T,k}}) + t)}{c_1 \min(c_2^2 \delta^2_{\mathcal{F}_{T,k}} \Gamma_2, c_2 \delta_{\mathcal{F}_{T,k}} \Gamma_\infty)}, \\ \text{if } k \leq \lfloor \log_2 N \rfloor, \\ \frac{2T(\ln 2 + k \ln 4 + \ln((c_3 N)/k) + Tk \ln(12/\delta_{\mathcal{F}_{T,k}}) + t)}{c_1 \min(c_2^2 \delta^2_{\mathcal{F}_{T,k}} \Gamma_2, c_2 \delta_{\mathcal{F}_{T,k}} \Gamma_\infty)}, \\ \text{if } k > \lfloor \log_2 N \rfloor. \end{cases} \tag{3.36}$$

From this one theorem 4 can be easily derived. For both cases, the bound can be written as $TM = \mathcal{O}(\frac{T^2 k + T \log(N/k)}{\min(\Gamma_2, \Gamma_\infty)})$.

CHAPTER 4

Dynamic Gradient Sparsity in Remote Sensing Image Fusion

In this chapter, we introduce a new sparsity induce term called Dynamic Gradient Sparsity. We also show how this method can be used to image fusion in remote sensing. Finally, this method is evaluated via extensive experiments with comparisons to the existing approaches. This work was presented under a slightly modification from [76].

4.1 Introduction

Multispectral (MS) images are widely used in many fields of remote sensing such as environmental monitoring, agriculture, mineral exploration etc. However, the design of MS sensors with high resolution is confined by infrastructure limits in onboard storage and bandwidth transmission [105]. In contrast, panchromatic (Pan) gray-scaled images with high spatial resolution can be obtained more conveniently because they are composed of much reduced numbers of pixels. The combinations of Pan images with high spatial resolution and MS images with high spectral resolution can be acquired simultaneously from most existing satellites. Therefore, we expect to obtain images in both high spatial resolution and high spectral resolution via image fusion (also called pan-sharpening). A fusion example on Quickbird satellite images is shown in Figure 4.1.

Image fusion is a typical inverse problem and generally difficult to solve. A number of conventional methods use projection and substitution, which include principal component analysis (PCA) [106], intensity hue saturation (IHS) [107], wavelet [108]

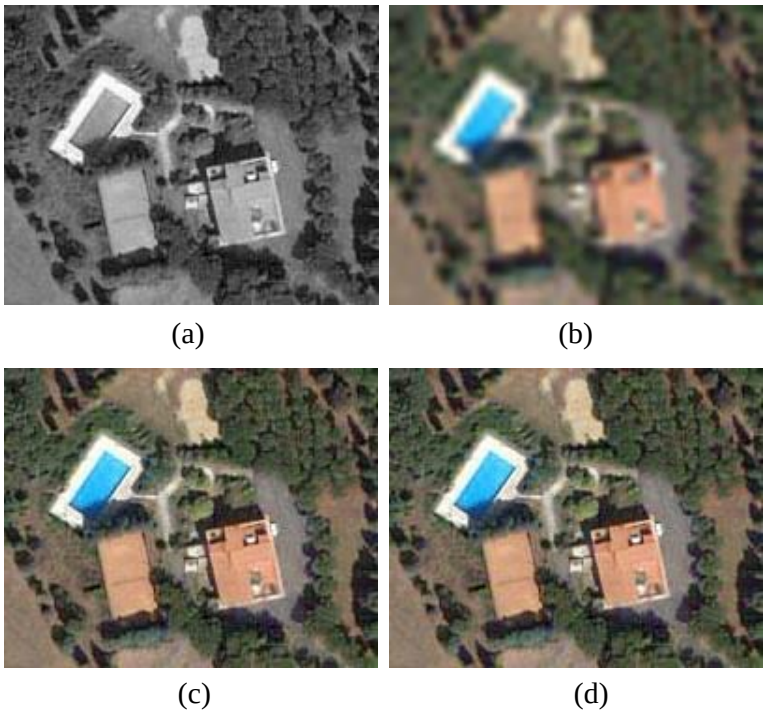


Figure 4.1. (a) A high resolution panchromatic image. (b) The corresponding low resolution multi-spectral image. (c) Our fusion result. (d) The ground-truth. Copyright DigitalGlobe.

and their combinations. These methods work in the following scheme: upsampling, forward transform, intensity matching, component substitution and reverse transform [109]. Other methods such as Brovey [110], assume the Pan image is a linear combination of all bands of the fused image. A detailed survey of existing methods can be found in [109]. While these previous methods provided some good visual results, they are very likely to suffer from spectral distortion since their strong assumptions are not realistic in remote sensing physics [105].

In order to overcome the issue caused by spectral distortion, a suite of variational approaches have emerged recently [111][112][113]. Each method formulates an energy function based on somewhat weak assumptions, and minimizing such a function leads to the optimum. The first variational method P+XS [111] is based on

the linear combination assumption in Brovey [110] and also assumes the upsampled MS image is the fusion result after blurring. As an accurate blur kernel is difficult to pre-estimate, AVWP [112] replaces this term with a spectral ratio constraint to preserve spectral information. It also forces the fused image to be close to the wavelet fused image [108]. Another variational model is engaged in estimating the fused image and the blurring model parameters iteratively [113]. Promising results have been achieved in these variational methods, especially they can reduce spectral distortion. However, due to the lack of an effective model to preserve spatial information, visible artifacts may appear on the fusion results.

In this chapter, we propose a new variational model for image fusion to bridge this gap. Motivated by the geographical relationship between the fused image and Pan image, a dynamic gradient sparsity property is discovered, defined and then exploited to improve spatial quality. In addition, we assume the fused image after downsampling should be close the MS image, which is formulated as a least squares fitting term to keep spectral information. The combined model does not violate remote sensing physics. This is a key difference compared with previous methods. Moreover, our method incorporates the inherent correlation of different bands, which has not been considered before. To optimize our entire energy function, a new algorithm is proposed in the fast iterative shrinkage-thresholding algorithm (FISTA) [42] framework, with a very fast convergence rate. Extensive experimental results demonstrate our method can significantly reduce spectral distortion while preserving sharp objects boundaries in the fused images.

4.2 Notations and Related Work

Scalers are denoted by lowercase letters. Bold letters denote matrices. Specially, $\mathbf{P} \in \mathbb{R}^{m \times n}$ denotes the Pan image and $\mathbf{M} \in \mathbb{R}^{\frac{m}{c} \times \frac{n}{c} \times s}$ denotes the low resolution MS

image. c is a constant. For example $c = 4$ when the resolution of Pan image is 0.6m and that of MS image is 2.4m in Quickbird acquisition. Let $\bar{\mathbf{M}} \in \mathbb{R}^{m \times n \times s}$ denote the upsampled MS image and the image to be fused is denoted by $\mathbf{X} \in \mathbb{R}^{m \times n \times s}$. Operator $\cdot$ denotes the element-wise product and $\cdot/$ denotes the element-wise division. Operator $*$ denotes convolution. $||\cdot||_F$ denotes the Frobenius norm. For simpleness, $\mathbf{X}_{i,j,d}$ denotes the element in i-th row, j-th column and d-th band in $\mathbf{X}$. And $\mathbf{X}_d$ denotes the whole d-th band, which is therefore a matrix.

4.2.1 P+XS

Ballester et al. proposed the first variational P+XS for pan-sharpening, where P and XS stand for Pan and MS images, respectively [111]. They formulate the pan-sharpening as an optimization problem:

$$\min_{\mathbf{X}} \sum_{d=1}^{s} ||\theta^{\perp} \cdot \nabla \mathbf{X}_d||_F^2 + \lambda || \sum_{d=1}^{s} \alpha_d \mathbf{X}_d - \mathbf{P}||_F^2 + \mu \sum_{d=1}^{s} ||\Pi_s (k_d * \mathbf{X}_d - \bar{\mathbf{M}}_d)||_F^2, \quad (4.1)$$

where μ and λ are two positive parameters. Π_s is a Dirac comb which selects multispectral pixel values. k_d denotes a blurring kernel for band d. α_d denotes a weight for band d and $\sum_{d=1}^{s} \alpha_d = 1$. $\theta = \nabla \mathbf{P} \cdot / |\nabla \mathbf{P}|$ represents the gradient direction of $\mathbf{P}$ and $\theta^{\perp}$ denotes its rotation by $\pi/2$.

It is based on three assumptions: 1) in the ideal case, the gradients of the fused images should be in the same directions as those of the Pan image, which is equivalent to $\theta^{\perp} \cdot \nabla \mathbf{X}_d = \mathbf{0}$; 2) the Pan image is assumed to be a linear combination of the bands of $\mathbf{X}$; 3) the upsampled MS image can be viewed as the fused image $\mathbf{X}$ after blurring. The classical gradient descent method is used to solve this problem.

4.2.2 AVWP

Derived from P+XS, a wavelet-based variational pan-sharpening called AVWP is proposed [112]. To reduce blurriness, AVWP constrains the fused image to be close to the wavelet fused image. In addition to preserve spectral correlation, it assumes the ratios between two spectral bands should be consistent for fused image $\mathbf{X}$ and the MS image, that is, $\mathbf{X}_d \cdot /\mathbf{X}_r = \bar{\mathbf{M}}_d \cdot /\bar{\mathbf{M}}_r$ for every pixel. The problem is to solve:

$$\min_{\mathbf{X}} \sum_{d=1}^{s} ||\theta^{\perp} \cdot \nabla \mathbf{X}_d||_F^2 + \lambda \sum_{d=1}^{s} ||\mathbf{X}_d - \mathbf{Z}_d||_F^2 + \mu \sum_{d,r=1,d<r}^{s} ||\mathbf{X}_d \cdot \bar{\mathbf{M}}_r - \mathbf{X}_r \cdot \bar{\mathbf{M}}_d||_F^2, \quad (4.2)$$

where $\mathbf{Z} \in \mathbb{R}^{m \times n \times s}$ denotes a combination of the MS image and the wavelet fused image.

4.2.3 FVP

Very recently, Fang et al. proposed a new variational approach [113], which we call Fang's variational pan-sharpening (FVP). The problem can be formulated as:

$$\min_{\mathbf{X},\mathbf{a},\mathbf{b}} |\sum_{d=1}^{s} \alpha_d \mathbf{X}_d - \mathbf{P}|_{TV} + \tau \sum_{d=1}^{s} ||\mathbf{a}_d||_F^2 + \lambda \sum_{d=1}^{s} ||\mathbf{a}_d \cdot (uk * \mathbf{X}_d) + \mathbf{b}_d - \bar{\mathbf{M}}_d||_F^2 + \mu \sum_{d=1}^{s-1} \sum_{r=s+1}^{s} ||\mathbf{X}_d - \mathbf{X}_r - \bar{\mathbf{M}}_d + \bar{\mathbf{M}}_r||_F^2, \quad (4.3)$$

where the total variation (TV) is defined as $\|\mathbf{X}\|_{TV} = \sum_{i=1}^{m} \sum_{j=1}^{n} \sqrt{(\nabla_1 \mathbf{X}_{i,j})^2 + (\nabla_2 \mathbf{X}_{i,j})^2}$. ∇_1 and ∇_2 denote the forward finite difference operators on the first and second coordinates, respectively. $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{m \times n \times s}$ are two deblurring model parameters. uk is a unit constant kernel. Different from the pre-defined deblurring kernel in P+XS, the deblurring model is updated during each iteration. The third term is derived by the spectral gradient similarity assumption: $\nabla_d \mathbf{X} = \nabla_d \bar{\mathbf{M}}$, where ∇_d denotes the gradient in the spectrum direction. Both (4.2) and (4.3) are solved by the spilt Bregman method [114].

4.2.4 Analysis

Although these methods provide remarkable results for pan-sharpening over the conventional ones, a common drawback of these three methods is the use of upsampled MS image as prior information. However, $\bar{\mathbf{M}}$ is not accurate to be a good reference image (see Figure 4.1). Secondly, these methods place constraints on the estimated blurring kernel and wavelet fused image, both of which will result in errors in the final output. Finally, all these methods are based on band-by-band fusion. The intra-correlations among multiple spectral bands are neglected. In the following section, we will show that *all* of these shortcomings are overcome in our method.

4.3 Proposed Method

4.3.1 Local Spectral Consistency

As discussed above, we'd like to avoid considering an upsampled MS image as prior knowledge to preserve spectral information. Therefore, we only assume the fused image after downsampling is close to the original MS image. Least squares fitting is used to model this relationship:

$$E_1 = \frac{1}{2}||\psi\mathbf{X} - \mathbf{M}||_F^2, \tag{4.4}$$

where ψ denotes a downsampling operator. Local spectral information is forced to be consistent with each MS pixel. Similar as in previous works, the input images are assumed to be geometrically registered during preprocessing.

Minimizing E_1 would be a severely ill-posed problem, due to the very low undersampling rate (e.g. $1/16$ when $c = 4$). Without strong prior information, $\mathbf{X}$ is almost impossible to be estimated accurately. This may be the reason that all previous methods do not use this energy function.

4.3.2 Dynamic Gradient Sparsity

Fortunately, the Pan image provides such prior information. Due to the strong geographical correlation with the fused image $\mathbf{X}$, the Pan image has already provided us with clear boundary information of land objects. Many researchers attempt to build this relationship mathematically. From recent reviews [109][105], however, no model exists that can effectively characterize this relationship.

As remotely sensed images are often piece-wise smooth, their gradients tend to be sparse and the non-zeros corresponds to the boundaries. In addition, the positions of such boundaries should be the same as those on the Pan image. It demonstrates that the sparsity property is not fixed but dynamic according to a reference image. This property has not been studied in sparsity theories yet. We call the data with such a property a dynamic gradient sparse signal/image.

Definition: *Let* $x \in \mathbb{R}^N$ *and* $r \in \mathbb{R}^N$ *denote the signal and the reference signal.* Ω_x *and* Ω_r *denote the support sets*[1] *of their gradients, respectively. The set of dynamic gradient sparse signals is defined as:*

$\mathcal{S}_x = \{x \in \mathbb{R}^N : |\Omega_x| = K, \Omega_x = \Omega_r, \text{ with } K << N\}$.

Using similar logic, it can be extended to multi-channel/spectral signals and images. The first term in P+XS [111] and AVWP [112] does not induce sparseness and tends to over-smooth the image by penalizing large values. In FVP [113], the first term is derived from the linear combination assumption in P+XS; it does not promote sparsity for each band. Different from previous work, dynamic gradient sparsity is encouraged in our method. Beside the prior information that previous methods attempt to use, we also notice the intra- correlations across different bands as they are the representations of the same land objects. Therefore, the gradients of different

[1]Here we mean the indices of the non-zero components.

bands should be group sparse. It is widely known that the ℓ_1 norm encourages sparsity and the $\ell_{2,1}$ norm encourages group sparsity [81]. Thus we propose a new energy function to encourage dynamic gradient sparsity and group sparsity simultaneously:

$$E_2 = ||\nabla \mathbf{X} - \nabla D(\mathbf{P})||_{2,1} \tag{4.5}$$

$$= \sum_i \sum_j \sqrt{\sum_d \sum_q (\nabla_q \mathbf{X}_{i,j,d} - \nabla_q \mathbf{P}_{i,j})^2}, \tag{4.6}$$

where $q = 1, 2$ and $D(\mathbf{P})$ means duplicating $\mathbf{P}$ to s bands. Interestingly, when there is no reference image, i.e. $\mathbf{P} = \mathbf{0}$, the result is identical to that of vectorial total variation (VTV) [115], which is widely used in color image denoising/deblurring.

To demonstrate why E_2 encourages dynamic gradient sparsity, we show a simple example on a 1D multi-channel signal in Figure 4.2. We could observe that, if the solution has a different support set from the reference, the total sparsity of the gradients will be increased. Cases (a)-(d) have group sparsity number 8, 4, 4, 2 respectively. Therefore, (a)-(c) will be penalized because they are not the sparsest solution in our method.

Combining the two energy functions, the image fusion problem can be formulated as:

$$\min_{\mathbf{X}} \{E(\mathbf{X}) = E_1 + \lambda E_2$$

$$= \frac{1}{2}||\psi \mathbf{X} - \mathbf{M}||_F^2 + \lambda ||\nabla \mathbf{X} - \nabla D(\mathbf{P})||_{2,1}\}, \tag{4.7}$$

where λ is a positive parameter.

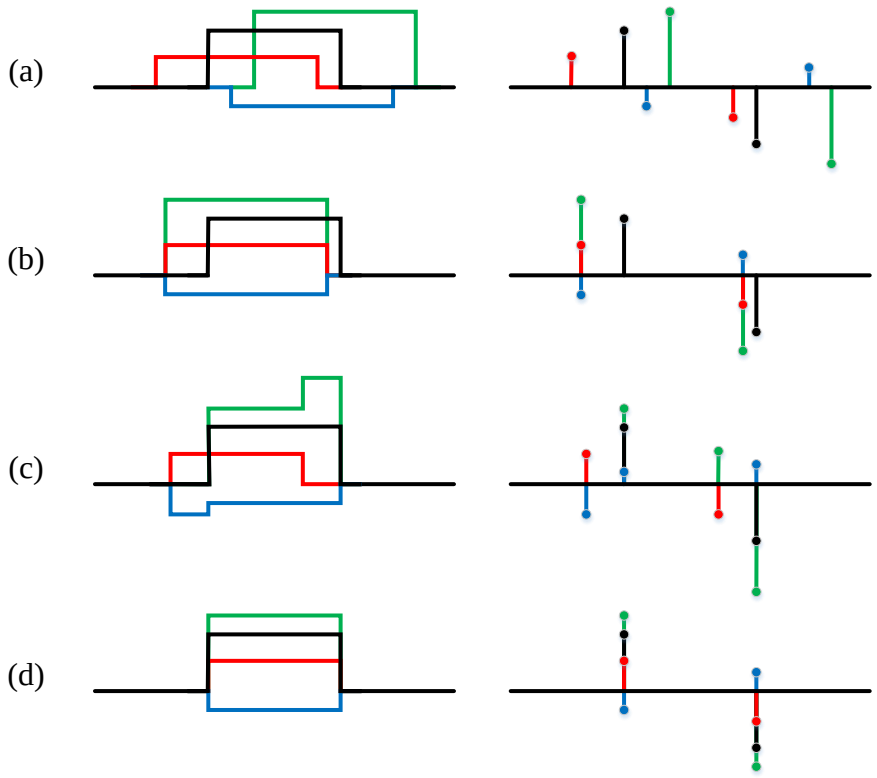


Figure 4.2. Illustration of possible solutions for different gradient based penalties. The black denotes a reference signal. RGB color lines denotes the solutions of different models. Left: 1D signals. Right: the corresponding gradients. (a) A possible solution of TV: the gradients of RGB channels are sparse but may not be correlated. (b) A possible solution of VTV: the gradients of R, G, B channels are group sparse, but may not be correlated to the reference signal. (c) A possible solution of (4.3): it does not encourage sparseness for each channel individually. (d) A possible solution of dynamic gradient sparsity regularization: the gradients can only be group sparse following the reference.

Comparing our method with existing methods [111][112][113], the first benefit of our method comes from the local spectral constraint. It does not rely on the upsampled MS image and linear-combination assumption. Therefore, only accurate spectral information is kept. It can be further applied to image fusion from different sources or different time acquisitions. Second, the proposed dynamic gradient sparsity only forces the support sets to be the same, while the sign of the gradients as well as

the magnitudes of the signal are not required to be the same. These properties make it invariant under contrast inversion [105] and not sensitive to illumination conditions. Last but not least, only our method can jointly fuse multiple bands simultaneously, which provides robustness to noise. These advantages exist in our method uniquely.

4.3.3 Algorithm

It is obviously that problem (4.7) is convex and has a global optimal solution. The first term is smooth while the second term is non-smooth. This motivates us to solve the problem in the FISTA framework [42]. It has been proven that FISTA can achieve the optimal convergence rate for first order methods. That is, $E(\mathbf{X}^k) - E(\mathbf{X}^*) \sim \mathcal{O}(1/k^2)$, where $\mathbf{X}^*$ is the optimal solution and k is the iteration counter. We summarize the proposed algorithm for pan-sharpening in Algorithm 5.

Here ψ^T denotes the inverse operator of ψ. L is the Lipschitz constant for $\psi^T(\psi\mathbf{X} - \mathbf{M})$, which can be set as 1 in this problem. We could observe that the solution is updated based on both $\mathbf{X}^k$ and $\mathbf{X}^{k-1}$, while the Bregman method that used in previous methods [111][113] updates $\mathbf{X}$ only based on $\mathbf{X}^k$. This is a reason why our method converges faster. For the second step, $L = 1$ and

$$\mathbf{X}^k = \arg\min_{\mathbf{X}} \{\frac{1}{2}\|\mathbf{X} - \mathbf{Y}\|_F^2 + \lambda||\nabla\mathbf{X} - \nabla D(\mathbf{P})||_{2,1}\}. \tag{4.8}$$

Let $\mathbf{Z} = \mathbf{X} - D(\mathbf{P})$ and we can rewrite the problem:

$$\mathbf{Z}^k = \arg\min_{\mathbf{Z}} \{\frac{1}{2}\|\mathbf{Z} - (\mathbf{Y} - D(\mathbf{P}))\|_F^2 + \lambda||\nabla\mathbf{Z}||_{2,1}\}. \tag{4.9}$$

This alternative problem is therefore a VTV denoising problem [115] and $\mathbf{X}^k$ can be updated by $\mathbf{Z}^k + D(\mathbf{P})$. The slow version of the VTV denoising algorithm [115] is accelerated based on FISTA framework to solve (4.8), which is summarized in Algorithm 6. It can be proven that Algorithm 6 reaches the optimal convergence rate if the step size is set as $1/(8\lambda)$.

Algorithm 5 DGS-Fusion

Input: L, λ, $t^1 = 1$, $\mathbf{Y}^0$

for $k = 1$ **to** $Maxiteration$ **do**

$\mathbf{Y} = \mathbf{Y}^k - \psi^T(\psi\mathbf{X} - \mathbf{M})/L$

$\mathbf{X}^k = \arg\min_{\mathbf{X}}\{\frac{L}{2}\|\mathbf{X} - \mathbf{Y}\|_F^2 + \lambda||\nabla\mathbf{X} - \nabla D(\mathbf{P})||_{2,1}\}$

$t^{k+1} = [1 + \sqrt{1 + 4(t^k)^2}]/2$

$\mathbf{Y}^{k+1} = \mathbf{X}^k + \frac{t^k - 1}{t^{k+1}}(\mathbf{X}^k - \mathbf{X}^{k-1})$

end for

The linear operator is defined as: $\mathcal{L}(\mathbf{R}, \mathbf{S})_{i,j,d} = \mathbf{R}_{i,j,d} - \mathbf{R}_{i-1,j,d} + \mathbf{S}_{i,j,d} - \mathbf{S}_{i,j-1,d}$ The corresponding inverse operator is defined as $\mathcal{L}^T(\mathbf{X}) = (\mathbf{R}, \mathbf{S})$ with $\mathbf{R}_{i,j,d} = \mathbf{X}_{i,j,d} - \mathbf{X}_{i+1,j,d}$ and $\mathbf{S}_{i,j,d} = \mathbf{X}_{i,j,d} - \mathbf{X}_{i,j+1,d}$. $\mathbb{P}$ is a projection operator used to ensure that $\sum_{d=1}^{s}(\mathbf{R}_{i,j,d}^2 + \mathbf{S}_{i,j,d}^2) \leq 1$, $|\mathbf{R}_{i,n,d}| \leq 1$, and $|\mathbf{S}_{m,j,d}| \leq 1$.

Algorithm 6 VTV-Denoising

Input: λ, $\mathbf{Y}$, $\mathbf{P}$, $(\mathbf{U}, \mathbf{V}) = (\mathbf{R}, \mathbf{S}) = (\mathbf{0}, \mathbf{0})$, $t^1 = 1$

$\mathbf{B} = \mathbf{Y} - D(\mathbf{P})$

for $k = 1$ **to** $Maxiteration$ **do**

$(\mathbf{R}^k, \mathbf{S}^k) = \mathbb{P}[(\mathbf{U}^k, \mathbf{V}^k) + \frac{1}{8\lambda}\mathcal{L}^T(\mathbf{B} - \lambda\mathcal{L}(\mathbf{U}^k, \mathbf{V}^k))]$

$t^{k+1} = \frac{1+\sqrt{1+4(t^k)^2}}{2}$

$(\mathbf{U}^{k+1}, \mathbf{V}^{k+1}) = (\mathbf{R}^k, \mathbf{S}^k) + \frac{t^k - 1}{t^{k+1}}(\mathbf{R}^k - \mathbf{R}^{k-1}, \mathbf{S}^k - \mathbf{S}^{k-1})$

end for

$\mathbf{Z} = \mathbf{B} - \lambda\mathcal{L}(\mathbf{R}^k, \mathbf{S}^k)$

$\mathbf{X} = \mathbf{Z} + D(\mathbf{P})$

If the subproblem is solved exactly by Algorithm 6, Algorithm 5 also can reach the optimal convergence rate $\mathcal{O}(1/k^2)$. However, due to the tradeoff between convergence rate and computational cost, the inner loop of Algorithm 6 only runs 3 iterations in all experiments.

Table 4.1. Comparison of Different Algorithms for Pan-sharpening. T_w denotes the time for wavelet fusion. R denotes the size of the blurring kernel.

Method	Scheme	Convergence Rate	Time Complexity
P+XS	Gradient Descent	very slow	$\mathcal{O}(N \log R)$
AVWP	Split Bregman	$< \mathcal{O}(1/k)$	$\mathcal{O}(N) + T_w$
FVP	Split Bregman	$< \mathcal{O}(1/k)$	$\mathcal{O}(N \log R)$
Proposed	FISTA	$\mathcal{O}(1/k^2)$	$\mathcal{O}(N)$

The comparison for different algorithms is presented in Table 1. In terms of convergence rate, FISTA [42] is much faster than the split Bregman method [114]. Also, only our method has linear complexity which makes it scalable for large datasets. There is only one insensitive parameter that needs to be tuned for our method while there are 2,4,4 parameters for the algorithms using X+PS [111], AVWP [112] and FVP [113], respectively. Without the need to estimate the blurring kernel and wavelet fused image, our method can be applied on different tasks more easily.

4.4 Experiment

The proposed method is validated on datasets from Quickbird, Geoeye, SPOT and IKONOS satellites. The resolution of Pan images ranges from 0.41 m to 1.5 m. All the corresponding MS images have lower resolutions with $c = 4$ and contain blue, green, red and near-infrared bands. For convenience, only the RGB bands

are presented. Due to the lack of multi-resolution images of the same scene, low resolution images are downsampled from the ground-truth. This strategy is common for comparing fusion algorithms (e.g. [116][117][111][112][113]).

We compare our method with classical methods PCA [106], IHS [107], wavelet [108], Brovey [110] and variation methods P+XS [111], AVWP [112]. We do not include FVP [113] as there is a serious mistake in the algorithm (Eq. (23) of [113]) and we can not reproduce their results. The parameters for each method are tuned individually according to the authors' suggestion and the best set is selected for each method, respectively. All experiments are conducted using Matlab on a desktop with 3.4GHz Intel core i7 3770 CPU.

4.4.1 Visual Comparison

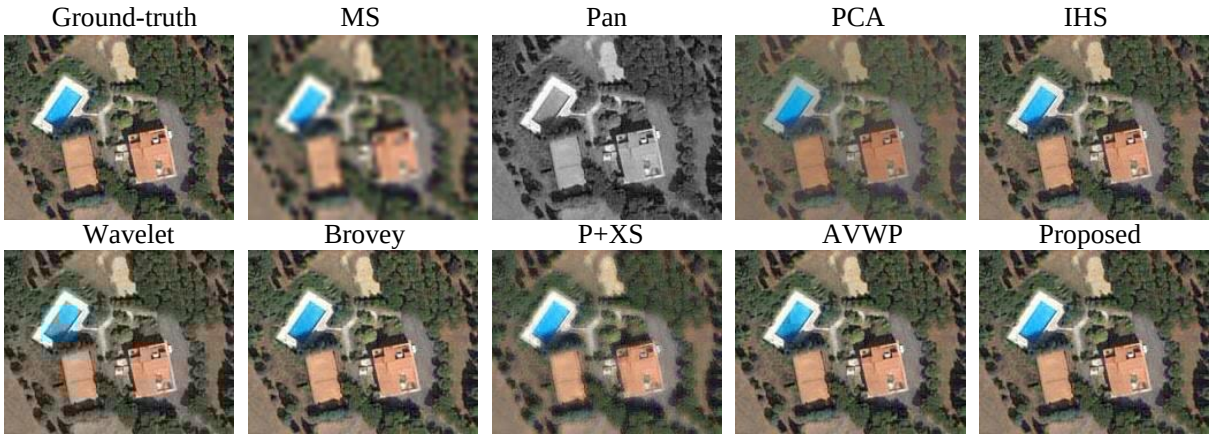


Figure 4.3. Fusion Results comparison (source: Quickbird). The Pan image has 200×160 pixels. Copyright DigitalGlobe .

First, we compare the fusion result by our method with those of previous works [106][107][108][110][111][112]. Figure 4.3 shows the results as well as the original

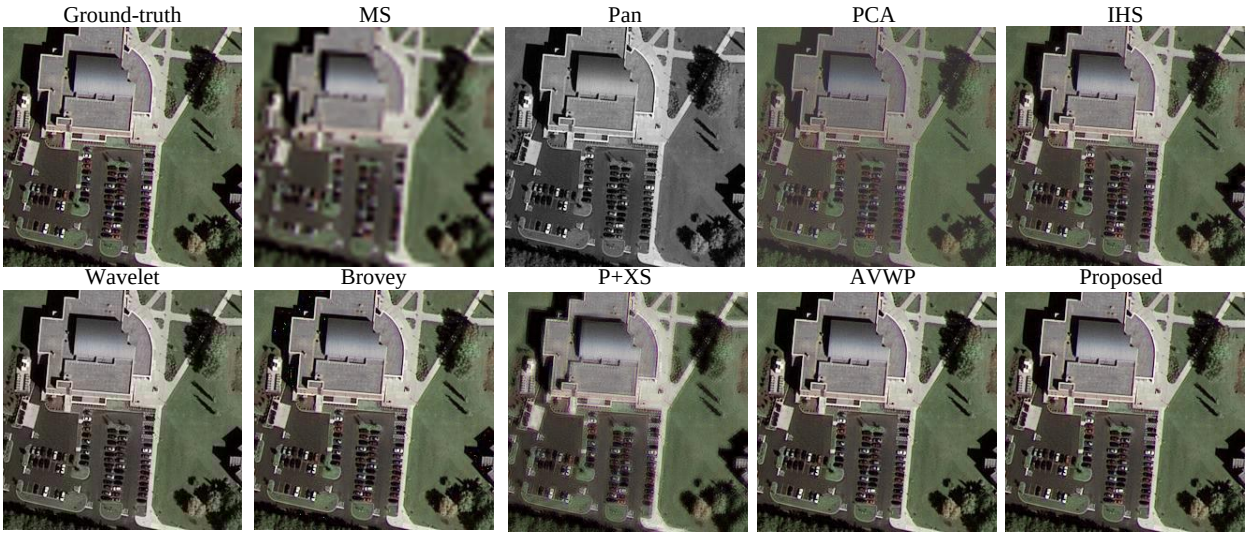


Figure 4.4. Fusion Results comparison (source: IKONOS). The Pan image has 256 × 256 pixels. Copyright DigitalGlobe.

images captured by the Quickbird satellite. All the methods can produce much better visual images than the original MS image. Obviously, PCA [106] performs the worst. No artifacts can be found on the images produced by IHS [107] and Brovey [110]. However, a closer look shows that the color on these images tends to change, especially on the trees and grass. This is a sign of spectral distortion [105]. Wavelet fusion [108] suffers from both spectral distortion and blocky artifacts (e.g. on the swimming poor). Blurred edges is a general issue in the image fused by P+XS [111]. AVWP [112] performs much better than all of them but it inherits the blocky artifacts of the wavelet fusion. The results of another experiment on a IKONOS image are shown in Figure 4.4, with similar performance by each algorithm. The difference is that some visible bright pixels can be found at the top-left corner of Brovey.

For better visualization, the error images compared with the ground-truth are presented in Figure 4.5 and Figure 4.6 at the same scale. From these error images, the spectral distortion, blocky artifacts, and blurriness can be clearly observed. These

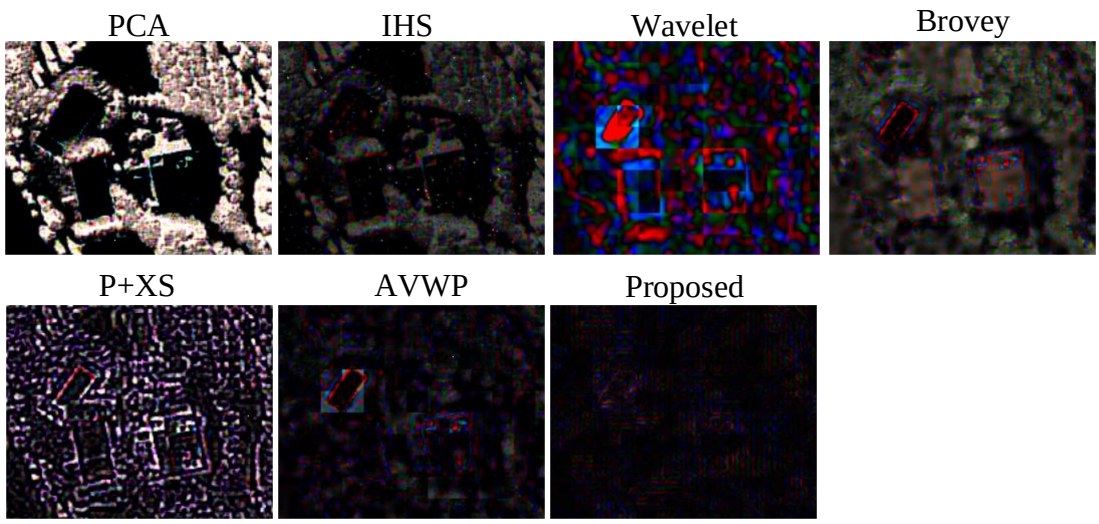


Figure 4.5. The corresponding error images to those in Figure 4.3. Brighter pixels represent larger errors.

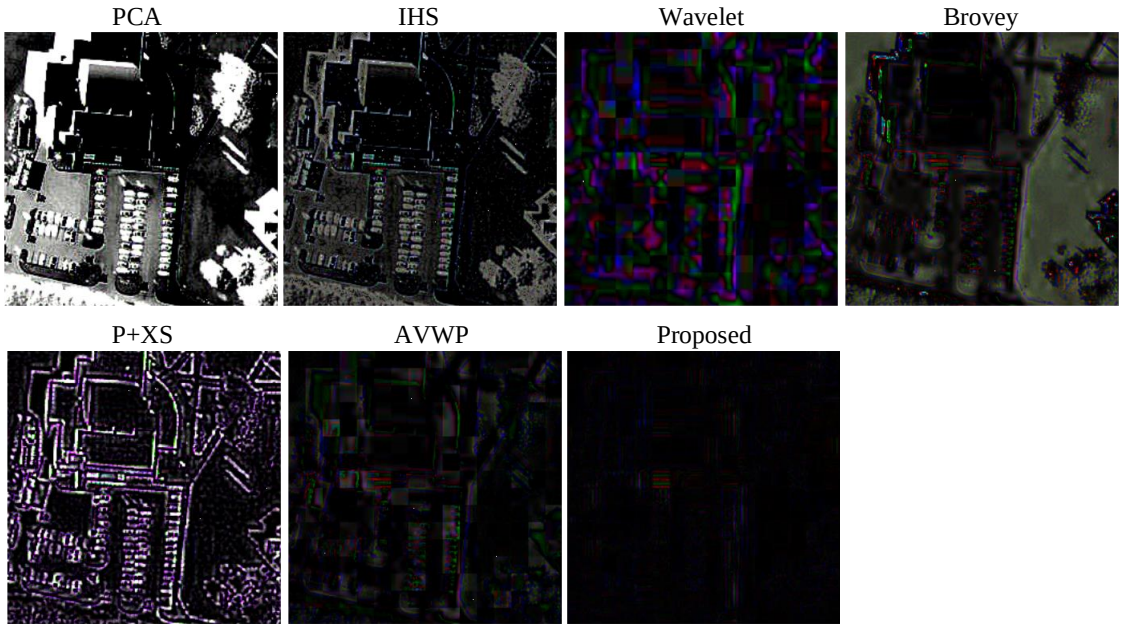


Figure 4.6. The corresponding error images to those in Figure 4.4. Brighter pixels represent larger errors.

results are consistent with those presented in previous work [112]. Due to the spectral distortion, the conventional methods are not adapted to vegetation study [105]. Previous variational methods [111][112] try to break such hard assumptions by combining a few weak assumptions. However, their assumption on the upsampled MS image always results in inaccuracy. In contrast, we only constrain the spectral information of the fused image to be locally consistent with the original MS image. The fusion results are impressively good on these two images.

4.4.2 Quantitative Analysis

Figure 4.7. Example images used in our experiments. Copyright DigitalGlobe for Quickbird, Geoeye and IKONOS. Copyright CNES for SPOT.

In addition to the two images used previously, 156 test images of different sizes (from 128×128 to 512×512) are cropped from Quickbird, Geoeye, IKONOS and SPOT datasets, which contain vegetation (e.g. forest, farmland), bodies of water (e.g. river, lake) and urban scenes (e.g. building, road). This test set is much larger

than the size of all datasets considered in previous variational methods (31 images in [111], 7 images in [112] and 4 images in [113]). Example images are shown in Figure 4.7.

To evaluate the fusion quality of different methods, we use four metrics that measure spectral quality and one metric that measures spatial quality. The spectral metrics include the relative dimensionless global error in synthesis (ERGAS) [118], spectral angle mapper (SAM) [118], universal image quality index (Q-average) [119] and relative average spectral error (RASE) [120]. The filtered correlation coefficients (FCC) [108] is used as spatial quality metric. In addition, peak signal-to-noise ratio (PSNR), and root mean squared error (RMSE) and mean structural similarity (MSSIM) [121] are used to evaluate the fusion accuracy when compared with the ground-truth.

The average results and the variance on this test set are listed in Table 2. The ideal value for each metric is shown in the last row. The results of variational methods [111][112] have much lower values in ERGAS and RASE than those of conventional methods [106][107][108][110]. From QAVE and SAM, the results are comparable to conventional methods. We can conclude that these variational methods can preserve more spectral information. Due to the blurriness, P+XS has the worse spatial resolution in terms of FCC. In terms of error and similarity metrics (PSNR, MSSIM, RMSE), AVWP and P+XS are always the second best and second worst, respectively. Except for the same FCC as the wavelet fusion, our method is consistently better than all previous methods in terms of all metrics. These results are enough to demonstrate the success of our method, where the dynamic gradient sparsity can preserve sharp edges and the spectral constraint keeps accurate spectral information. In terms of PSNR, it can outperform the second best method AVWP by more than 7 dB.

Table 4.2. Performance Comparison on the 158 remotely sensed images.

Method	ERGAS	QAVE	RASE	SAM	FCC	PSNR	MSSIM	RMSE
PCA [106]	5.67±1.77	0.664±0.055	22.3±6.8	2.11±1.35	0.972±0.014	20.7±2.7	0.799±0.067	24.1±6.7
IHS [107]	1.68±0.86	0.734±0.011	6.63±3.4	0.79±0.54	0.989±0.006	31.2±4.6	0.960±0.035	8.1±4.2
Wavelet[108]	1.18±0.45	0.598±0.113	4.50±1.6	2.45±1.18	**0.997±0.002**	36.1±3.6	0.983±0.009	4.5±1.9
Brovey [110]	1.22±1.08	0.733±0.011	5.18±4.6	0.61±0.58	0.940±0.170	38.2±5.6	0.989±0.008	9.1±19.7
P+XS[111]	0.89±0.33	0.720±0.036	3.47±1.3	0.66±0.36	0.898±0.024	25.9±3.5	0.854±0.051	14.7±5.4
AVWP[112]	0.46±0.17	0.733±0.013	1.81±0.6	0.69±0.70	0.996±0.002	40.0±3.5	0.991±0.006	2.9±1.0
Proposed	**0.07±0.03**	**0.746±0.004**	**0.3±0.1**	**0.18± 0.11**	**0.997±0.002**	**47.5±3.6**	**0.998±0.001**	**1.1±0.5**
Reference Value	0	1	0	0	1	$+\infty$	1	0

If we consider the prior information that is used, the performance of each algorithm is easy to explain. Conventional projection-substitution methods only treat the input images as vectorial information (i.e. 1D). The difference is the substitution performed on various projection spaces. However, 2D information such as boundaries is not utilized. The boundary information has been considered in both variational methods P+XS [111] and AVWP [112], although their models can not effectively exploit this prior information. Promising results, especially by AVWP, have already achieved over conventional methods. By using the proposed dynamic gradient sparsity, our method has successfully learned the prior knowledge provided by the Pan image. Due to the group sparsity across different bands, our method is less sensitive to noise. These are why our method consistently outperforms the others.

4.4.3 Efficiency Comparison

To evaluate the efficiency of the proposed method, we compare the proposed method with previous variational methods P+XS [111] and AVWP [112] in terms of both accuracy and computational cost. PSNR is used to measure fusion accuracy. Figure 4.8 demonstrates the convergence rate comparison of these algorithms corresponding to the images in Figure 4.3 and 4.4. Inheriting the benefit of the FISTA [42] framework, our method often converges in 100 to 150 outer iterations. AVWP often converges in 200 to 400 iterations. P+XS that uses classic gradient decent method has not converged even with 600 iterations. After each algorithm converged, our method can approximately outperform AVWP by more than 5 dB and 8 dB on these two image in terms of PSNR. Note that the later one is the second best method from previous analysis.

The average computational costs of these three methods are listed in Table 3 for different sizes of test images. Both the proposed method and AVWP terminate

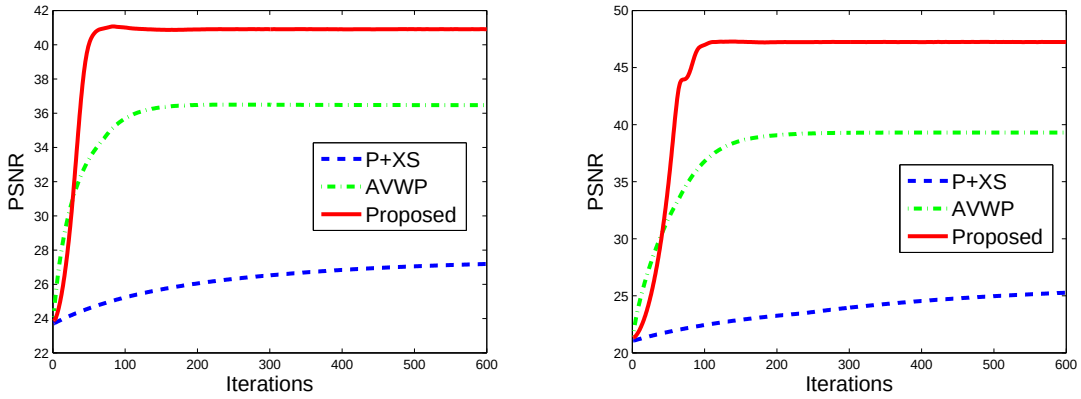


Figure 4.8. Convergence rate comparison among P+XS, AVWP and the proposed method. (a) Result corresponds to Figure 4.3. (b) Result corresponds to Figure 4.4.

when a fixed tolerance is reached (e.g. 10^{-3} of the relative change on $\mathbf{X}$). The computational cost of our method tend to be linear from these results. Even the second fastest method AVWP takes about 50% more time than ours on an image of 512 by 512 pixels. These comparisons are sufficient to demonstrate the efficiency and effectiveness of our method.

Table 4.3. Computational time (second) comparison.

	128×128	256×256	384×384	512×512
P+XS	6.7	16.0	48.3	87.4
AVWP	1.7	8.3	28.2	54.7
Proposed	**1.4**	**5.0**	**19.3**	**36.8**

4.5 Summery

We have proposed a novel and powerful variational model for pan-sharpening with local spectral consistency and dynamic gradient sparsity. The model naturally

incorporates the gradient prior information from a high resolution Pan image and spectral information from an MS image. Moreover, our model also exploits the band correlations of the fused image itself, which has not yet considered in any previous method. An efficient optimization algorithm has been devised to solve the problem, with sufficient implementation details. Extensive experiments are conducted on 158 images stemming from a variety of sources. Due to the proposed unique techniques, our methods is corroborated to consistently outperforms the state-of-the-arts in terms of both spatial and spectral qualities.

The proposed method is not sensitive to an exact downsampling scheme. When the actual downsampling is cubic interpolation, the same result can be obtained if we use either cubic or bilinear interpolation. It is only sightly worse when we use an averaging strategy.

Parallel programming can further accelerate the speed of our method. No information from another patch is required for fusion. As our method does not require strong correlation between the input images, it can be used in fusing images from different capture times (shadows should be taken into account), different sources (e.g. images from different satellites). Conventional TV in medical image fusion (e.g. CT and MRI) [122] may be enhanced by the proposed dynamic gradient sparsity.

CHAPTER 5

Deep Sparse Representation for Robust Image Registration

This chapter introduces a hierarchical sparsity model called deep sparse representation for image registration. The proposed method is motivated by that the optimally registered images can be deeply sparsified in the gradient domain and frequency domain, with the separation of a sparse tensor of errors. One of the key advantages of the proposed similarity measure is its robustness to severe intensity distortions, which widely exist on medical images, remotely sensed images and natural photos due to the difference of acquisition modalities or illumination conditions. This work was presented under a slightly modification from [123].

5.1 Introduction

Image registration is a fundamental task in image processing and computer vision [124, 125, 126, 6, 127]. It aims to align two or more images into the same coordinate system, and then these images can be processed or compared. Accuracy and robustness are two of the most important metrics to evaluate a registration method. It has been shown that a mean geometric distortion of only 0.3 pixel will result in noticeable effect on a pixel-to-pixel image fusion process [128]. Robustness is defined as the ability to get close to the accurate results on different trials under diverse conditions. Based on the feature used in registration, existing methods can be classified into feature-based registration (e.g., [129, 130]) and pixel-based registration (e.g., [131, 9, 11, 132]). Feature-based methods rely on the landmarks extracted from the images. However, extracting reliable features is still an open problem and an

active topic of research [127]. In this chapter, we are interested in image registration by directly using their pixel values. In addition, we wish to successfully register the images from a variety of applications in subpixel-level accuracy, as precisely as possible.

One key component for image registration is the energy function to measure (dis)similarity. The optimized similarity should lead to the correct spatial alignment. However, finding a reliable similarity measure is quite challenging due to the unpredicted variations of the input images. In many real-world applications, the images to be registered may be acquired at different times and locations, under various illumination conditions and occlusions, or by different acquisition modalities. As a result, the intensity fields of the images may vary significantly. For instance, slow-varying intensity bias fields often exist in brain magnetic resonance images [133]; the remotely sensed images may even have inverse contrast for the same land objects, as multiple sensors have different sensitivities to wavelength spectrum [105]. Unfortunately, many existing pixel-based similarity measures are not robust to these intensity variations, e.g., the widely used sum-of-squared-difference (SSD) [6].

Recently, the sparsity-inducing similarity measures have been repeatedly successful in overcoming such registration difficulties [7, 134, 135, 136]. In RASL [134] (robust alignment by sparse and low-rank decomposition), the images are vectorized to form a data matrix. The transformations are estimated to seek a low rank and sparse representation of the aligned images. Two online alignment methods, ORIA [135] (online robust image alignment) and t-GRASTA [136] (transformed Grassmannian robust adaptive subspace tracking algorithm), are proposed to improve the scalability of RASL. All of these methods assume that the large errors among the images are sparse (e.g., caused by shadows, partial occlusions) and separable. However, as we will show later, many real-world images contain severe spatially-varying intensity

distortions. These intensity variations are not sparse and therefore difficult to be separated by these methods. As a result, the above measures may fail to find the correct alignment and thus are less robust in these challenging tasks.

The residual complexity (RC) [7] is one of the best measures for registering two images corrupted by severe intensity distortion [137], which uses the discrete cosine transform (DCT) to sparsify the residual of two images. For a batch of images, RC has to register them pair-by-pair and the solution may be sub-optimal. In addition, DCT and inverse DCT are required in each iteration, which slows down the overall speed of registration. Finally, although RC is robust to intensity distortions, the ability of RC to handle partial occlusions is unknown.

Unlike previous works that vectorize each image into a vector [134, 135, 136], we arrange the input images into a 3D tensor to keep their spatial structure. With this arrangement, the optimally registered image tensor can be deeply sparsified into a sparse frequency tensor and a sparse error tensor (see Fig. 5.1 for more details). Severe intensity distortions and partial occlusions will be sparsified and separated out in the first and second layers, while any misalignment will increase the sparseness of the frequency tensor (third layer). We propose a novel similarity measure based on such deep sparse representation of the natural images. Compared with the low rank similarity measure which requires a batch of input images, the proposed similarity measure still works even when there are only two input images. An efficient algorithm based on the Augmented Lagrange Multiplier (ALM) method is proposed for the batch mode, while the gradient descent method with backtracking is presented to solve the pair registration problem. Both algorithms have very low computational complexity in each iteration. We compare our method with 9 traditional and state-of-the-art algorithms on a wide range of natural image datasets, including medical images, remotely sensed images and photos. Extensive results demonstrate that our

method is more robust to different types of intensity variations and always achieves higher sub-pixel accuracy over all the tested methods.

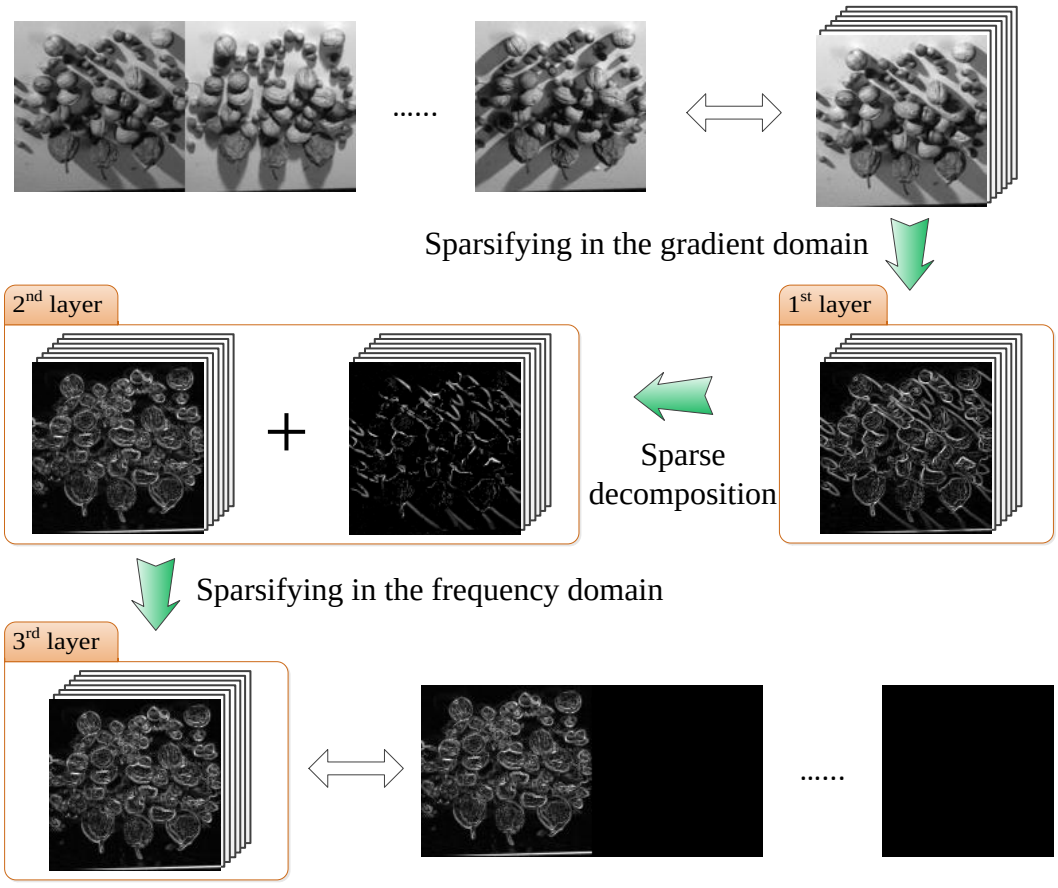


Figure 5.1. Deep sparse representation of the optimally registered images. First we sparsify the image tensor into the gradient tensor (1st layer). The sparse error tensor is then separated out in the 2nd layer. The gradient tensor with repetitive patterns are sparsified in the frequency domain. Finally we obtain an extremely sparse frequency tensor (composed of Fourier coefficients) in the 3rd layer.

5.2 Image registration via deep sparse representation

In this chapter, we use bold letters denote multi-dimensional data. For example, $\mathbf{x}$ denotes a vector, $\mathbf{X}$ denotes a matrix and $\boldsymbol{\mathcal{X}}$ is a 3D or third-order tensor. $\boldsymbol{\mathcal{X}}_{(i,j,t)}$

denotes the entry in the i-th row, j-th column and t-th slice. $\boldsymbol{\mathcal{X}}_{(:,:,t)}$ denotes the whole t-th slice, which is therefore a matrix. The ℓ_1 norm is the summation of absolute values of all entries, which applies to vector, matrix and tensor.

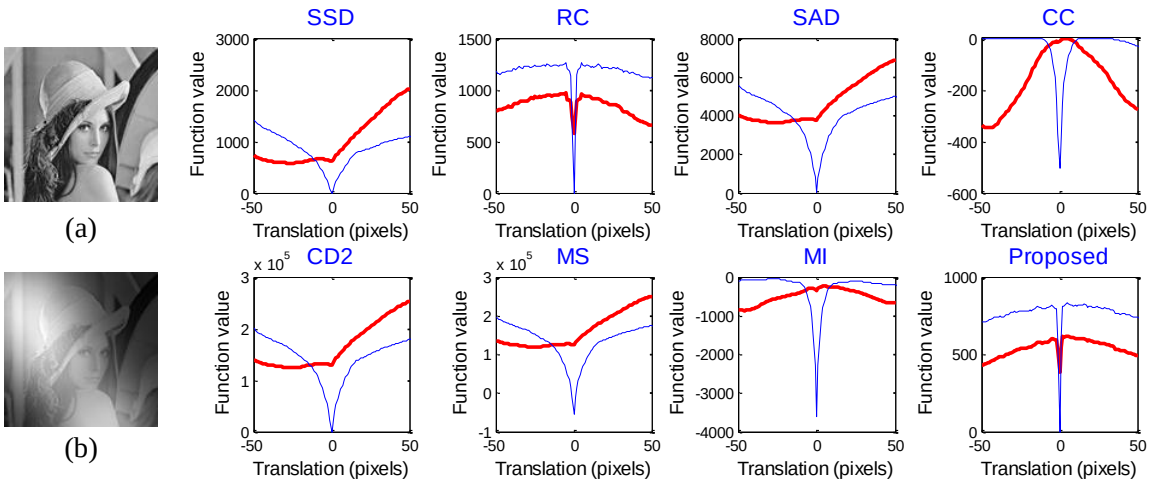


Figure 5.2. A toy registration example with respect to horizontal translation using different similarity measures (SSD [6], RC [7], SAD [6], CC [8], CD2 [9], MS [10], MI [11] and the proposed pair mode). (a) The Lena image (128×128). (b) A toy Lena image under a severe intensity distortion. Blue curves: registration between (a) and (a); red curves: registration between (b) and (a).

5.2.1 Batch mode

We introduce our deep sparsity architecture in the inverse order for easy understanding. Suppose we have a batch of grayscale images $\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_N \in \mathbb{R}^{w \times h}$ to be registered, where N denotes the total number of images. First, we consider the simplest case that all the input images are identical and perturbed from a set of transformations $\tau = \{\tau_1, \tau_2, ..., \tau_N\}$ (it can be affine, non-rigid, etc.).

We arrange the input images into a 3D tensor $\boldsymbol{\mathcal{D}} \in \mathbb{R}^{w \times h \times N}$, with

$$\boldsymbol{\mathcal{D}}_{(:,:,t)} = \mathbf{I}_t, \quad t = 1, 2, ..., N, \tag{5.1}$$

After removing the transformation perturbations, the slices show repetitive patterns. Such periodic signals are extremely sparse in the frequency domain. Ideally the Fourier coefficients from the second slice to the last slice should be all zeros. We can minimize the ℓ_1 norm of the Fourier coefficients to seek the optimal transformations:

$$\min_{\boldsymbol{\mathcal{A}},\tau} \ ||\mathcal{F}_N \boldsymbol{\mathcal{A}}||_1, \ s.t. \ \boldsymbol{\mathcal{D}} \circ \tau = \boldsymbol{\mathcal{A}}, \tag{5.2}$$

where $\mathcal{F}_N$ denotes the Fourier transform in the third direction.

The above model can be hardly used on practical cases, due to the corruptions and partial occlusions in the images. Similar as previous work [134], we assume the noise is negligible in magnitude as compared to the error caused by occlusions. Let $\boldsymbol{\mathcal{E}}$ be the error tensor. We can separate it from the image tensor if it is sparse enough. Similar, we use the ℓ_1 norm to induce sparseness:

$$\min_{\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{E}},\tau} \ ||\mathcal{F}_N \boldsymbol{\mathcal{A}}||_1 + \lambda ||\boldsymbol{\mathcal{E}}||_1, \ s.t. \ \boldsymbol{\mathcal{D}} \circ \tau = \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{E}}, \tag{5.3}$$

where $\lambda > 0$ is a regularization parameter.

The above approach requires that the error $\boldsymbol{\mathcal{E}}$ is sparse. However, in many real-world applications, the images are corrupted with spatially-varying intensity distortions. Existing methods such as RASL [134] and t-GRASTA [136] may fail to separate these non-sparse errors. The last stage of our method comes from the intuition that the locations of the image gradients (edges) should almost keep the same, even under severe intensity distortions. Therefore, we register the images in the gradient domain:

$$\min_{\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{E}},\tau} \ ||\mathcal{F}_N \boldsymbol{\mathcal{A}}||_1 + \lambda ||\boldsymbol{\mathcal{E}}||_1, \ s.t. \ \nabla\boldsymbol{\mathcal{D}} \circ \tau = \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{E}}, \tag{5.4}$$

where $\nabla\boldsymbol{\mathcal{D}} = \sqrt{(\nabla_x \boldsymbol{\mathcal{D}})^2 + (\nabla_y \boldsymbol{\mathcal{D}})^2}$ denotes the gradient tensor along the two spatial directions. This is based on a mild assumption that the intensity distortion fields of natural images often change smoothly.

With this rationale, the input images can be sparsely represented in a three layer architecture, which is shown in Fig. 5.1. We call it deep sparse representation of images. Comparing with existing popular low rank representation [134], our modeling has two major advantages. First, the low rank representation treats each image as a 1D signal, while our modeling exploits the spatial prior information (piece-wise smoothness) of natural images. Second, when the number of input images is not sufficient to form a low rank matrix, our method is still effective. Next, we will demonstrate how does our method register only two input images.

5.2.2 Pair mode

For registering a pair of images, our model can be simplified and the registration can be accelerated. After two-point discrete Fourier transform (DFT), the first entry is the sum and the second entry is the difference. The difference term is much sparser than the sum term when the two images have been registered. We can discard the sum term to seek a sparser representation. Let $\mathbf{I}_1$ be the reference image, and $\mathbf{I}_2$ be the source image to be registered. The problem (5.4) can be simplified to

$$\begin{aligned} &\min_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{E},\tau} ||\mathbf{A}_1 - \mathbf{A}_2||_1 + \lambda||\mathbf{E}||_1, \\ &\quad s.t.\ \nabla\mathbf{I}_1 = \mathbf{A}_1, \nabla\mathbf{I}_2 \circ \tau = \mathbf{A}_2 + \mathbf{E}. \end{aligned} \tag{5.5}$$

Both ℓ_1 norms in (5.5) implies the same property, i.e., sparseness of the residual image $\mathbf{E}$. Therefore, we can further simplify the above energy function:

$$\min_{\tau} ||\nabla\mathbf{I}_1 - \nabla\mathbf{I}_2 \circ \tau||_1. \tag{5.6}$$

It's interesting that (5.6) is equivalent to minimizing the total variation (TV) of the residual image. The TV has been successfully utilized in many image denoising,

debluring, and reconstruction problems. To our best knowledge, this is the first attempt to define TV as a similarity measure[1].

We compare the proposed similarity measure with SSD [6], RC [7], sum-of-absolute value (SAD) [6], correlation coefficient (CC) [8], CD2 [9], MS [10] and mutual information (MI) [11] on a toy example. The Lena image is registered with itself with respect to the horizontal translations. The blue curves in Fig. 5.2 show the responses of different measures, all of which can find the optimal alignment at the zero translation. After adding intensity distortions and rescaling, the appearance of source image shown in Fig. 5.2(b) is not consistent with that of the original Lena image. The results denoted by the red curves show that only RC and the proposed pair mode can handle this intensity distortion while other methods fail.

5.3 Algorithms

5.3.1 Batch mode

Problem (5.4) is difficult to solve directly due to the non-linearity of the transformations τ. We use the local first order Taylor approximation for each image:

$$\nabla \mathbf{I}_t \circ (\tau_t + \triangle \tau_t) \approx \nabla \mathbf{I}_t \circ \tau_t + \boldsymbol{\mathcal{J}}_t \otimes \triangle \tau_t \tag{5.7}$$

for $t = 1, 2, ..., N$, where $\boldsymbol{\mathcal{J}}_t = \frac{\partial}{\partial \zeta}(\nabla \mathbf{I}_t \circ \zeta)|_{\zeta=\tau_t} \in \mathbb{R}^{w \times h \times p}$ when τ_t is defined by p parameters. The *Tensor-Vector Product* of the last term is defined by:

Definition 1. *Tensor-Vector Product.* The product of a tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and a vector $\mathbf{b} \in \mathbb{R}^{n_3}$ is a matrix $\mathbf{C} \in \mathbb{R}^{n_1 \times n_2}$. It is given by $\mathbf{C} = \boldsymbol{\mathcal{A}} \otimes \mathbf{b}$, where $\mathbf{C}_{(i,j)} = \sum_{t=1}^{n_3} \boldsymbol{\mathcal{A}}_{(i,j,t)} \mathbf{b}_{(t)}$, for $i = 1, 2, ..., n_1$ and $j = 1, 2, ..., n_2$.

[1]It is substantially different from the TV regularization methods [138, 139], where the widely used SSD is their actual similarity measure.

Based on this, the batch mode (5.4) can be rewritten as:

$$\begin{aligned} &\min_{\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{E}},\triangle\tau} ||\mathcal{F}_N\boldsymbol{\mathcal{A}}||_1 + \lambda||\boldsymbol{\mathcal{E}}||_1, \\ &\quad s.t.\ \nabla\boldsymbol{\mathcal{D}} \circ \tau + \boldsymbol{\mathcal{J}} \otimes \triangle\tau = \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{E}}, \end{aligned} \tag{5.8}$$

This constrained problem can be solved by the augmented Lagrange multiplier (ALM) algorithm [134, 140]. The augmented Lagrangian problem is to iteratively update $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}, \triangle\tau$ and $\boldsymbol{\mathcal{Y}}$ by

$$\begin{aligned} (\boldsymbol{\mathcal{A}}^{k+1}, \boldsymbol{\mathcal{E}}^{k+1}, \triangle\tau^{k+1}) &= \arg\min_{\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{E}},\triangle\tau} \mathcal{L}(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}, \triangle\tau, \boldsymbol{\mathcal{Y}}), \\ \boldsymbol{\mathcal{Y}}^{k+1} &= \boldsymbol{\mathcal{Y}}^k + \mu^k h(\boldsymbol{\mathcal{A}}^k, \boldsymbol{\mathcal{E}}^k, \triangle\tau^k), \end{aligned} \tag{5.9}$$

where k is the iteration counter and

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}, \triangle\tau, \boldsymbol{\mathcal{Y}}) =& < \boldsymbol{\mathcal{Y}}, h(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}, \triangle\tau) > + ||\mathcal{F}_N\boldsymbol{\mathcal{A}}||_1 \\ &+ \lambda||\boldsymbol{\mathcal{E}}||_1 + \frac{\mu}{2}||h(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}, \triangle\tau)||_F^2, \end{aligned} \tag{5.10}$$

where the inner product of two tensors is the sum of all the element-wise products and

$$h(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}, \triangle\tau) = \nabla\boldsymbol{\mathcal{D}} \circ \tau + \boldsymbol{\mathcal{J}} \otimes \triangle\tau - \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{E}}. \tag{5.11}$$

A common strategy to solve (5.9) is to minimize the function against one unknown at one time. Each of the subproblem has a closed form solution:

$$\begin{aligned} \boldsymbol{\mathcal{A}}^{k+1} &= \mathcal{T}_{1/\mu^k}(\nabla\boldsymbol{\mathcal{D}} \circ \tau + \boldsymbol{\mathcal{J}} \otimes \triangle\tau + \frac{1}{\mu^k}\boldsymbol{\mathcal{Y}}^k - \boldsymbol{\mathcal{E}}^k) \\ \boldsymbol{\mathcal{E}}^{k+1} &= \mathcal{T}_{\lambda/\mu^k}(\nabla\boldsymbol{\mathcal{D}} \circ \tau + \boldsymbol{\mathcal{J}} \otimes \triangle\tau + \frac{1}{\mu^k}\boldsymbol{\mathcal{Y}}^k - \boldsymbol{\mathcal{A}}^{k+1}) \\ \triangle\tau_t^{k+1} &= \boldsymbol{\mathcal{J}}_t^T \otimes (\boldsymbol{\mathcal{A}}_{(:,:,t)}^{k+1} + \boldsymbol{\mathcal{E}}_{(:,:,t)}^{k+1} - \nabla\boldsymbol{\mathcal{D}}_{(:,:,t)} \circ \tau \\ &\quad - \frac{1}{\mu^k}\boldsymbol{\mathcal{Y}}_{(:,:,t)}^k), \quad \text{for } t = 1, 2, ..., N \end{aligned} \tag{5.12}$$

where the $\mathcal{T}_\alpha()$ denotes the soft thresholding operation with threshold value α. In the third equation of (5.12), we use the *Tensor-Matrix Product* and *Tensor Transpose* defined as follows:

Definition 2. *Tensor-Matrix Product*. The product of a tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and a matrix $\mathbf{B} \in \mathbb{R}^{n_2 \times n_3}$ is a vector $\mathbf{c} \in \mathbb{R}^{n_1}$. It is given by $\mathbf{c} = \boldsymbol{\mathcal{A}} \otimes \mathbf{B}$, where $\mathbf{c}_{(i)} = \sum_{j=1}^{n_2} \sum_{t=1}^{n_3} \boldsymbol{\mathcal{A}}_{(i,j,t)} \mathbf{B}_{(j,t)}$, for $i = 1, 2, ..., n_1$.

Definition 3. *Tensor Transpose*. The transpose of a tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is the tensor $\boldsymbol{\mathcal{A}}^T \in \mathbb{R}^{n_3 \times n_1 \times n_2}$.

The registration algorithm for the batch mode is summarized in Algorithm 7. Let $M = w \times h$ be the number of pixels of each image. We set $\lambda = 1/\sqrt{M}$ and $\mu_k = 1.25^k \mu_0$ in the experiments, where $\mu_0 = 1.25/||\nabla D||_2$. For the inner loop, applying the fast Fourier transform (FFT) costs $\mathcal{O}(N \log N)$. All the other steps cost $\mathcal{O}(MN)$. Therefore, the total computation complexity of our method is $\mathcal{O}(N \log N + MN)$, which is significantly faster than $\mathcal{O}(N^2 M)$ when applying SVD decomposition in RASL (if $M >> N$).

5.3.2 Pair mode

Similar as that in the batch mode, we have:

$$\nabla \mathbf{I}_2 \circ (\tau + \triangle\tau) \approx \nabla \mathbf{I}_2 \circ \tau + \boldsymbol{\mathcal{J}}_p \otimes \triangle\tau \tag{5.13}$$

where $\boldsymbol{\mathcal{J}}_p \in \mathbb{R}^{w \times h \times p}$ denotes the Jacobian. Thus, the pair mode (5.6) is to minimize the energy function with respect to $\triangle\tau$:

$$E(\triangle\tau) = ||\nabla \mathbf{I}_1 - \nabla \mathbf{I}_2 \circ \tau - \boldsymbol{\mathcal{J}}_p \otimes \triangle\tau||_1 \tag{5.14}$$

The ℓ_1 norm in (5.14) is not smooth. We can have a tight approximation for the absolute value: $|x| = \sqrt{x^2 + \epsilon}$, where ϵ is a small constant (e.g. 10^{-10}). Let $\mathbf{r} =$

Algorithm 7 Image registration via DSR - batch mode

input: Images $\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_N$, initial transformations $\tau_1, \tau_2, ..., \tau_N$, regularization parameter λ.

repeat

1) Compute $\boldsymbol{\mathcal{J}}_t = \frac{\partial}{\partial \zeta}(\nabla \mathbf{I}_t \circ \zeta)|_{\zeta=\tau_t}$, $t = 1, 2, ..., N$;

2) Warp and normalize the gradient images:

$\nabla \boldsymbol{\mathcal{D}} \circ \tau = [\frac{\nabla \mathbf{I}_1 \circ \tau_1}{||\nabla \mathbf{I}_1 \circ \tau_1||_F}; ...; \frac{\nabla \mathbf{I}_N \circ \tau_N}{||\nabla \mathbf{I}_N \circ \tau_N||_F}]$;

3) Use (5.12) to iteratively solve the minimization problem of ALM:

$\boldsymbol{\mathcal{A}}^*, \boldsymbol{\mathcal{E}}^*, \triangle\tau^* = \arg\min \mathcal{L}(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}, \triangle\tau, \boldsymbol{\mathcal{Y}})$;

4) Update transformations: $\tau = \tau + \triangle\tau^*$;

until Stop criteria

$\nabla \mathbf{I}_1 - \nabla \mathbf{I}_2 \circ \tau - \boldsymbol{\mathcal{J}}_p \otimes \triangle\tau$, and we can obtain the gradient of the energy function by the chain rule:

$$\nabla E(\triangle\tau) = \boldsymbol{\mathcal{J}}_p^T \otimes \frac{\mathbf{r}}{\sqrt{\mathbf{r} \circ \mathbf{r} + \epsilon}} \tag{5.15}$$

where $\circ$ denotes the Hadamard product. Note that the division in (5.15) is element-wise.

Gradient descent with backtracking is used to minimize the energy function (5.14), which is summarized in Algorithm 8. We set the initial step size $\mu^0 = 1$ and $\eta = 0.8$. The computational complexity of each iteration is $\mathcal{O}(M)$, which is much faster than $\mathcal{O}(M \log M)$ in RC when fast cosine transform (FCT) is applied [7]. Similar as the batch mode, we use the normalized images to rule out the trivial solutions. For non-rigid registration, the transformation is implemented using the free form deformation (FFD) transformation with B-spline control points [141]. We use a coarse-to-fine hierarchical registration architecture for both the batch mode and pair mode [142].

Algorithm 8 Image registration via DSR - pair mode

input: $\mathbf{I}_1$, $\mathbf{I}_2$, $\eta < 1$, τ, μ^0.

repeat

1) Warp and normalize $\mathbf{I}_2$ with τ;

2) $\mu = \mu^0$;

3) Compute $\triangle\tau = -\mu\nabla E(\mathbf{0})$;

4) If $E(\triangle\tau) > E(\mathbf{0})$,

set $\mu = \eta\mu$ and go back to 3);

5) Update transformation: $\tau = \tau + \triangle\tau$;

until Stop criteria

5.4 Experimental results

In this section, we validate our method on a wide range of applications. We compare our batch mode with RASL [134] and t-GRASTA [136], and compare our pair mode with RC [7] and SSD [6]. One of the most important advantages of our method is its robustness and accuracy on natural images under spatially-varying intensity distortions. As shown in [7] and Fig. 5.2, SAD [6], CC [8], CD2 [9], MS [10], MI [11] are easy to fail in such cases. We do not include them in the following experiments. All experiments are conducted on a desktop computer with Intel i7-3770 CPU with 12GB RAM.

5.4.1 Batch image registration

To evaluate the performance of our batch mode, we use a popular database of naturally captured images [143]. We choose the four datasets with the largest lighting variations: "NUTS", "MOVI", "FRUITS" and "TOY". These datasets are very challenging to register, as they have up to 20 different lighting conditions and are

occluded by varying shadows. Random translations on both directions are applied on the four datasets, which are drawn from a uniform distribution in a range of 10 pixels.

After registration on the "NUTS" dataset, the two components of each algorithm is shown in Fig. 5.3. RASL [134] and t-GRASTA [136] fail to separate the shadows and large errors, while we can successfully find the deep sparse representation of the optimally registered images. The average of perturbed images and results are shown in Fig. 5.4, where the average image by the proposed method has significantly sharper edges than those by the two existing methods. The quantitative comparisons on the four datasets are listed in Table 5.1 over 20 random runs. The overall average errors of our method are consistently lower than those of RASL and t-GRASTA. More importantly, only our method can always achieve subpixel accuracy. For 20 images with size 128×128 pixels, the registration time is around 7 seconds for both RASL and our method, while t-GRAST costs around 27 seconds. RASL should be much slower on larger datasets due to the higher complexity of SVD, although we did not test.

Table 5.1. The mean/max registration errors in pixels of RASL, t-GRASTA and our method on the four lighting datasets. The first image is fixed to evaluate the errors.

	RASL	t-GRASTA	Proposed
NUTS	0.670/2.443	1.153/3.842	**0.061/0.488**
MOVI	0.029/ 0.097	0.568/ 2.965	**0.007/0.024**
FRUITS	0.050/0.107	1.094/4.495	**0.031/0.076**
TOY	0.105/ 0.373	0.405/2.395	**0.038/0.076**

We evaluate these three methods on the Multi-pie face database [144]. This database contains 20 images of each subject captured at different illumination con-

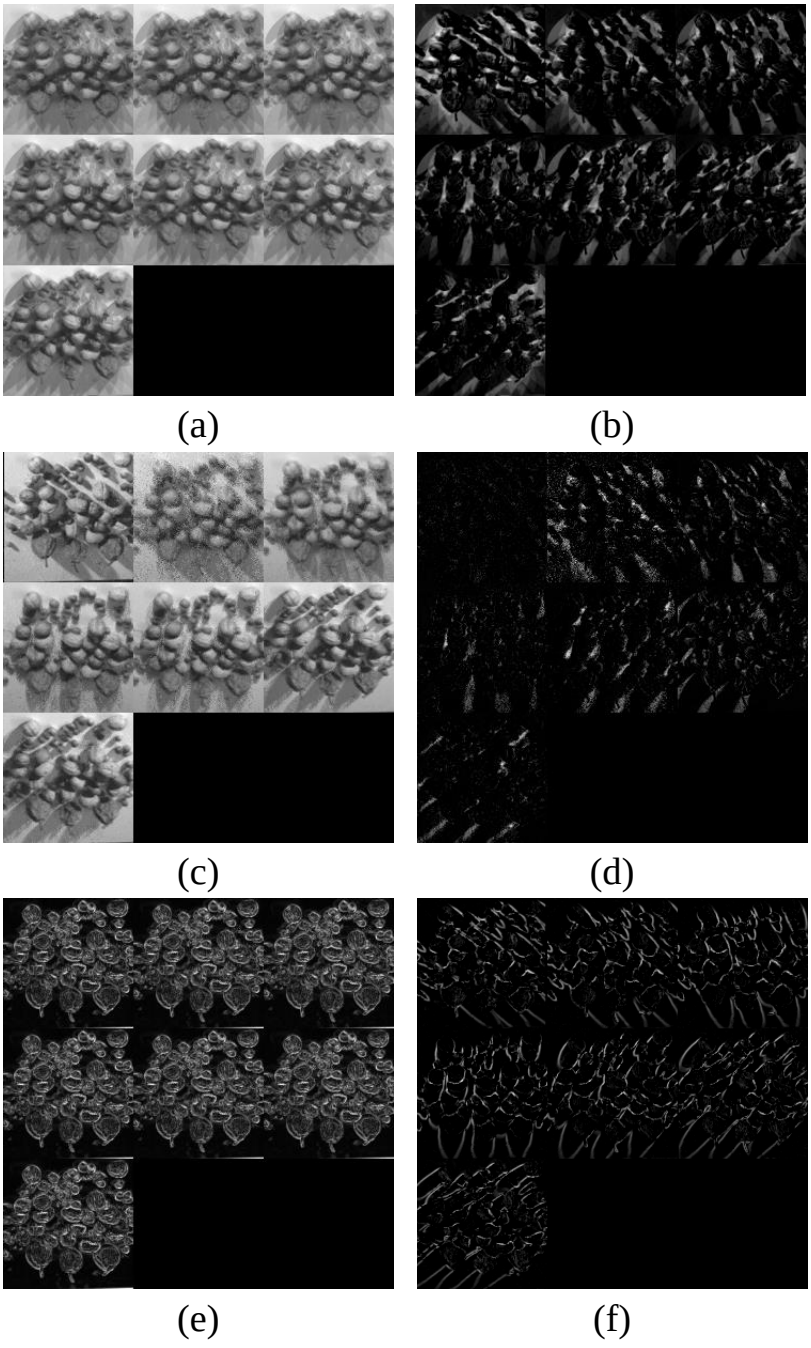


Figure 5.3. Batch image registration on the NUTS datasets. (a) The low rank component by RASL. (b) The sparse errors by RASL. (c) The subspace representation by t-GRASTA. (d) The sparse errors by t-GRASTA. (e) The visualization of $\boldsymbol{\mathcal{A}}$ by our method. (f) The sparse error $\boldsymbol{\mathcal{E}}$ by our method.

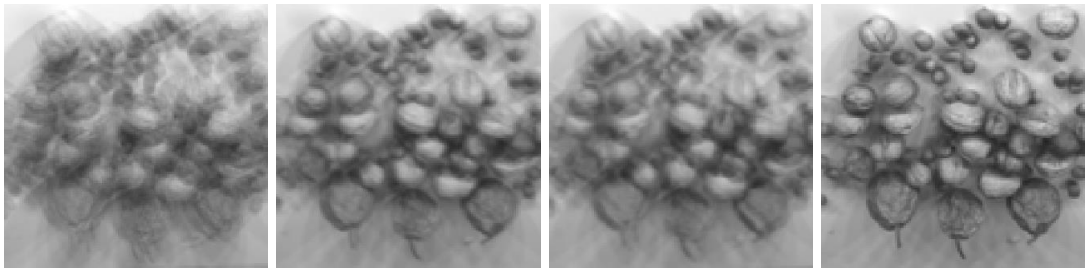


Figure 5.4. Registration results on the "NUTS" dataset. (a) The average image of perturbed images. (b) The average image by RASL. (c) The average image by t-GRASTA. (d) The average image by our method.

ditions. We add random artificial rotations in a range of 10° and translations in 10 pixels on the first 100 subjects from the Session 1. As the optimal alignment is not unique (e.g, all images shift by 1 pixel), we compare the standard derivation (STD) of the transformations after registration. Ideally, the STD should be zero when all the perturbations have been exactly removed. Fig. 5.5 shows the average registration results over 20 runs for each subject. Our method is more accurate than RASL and t-GRASTA for almost every subject.

5.4.2 Pair image registration

5.4.2.1 Simulations

We conduct a simulation on a brain MRI image [145]. The source image is warped by a non-rigid transformation, perturbed from random zero-mean Gaussians with three pixels standard deviation [146]. We add a few Gaussian intensity fields to simulate the distortion and rescale the images to [0,1]. Fig. 5.6 shows the input images and the results by RC [7] and the proposed method. SSD is not compared in non-rigid registration, as it always failed although different settings were tried. As we could see, both results are very close to the ground truth. A visible artifact can be

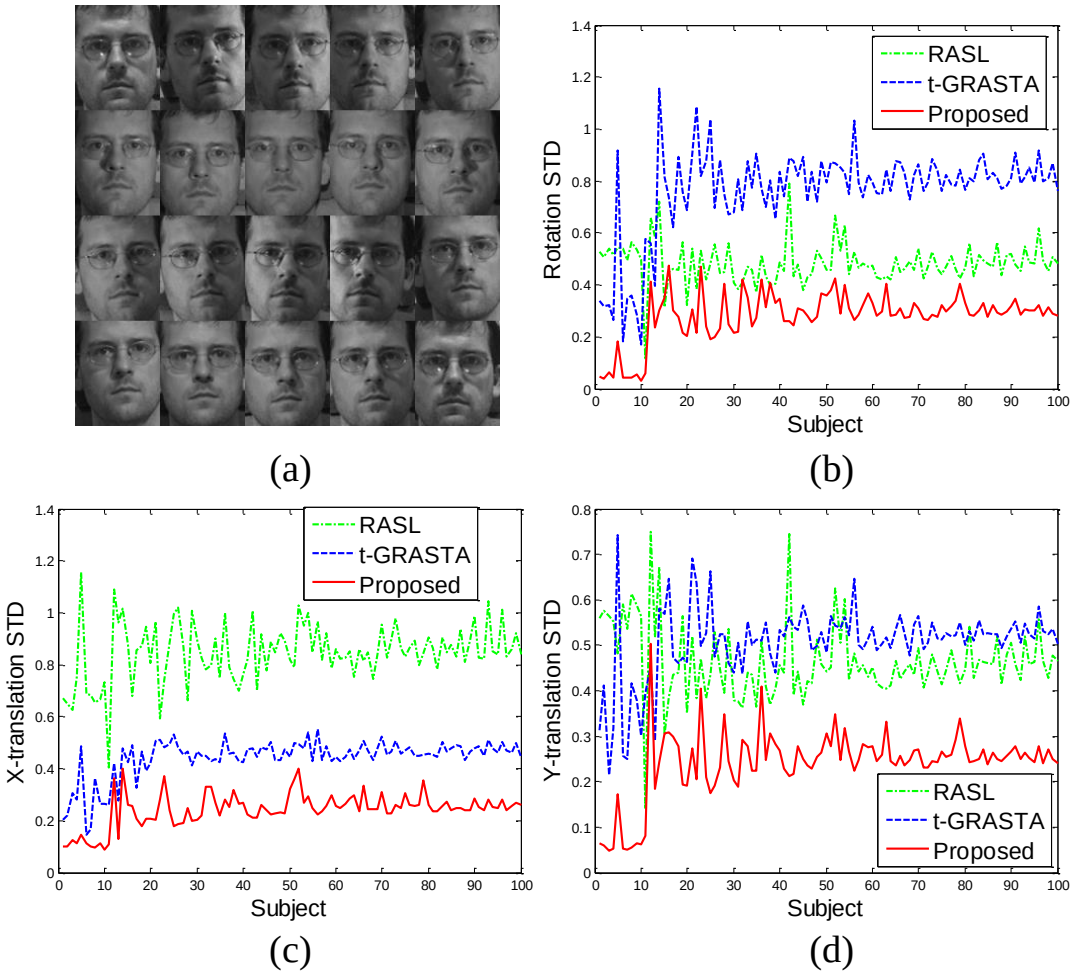


Figure 5.5. (a) An example input of the multipie image database. (b) The STD (in degrees) of rotations after registration. (c) The STD (in pixels) of X-translation after registration. (d) The STD (in pixels) of Y-translation after registration .

observed in the image recovered by RC, which is highlighted by the blue circle. The estimated transformation by our method is more smooth, and closer to the Gaussion perturbations.

For quantitative comparisons, we evaluate SSD, RC and the proposed method with random intensity distortions and random transformations like Fig. 5.6. The reference image without intensity distortions is used as ground-truth. Non-rigid transformations are applied on the brain image in Fig. 5.6, and random affine transformations

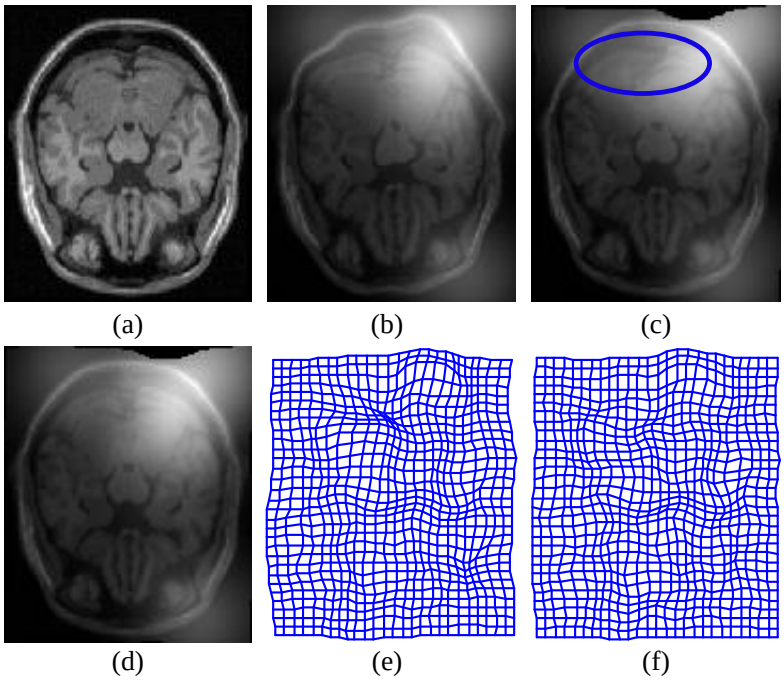


Figure 5.6. Synthetic experiment with non-rigid transformation. (a) The reference image. (b) The source image with intensity distortion. (c) Registration result by RC. (d) Registration by our method. (e) The transformation estimated by RC. (f) The transformation estimated by our method. Best viewed in $\times 2$ sized color pdf file.

are applied on the Lena image in Fig. 5.2 (with a similar range as the previous settings). The root-mean-square error (RMSE) is used as the metric for error evaluation of both image intensities and transformations.

The number of Gaussian intensity fields K is from 1 to 6. We run each setting 50 times and the results are plotted in Fig. 5.7. It can be observed that the proposed method is consistently better than SSD and RC, for both types of transformation. The registration speed of our method is often faster than that of RC. The average speed for the pair mode is 6.5 seconds per registration on the brain image (216×180) while that of RC is 13.7 seconds per registration.

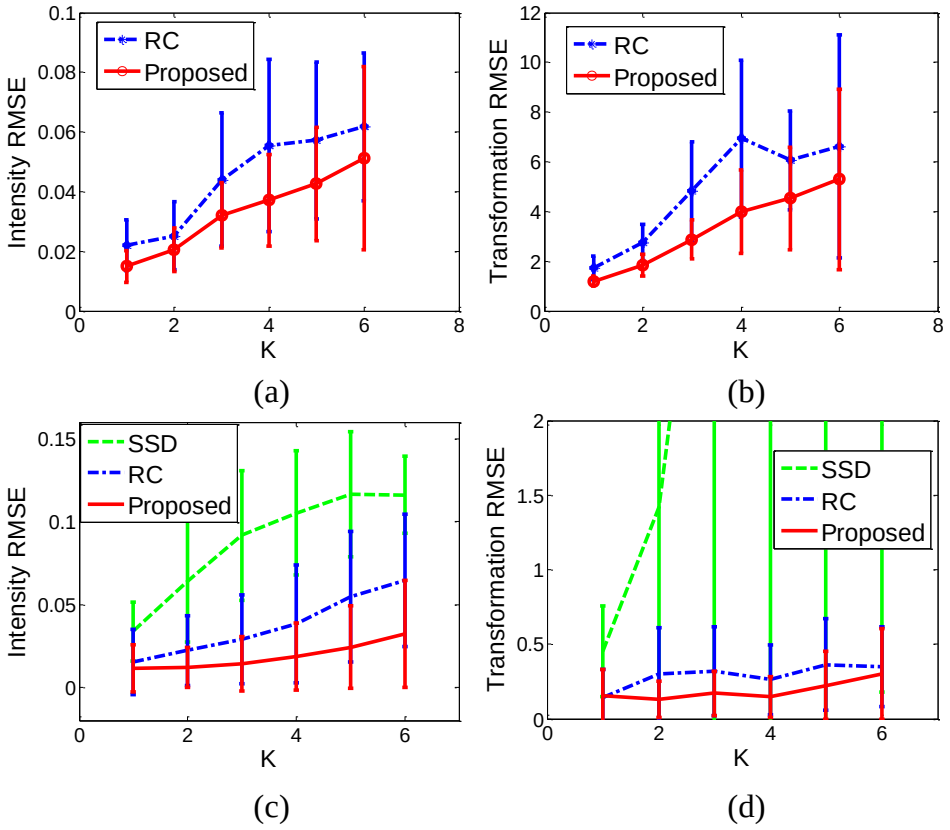


Figure 5.7. Registration performance comparisons with random transformation perturbations and random intensity distortions. (a) Intensity RMSE on the brain image. (b) Transformation (non-rigid) RMSE on the brain image. (c) Intensity RMSE on the Lena image. (d) Transformation (affine) RMSE on the Lena image.

5.4.2.2 Multisensor remotely sensed image registration

Multisensor image registration is a key preprocessing operation in remote sensing, e.g., for image fusion, change detection. The same land objects may be acquired at different times, under various illumination conditions by different sensors. Therefore, it is very possible that the input images have significant dissimilarity in terms of intensity values. Here, we register a panchromatic image to a multispectral image acquired by IKONOS multispectral imaging satellite [147], which have been pre-registered at their capture resolutions. The multispectral image has four bands:

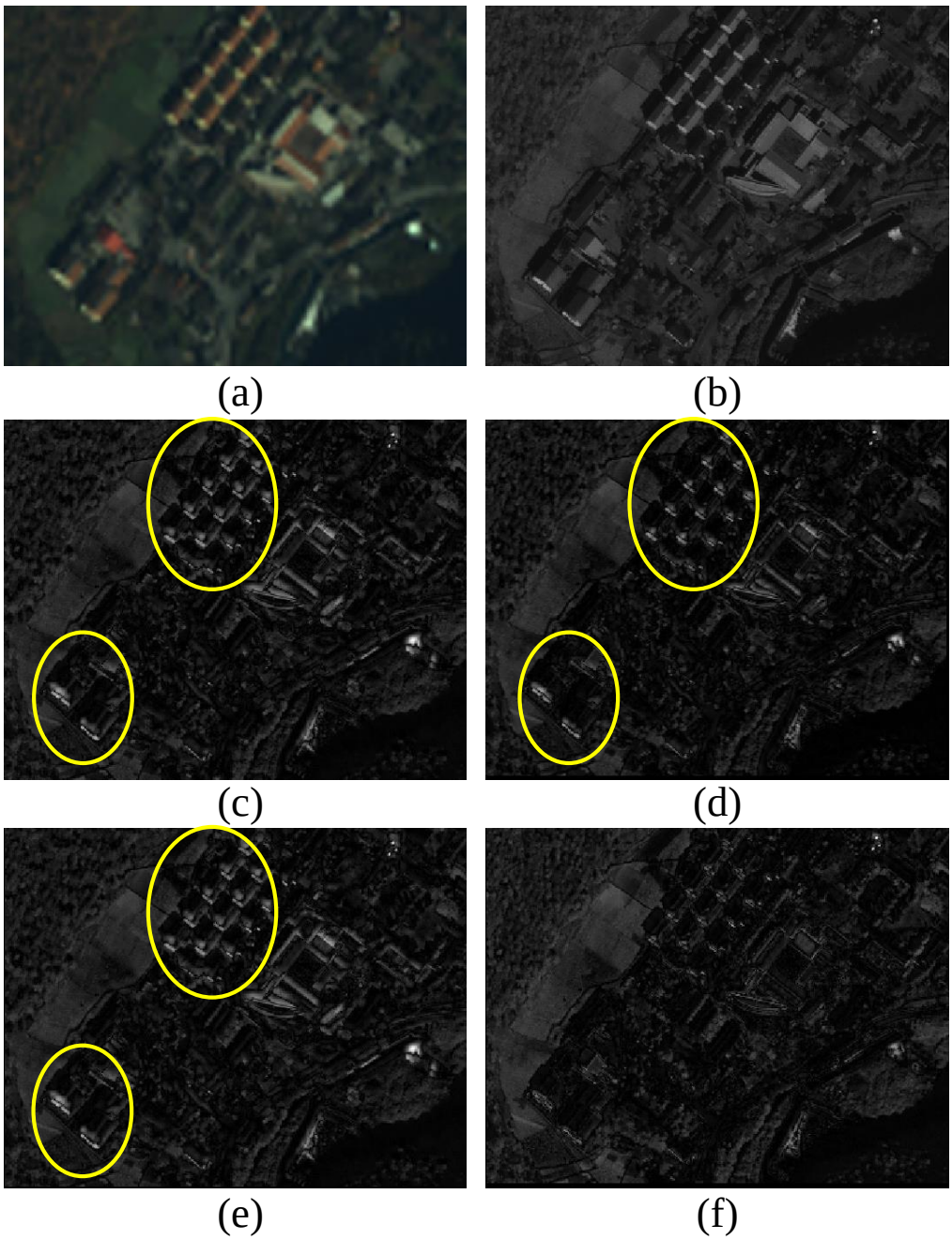


Figure 5.8. Registration of a multispectral image and a panchromatic image. (a) Reference image. (b) Source image. (c) The difference image before registration. (d) The difference image by SSD. (e) The difference image by RC. (f) The difference image by our method. Visible misalignments are highlighted by the yellow circles. Best viewed in $\times 2$ sized color pdf file.

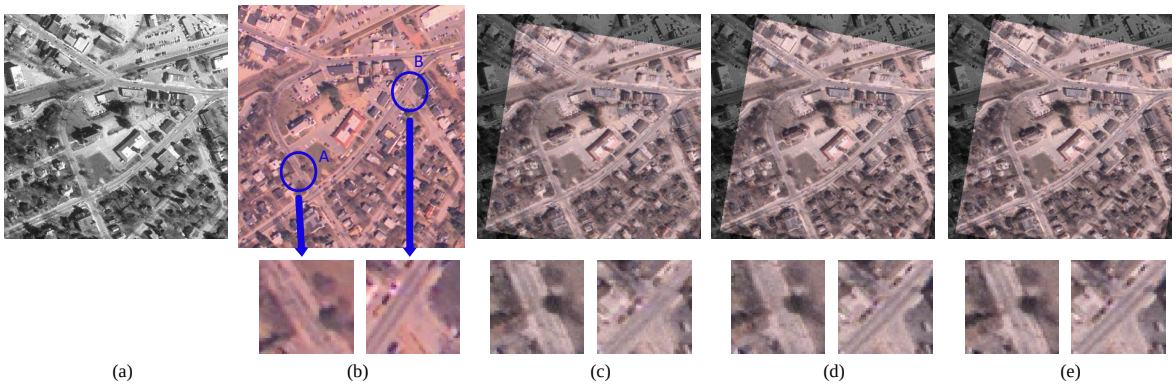


Figure 5.9. Registration of an aerial photograph and a digital orthophoto. From left to right, the images are: the reference image, the source image, the overlay by MATLAB, the overlay by RC, the overlay by our method. The second row shows the zoomed-in areas of streets A and B. Best viewed in $\times 2$ sized color pdf file.

blue, green, red and near-infrared, with 4 meter resolution (Fig. 5.8 (a)). The Pan image has 1 meter resolution (Fig. 5.8 (b)). The different image resolutions make this problem more difficult. From the difference image in Fig. 5.8 (c), we can observe that there exists misalignment in the northwest direction.

We compare our method with SSD [6] and RC [7], and the results are shown in Fig. 5.8 (d)-(f). It is assumed that the true transformation is formed by pure translation. Although we do not have the ground-truth, from the difference image, it can be clearly observed that our method can reduce the misalignment. In contrast, SSD and RC are not able to find better alignments than the preregistration method.

We register an aerial photograph to a digital orthophoto. The reference image is the orthorectified MassGIS georegistered orthophoto [148]. The source image is a digital aerial photograph, which does not have any particular alignment or registration with respect to the earth. The input images and the results are shown in Fig. 5.9. MATLAB uses manually selected control points for registration, while RC and our registrations are automatic. At the first glance, all the methods obtain registration

with good quality. A closer look shows that our method has higher accuracy than the others. In the source image, two lanes can be obviously observed in streets A and B. After registration and composition, street B in the result by MATLAB and street A in the result by RC are blurry due to the misalignment. Our method is robust to the local mismatches of vehicles.

5.4.2.3 Multimodal medical image registration

We further validate the performance of different methods on real-world medical images. Temporal and multimodal registration are performed on two retina images taken two years apart [12]. The reference image and source image are shown in Fig. 5.10 (a) and (b). These retina images are quite difficult to register with intensities. In order to avoid local minimum, we use affine transformation for preregistration and the result is shown in 5.10 (c). From the overlay in 5.10 (d), we could observe that there still exist misalignments for the vessels at the bottom half of the overlay. A local error can be found in the result by RC, while our method can eliminate the misalignments.

We validate the performance of our method on real-world applications. The proposed method is compared with RC on two images from a iris video sequence [7] (shown in Fig. 5.11). The deformation between the source image and reference image is highly nonlinear. The intensity artifact in the source image makes this problem more challenging. The composition image without registration is shown in Fig. 5.11(c) using green and magenta colors. The vessels are blurry due to the misalignment. After registration, both RC and the proposed DTV provide accurate alignments on the vessels. However, the image registered by RC has been partially distorted due to the severe intensity variance.

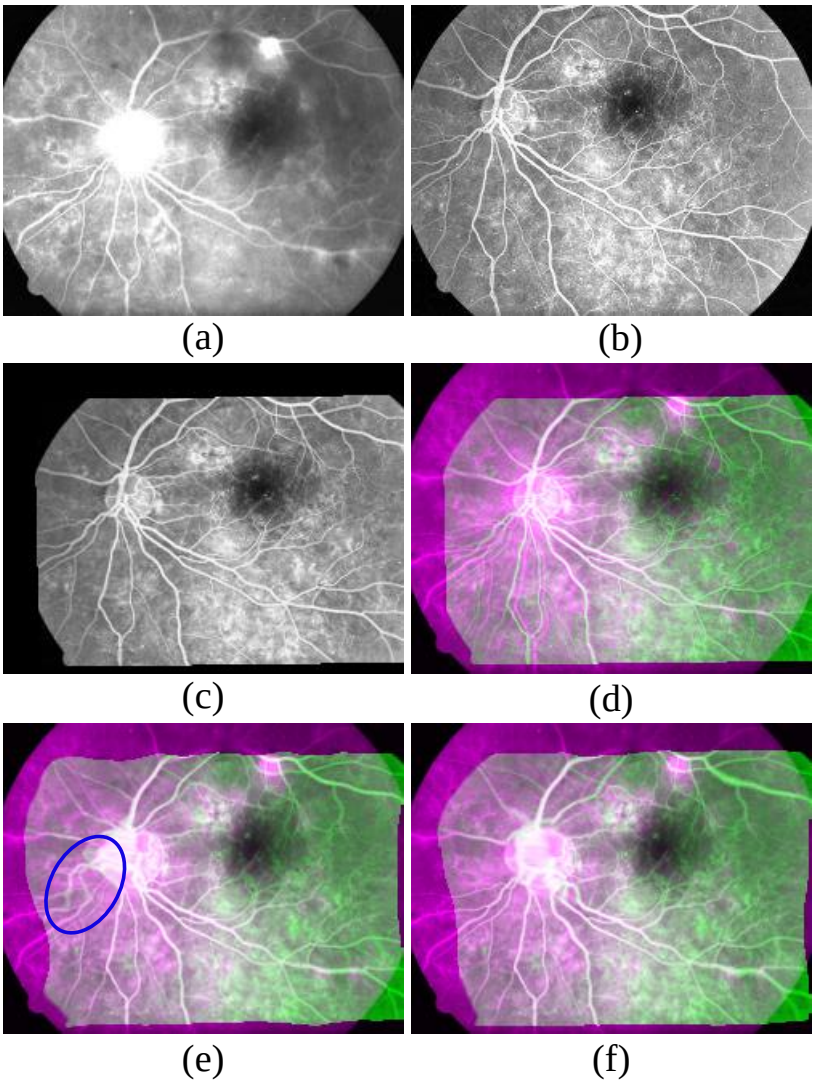


Figure 5.10. Registration of two retina images [12]. (a) Reference image. (b) Source image. (c) The source image after affine preregistration. (d) The overlay before registration. (e) The overlay after registration by RC. (f) The overlay after registration by our method. Visual artifact is highlighted by the blue circle.

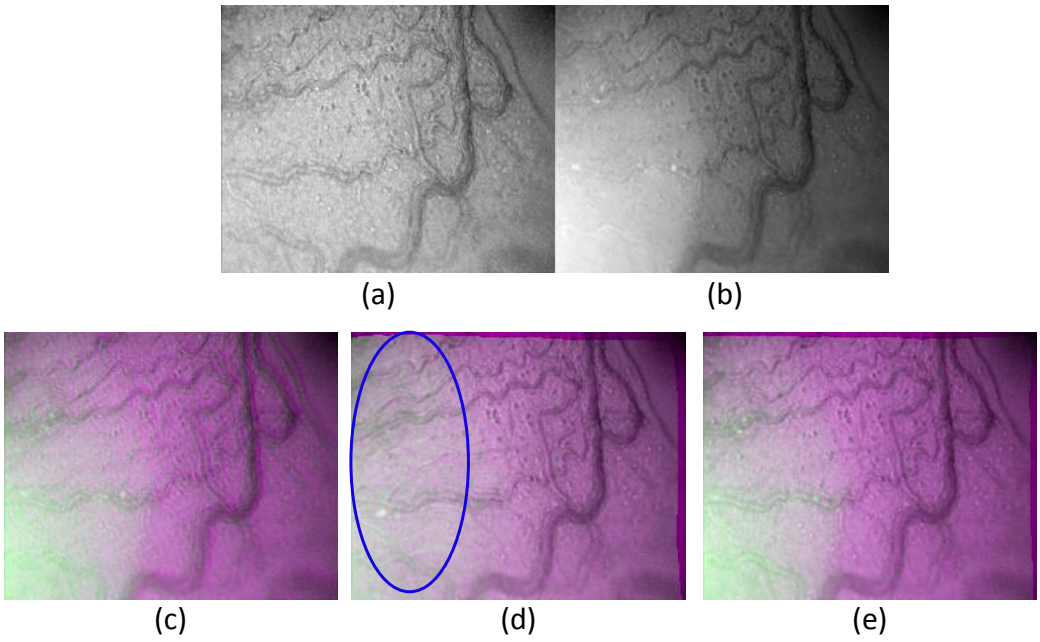


Figure 5.11. Registration of two iris images [7]. (a) Reference image. (b) Source image. (c) The overlay before registration. (d) The overlay after registration by RC. (e) The overlay after registration by our method. Visible artifact is highlighted by the blue circle. Best viewed in $\times 2$ sized color pdf file.

5.5 Summary

In this chapter, we have proposed a novel similarity measure for robust and accurate image registration. It is motivated by the deep sparse representation of the optimally registered images. The benefit of the proposed method is three fold: (1) compared with existing approaches, it can handle severe intensity distortions and partial occlusions simultaneously; (2) it can be used for registration of two images or a batch of images, with various types of transformations; (3) its low computational complexity makes it scalable to large datasets. We have conducted extensive experiments to test our method on multiple challenging datasets. The promising results demonstrate the robustness and accuracy of our method over the state-of-the-art batch registration methods and pair registration methods, respectively. We also show

that our method can be used to reduce the registration errors in many real-world applications.

Due to the local linearization in the optimization, our method as well as all the compared methods cannot handle large transformations. However, this is not a big issue for many real-world applications. For example, the remotely sensed images can be coarsely georegistered by their geographical coordinates. A good initialization also can be found on medical images, as the objects (e.g., brain) are easy to recognize and the image size is relatively smaller. For images with large transformations, we can use the FFT-based algorithm [132] to coarsely register the images and then apply our method as a refinement. Therefore, we did not test the maximum amount of transformations that our method can handle. So far, the proposed method can only be used for offline registration. How to extend this method to the online mode is an interesting topic of future research.

CHAPTER 6

Fast Iteratively Reweighted Least Squares Algorithms for Analysis-Based Sparsity Learning

In this chapter, we propose a novel algorithm for analysis-based sparsity reconstruction. It can solve the generalized problem by structured sparsity regularization with an orthogonal basis and total variation regularization. The proposed algorithm is based on the iterative reweighted least squares (IRLS) model, which is further accelerated by the preconditioned conjugate gradient method.This work was presented under a slightly modification from [158].

6.1 Introduction

Ill-posed problems widely exist in medical imaging and computer vision. In order to seek a meaningful solution, regularization is often used if we have certain prior knowledge. With the emerging of compressive sensing (CS) [37], sparsity regularization has been an active topic in recent years. If the original data is sparse or compressible, it can be recovered precisely from a small number of measurements. The ℓ_1 norm is usually used to induce sparsity and gains great success in many real applications. The optimization problems can be written as:

$$\min_x \{F(x) = \frac{1}{2}||Ax - b||_2^2 + \lambda||x||_1\}, \tag{6.1}$$

where $A \in \mathbb{R}^{M\times N}$ is the measurement matrix and $b \in \mathbb{R}^M$ is the vector of measurements; $x \in \mathbb{R}^N$ is the data to be recovered; λ is a positive parameter.

According to structured sparsity theories [29, 30], more benefits can be achieved if we could utilize more prior information about the sparsity patterns. For example,

the components of the data may be clustered in groups, which is called group sparse data. Components within the same group tend to be zeros or non-zeros. Sometimes one component may appear in several groups simultaneously, which corresponds to the overlapping group sparsity [52]. A favorable method would be replacing the ℓ_1 norm with $\ell_{2,1}$ norm to model the group sparsity [81]:

$$||x||_{2,1} = \sum ||x_{g_i}||_2, \quad i = 1, 2, ..., s, \tag{6.2}$$

where x_{g_i} denotes the components in i-th group and s is the total number of groups. It has been proven that, less measurements are required for structured sparsity recovery, or more precise solution can be obtained with the same number of measurements [29, 30, 149].

In many real-world applications, the data itself is not sparse, but it can be sparsely represented in some transformation domains. This leads to the analysis-based sparsity regularization problem:

$$\min_x \{F(x) = \frac{1}{2}||Ax - b||_2^2 + \lambda||\Psi x||_{2,1}\}, \tag{6.3}$$

where Ψ denotes some sparifying basis, e.g. the wavelet or finite difference basis. In this article, we are interested in this generalized sparsity regularization problem (6.3), which may contain overlapped groups. The standard sparsity and non-overlapping group sparsity minimization problem are special cases of problem (6.3). In our work, we focus on the image reconstruction applications, e.g. CS imaging [150], image inpainting [151], compressive sensing magnetic resonance imaging (CS-MRI) [1], where A is an undersampling matrix/operator.

When Ψ is an orthogonal basis, problem (6.3) corresponds to the Lasso problem [63]. In the literature, many efficient algorithms can be used to solve the standard Lasso and non-overlapping group Lasso, such as FISTA [42], SPGL1 [40], SpaRSA

[56], FOCUSS [17]. However, there are relatively much fewer algorithms for overlapping group Lasso, due to the non-smoothness and non-separableness of the overlapped $\ell_{2,1}$ penalty. SLEP [15, 152], GLO-pridu [153] solve the overlapping group sparsity problem by identifying active groups, and YALL1 [16] solves it with the alternating direction method (ADM). Both SLEP and GLO-pridu are based on the proximal gradient descent method (e.g. FISTA [42]), which cannot achieve a convergence rate better than $F(x^k) - F(x^*) \sim \mathcal{O}(1/k^2)$, where x^* denotes the optimal solution and k is the iteration number. YALL1 relaxes the original problem with augmented Lagrangian and iteratively minimizes the subproblems based on the variable splitting method. Generally, the convergence rate of ADM is no better than $\mathcal{O}(1/k)$ in sparse recovery problems. Although they are very efficient in each iteration, a large number of iterations may be required due to the relatively slow convergence rate. On the other hand, the iterative reweighted least squares (IRLS) algorithms have been proven that they converge exponentially fast [154] [155]. Unfortunately, conventional IRLS algorithms contain a large scale inverse operation in each step, which makes them still much more computationally expensive than the fastest proximal methods such as FISTA [149]. In addition, it is unknown how to extend these IRLS based algorithms to solve the overlapping group Lasso problems.

Another special case of (6.3) is the total variation (TV) reconstruction problem, where Ψ denotes the first-order finite difference matrices and is non-orthogonal. There are several efficient algorithms specially designed for TV reconstruction, including the RecPF [3] and SALSA [156]. Both of them are relaxed by the ADM. The efficient transformation in RecPF [3] requires that A^TA can be diagonalized by the Fourier transform, while SALSA [156] requires $AA^T = I$. Due to these restrictions, these two methods can not be applied to certain reconstruction applications, e.g. CS imaging [150]. Moreover, it is unknown how to extend them to solve the joint to-

tal variation (JTV) problems. Moreover, the ADM-based methods often have slower convergence rate. Of course, generalized minimization methods can be used, such as the split Bregman method [114], FISTA [39] and IRN [157], but they have their own inferiority without considering the special structure of undersampling matrix A in reconstruction.

In this article, we propose a novel scheme for the analysis-based sparsity reconstruction (6.3) based on the IRLS framework. It preserves the fast convergence performance of traditional IRLS, which only requires a few reweighted iterations to achieve an accurate solution. We call our method fast iterative reweighted least squares (FIRLS). Moreover, we propose a new "pseudo-diagonal" type preconditioner to significantly accelerate the inverse subproblem with preconditioned conjugate gradient (PCG) method. This preconditioner is based on the observation that A^TA is often diagonally dominant in the image reconstruction problems. With the same computation complexity, the proposed preconditioner provides more precise results than conventional Jacobi diagonal preconditioner. In addition, the proposed preconditioner can be applied even when A is an operator, e.g., the Fourier or wavelet transform, which is not feasible for most existing preconditioners of the PCG methods. Besides the efficiency and fast convergence rate, the proposed method can be easily applied to different sparsity patterns, e.g. overlapping group sparsity, TV and JTV. We validate the proposed method on CS-MRI for tree sparsity, joint sparsity, TV and JTV based reconstruction. Extensive experimental results demonstrate that the proposed algorithm outperforms the state-of-the-art methods in terms of both accuracy and computational speed. Part of results in this work has been presented in [158].

6.2 Related Work: IRLS

The conventional IRLS algorithms solve the standard sparse problem in this constrained form:

$$\min_x ||x||_1, \text{subject to } Ax = b. \tag{6.4}$$

In practice, the ℓ_1 norm is replaced by a reweighted ℓ_2 norm [155]:

$$\min_x x^T W x, \text{subject to } Ax = b. \tag{6.5}$$

The diagonal weight matrix W in the k-th iteration is computed from the solution of the current iteration x^k, in particular, the diagonal elements $W_i^k = |x_i^k|^{-1}$. With current weights W^k, we can derive the closed form solution for x^{k+1}:

$$x^{k+1} = (W^k)^{-1} A^T (A(W^k)^{-1} A^T)^{-1} b. \tag{6.6}$$

The algorithm can be summarized in Algorithm 9. It has been proven that the IRLS algorithm converges exponentially fast under mild conditions [154]:

$$||x^k - x^*||_1 \leq \mu ||x^{k-1} - x^*||_1 \leq \mu^k ||x^0 - x^*||_1, \tag{6.7}$$

where μ is a fixed constant with $\mu < 1$. However, this algorithm is rarely used in compressive sensing applications especially for large scale problems. That is because the inverse of $A(W^k)^{-1}A^T$ takes $\mathcal{O}(M^3)$ if A is a $M \times N$ sampling matrix. Even with higher convergence rate, traditional IRLS still cannot compete with the fastest first-order algorithms such as FISTA [42] (some results have been shown in [149]). Moreover, none of existing IRLS methods [159, 154, 17] could solve the overlapping group sparsity problems, which significantly limits the usage.

Algorithm 9 IRLS

Input: A,b,x^1,$k = 1$

while not meet the stopping criterion **do**

Update W: $W_i^k = |x_i^k|^{-1} \quad \forall\ W_i^k$

Update x: $x^{k+1} = (W^k)^{-1}A^T(A(W^k)^{-1}A^T)^{-1}b$

Update $k = k + 1$

end while

6.3 FIRLS for overlapping group sparsity

6.3.1 An alternative formulation for overlapping group sparsity

We consider the overlapping group Lasso problem first [81, 52]. The mixed $\ell_{2,1}$ norm in (6.3) may contain overlapping groups. It can be rewritten in the analysis-based sparsity form:

$$\min_x \{F(x) = \frac{1}{2}||Ax - b||_2^2 + \lambda||G\Phi x||_{2,1}\}, \tag{6.8}$$

where Φ denotes an orthogonal sparse basis and is optional. A good choice of Φ for natural images/signals would be the orthogonal wavelet transform. G is a binary matrix for group configuration, which is constructed by the rows of the identity matrix. With different settings of G, this model can handle overlapping group, non-overlaping group and standard sparsity problems. Tree sparsity can also be approximated by this model [160, 161, 47]. Simple examples of G for different types of group sparse problems are shown in Fig. 6.1. Although G may have large scales, it can be efficiently implemented by a sparse matrix. This kind of indexing matrix has been used in the previous work YALL1 [16]. With this reformulation, $\Psi = G\Phi$ and the $\ell_{2,1}$ norm in (6.8) is now non-overlapping.

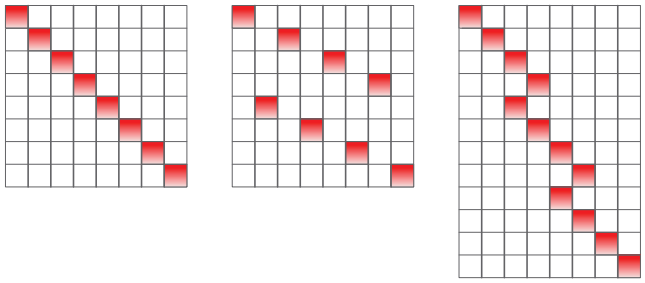


Figure 6.1. Examples of group configuration matrix G for a signal of size 8. The red elements denote ones and white elements denote zeros. (a) standard sparsity case where G is the identical matrix. (b) non-overlapping groups of [1,3,5,7] and [2,4,6,8]. (c) overlapping groups of [1,2,3,4], [3,4,5,6] and [5,6,7,8]. Their group sizes are 1,4 and 4, respectively.

Consider the Young's inequality holding for a general function $g(\cdot) : \mathbb{R} \longmapsto \mathbb{R}$:

$$\sqrt{g(x)} \leq \frac{\sqrt{g(y)}}{2} + \frac{g(x)}{2\sqrt{g(y)}}, \tag{6.9}$$

with $g(y) > 0$ and $g(x) \geq 0$. The equality holds only when $g(x) = g(y)$. Based on this, we have

$$||G\Phi x||_{2,1} = \sum_{i=1}^{s} ||(G\Phi x)_{g_i}||_2 \leq \sum_{i=1}^{s} \quad [\frac{||(G\Phi x^k)_{g_i}||_2}{2} + \frac{||(G\Phi x)_{g_i}||_2^2}{2||(G\Phi x^k)_{g_i}||_2}]. \tag{6.10}$$

Writing it in matrix form and we can majorize $F(x)$ by the majorization minimization (MM) method [162]:

$$Q(x, W^k) = \frac{1}{2}||Ax - b||_2^2 + \frac{\lambda}{2} x^T \Phi^T G^T W^k G \Phi x + \frac{\lambda}{2} \sum_{i=1}^{s} \frac{1}{W_{g_i}^k}, \tag{6.11}$$

where Φ^T denotes the inverse transform of Φ; W^k is the group-wise weights. The weight of i-th group $W_{g_i}^k$ can be obtained by:

$$W_{g_i}^k = (||(G\Phi x^k)_{g_i}||_2^2 + \epsilon)^{-1/2}. \tag{6.12}$$

ϵ is very small number (e.g. 10^{-10}) to avoid infinity. Suppose that the signal x to be recovered is of length N and G is a N'-by-N matrix, then W^k is a N'-by-N' diagonal matrix and has the following form:

$$W^k = \begin{Bmatrix} W^k_{g_1} & & & & & \\ & \ddots & & & & \\ & & W^k_{g_1} & & & \\ & & & \ddots & & \\ & & & & W^k_{g_s} & \\ & & & & & W^k_{g_s} \end{Bmatrix}, \tag{6.13}$$

where each group-wise weight $W^k_{g_i}$ is duplicated $|g_i|$ times and $|g_i|$ denotes the size of the i-th group. One can find that the group-wise weights are all related to G. With different settings of G, the group-wise weights are directly derived. Variant-size group sparsity problems also can be flexibly handled in this model. An interesting case would be the standard sparse problem, where each group contains only one element and the group-wise weight matrix W is the same as that in IRLS algorithm [154, 159].

Now the problem becomes:

$$x^{k+1} = \arg\min_x Q(x, W^k). \tag{6.14}$$

Note that $W^k_{g_i}$ is independent of x and can be considered as a constant. We iteratively update W^k with x^k and solve x^{k+1} based on current W^k. Our algorithm is also a IRLS type algorithm with exponentially fast convergence rate.

6.3.2 Accelerating with PCG

In each iteration, W^k can be easily updated with (6.13) and (6.12). To solve (6.14), a simple way is to let the first order derivative of $Q(x|x^k)$ be zero as it is a quadratic convex function:

$$(A^TA + \lambda\Phi^TG^TW^kG\Phi)x - A^Tb = 0. \tag{6.15}$$

The way to solve (6.15) determines the efficiency of the whole algorithm. The exact inverse of the system matrix $S = A^TA + \lambda\Phi^TG^TW^kG\Phi$ takes $\mathcal{O}(N^3)$ time. It is impractical to compute S^{-1} for many cases especially when the size of S is large. An alternative way is to approximate the solution of (6.15) with classical conjugate gradient (CG) decent method. It is much faster than computing the exact solution. Besides CG, a better way is the preconditioned conjugate gradient (PCG) method [163]. The design of preconditioner is problem-dependent, which should be as close as possible to the system matrix S and can be inversed efficiently. Therefore, it is not an easy task to design a good preconditioner in general due to this tradeoff. For signal/image reconstruction, such preconditioner has not been found in existing IRLS algorithms [159, 154, 17].

By observing that S is usually diagonally dominant in reconstruction problems, e.g. CS imaging, image inpainting and CS-MRI, we define a new preconditioner for best approximation in Frobenius norm $||\cdot||_F$:

$$P = \arg\min_{X\in\mathcal{D}} ||S - X||_F, \tag{6.16}$$

where $\mathcal{D}$ denotes the class of diagonal or "pseudo-diagonal" matrices. Here, the pseudo-diagonal matrix means a non-diagonal matrix whose inverse can be obtained efficiently like a diagonal matrix with $\mathcal{O}(N)$ time. Please note that the G^TW^kG is always diagonal for any kind of G in Fig. 6.1. Due to the strong constraint, the

possible diagonal or "pseudo-diagonal" candidates for (6.16) are enumerable. In addition, we observe that A^TA is often diagonally dominant in the image reconstruction problems. For example, in CS-MRI, $A = RF$ where F denotes the Fourier transform and $R \in \mathbb{R}^{M\times N}$ $(M < N)$ is a selection matrix containing M rows of the identity matrix. Therefore, $A^TA = F^TR^TRF$ is diagonally dominant as R^TR is diagonal. For the image inpainting problem, $A^TA = R^TR$ is diagonal. This structure also holds when A is a random projection matrix.

Based on this diagonally dominant effect, it is not hard to find an accurate solution

$$P = (\overline{A^TA}I + \lambda\Phi^TG^TW^kG\Phi), \tag{6.17}$$

where $\overline{A^TA}$ is the mean of diagonal elements of A^TA and I denotes the identity matrix. The preconditioning error in Frobenius norm $||S - P||_F$ is very small, due to diagonally dominant structure of A^TA.

As A is known for the application, $\overline{A^TA}$ can be pre-estimated before the first iteration and is fixed for each iteration. Therefore in each iteration, $P^{-1} = \Phi^T(\overline{A^TA}I + \lambda G^TW^kG)^{-1}\Phi$ can be obtained with linear time.

Several advantages of the proposed preconditioner can be found when compared with existing ones [164, 165]. To get the inverse, fast Fourier transforms are used in recent circulant preconditioners for image deblurring [164] [165], while our model only requires linear time to obtain P^{-1}. Compared with conventional Jacobi preconditioner, we do not discard all non-diagonal information and therefore the preconditioner is more accurate. Moreover, our model can also handle the case when A or Φ is an operator, while other preconditioners [164, 165, 157] cannot because they require the exact values of S. Interestingly, the conventional Jacobi preconditioner

can be derived by (6.16), when the original data is sparse (i.e. $\Phi = 1$) and A is a numerical matrix.

Algorithm 10 FIRLS_OG

Input: A,b,x^1, G, λ, $k = 1$

while not meet the stopping criterion **do**

Update W^k by (6.38) (6.12)

Update $S = A^T A + \lambda \Phi^T G^T W^k G \Phi$

Update $P = \Phi^T(\overline{A^T A} I + \lambda G^T W^k G)\Phi$,

$$P^{-1} = \Phi^T(\overline{A^T A} I + \lambda G^T W^k G)^{-1}\Phi$$

while not meet the PCG stopping criterion **do**

Update x^{k+1} by PCG for $Sx = A^T b$ with preconditioner P

end while

Update $k = k + 1$

end while

Our method can be summarized in Algorithm 10. We denote this overlapping group sparsity version as FIRLS_OG. Although our algorithm has double loops, we observe that only 10 to 30 PCG iterations are sufficient to obtain a solution very close to the optimal one for the problem (6.15). In each inner PCG iteration, the dominated cost is by applying S and P^{-1}, which is denoted by $\mathcal{O}(\mathcal{C}_S + \mathcal{C}_P)$. When A and Φ are dense matrices, $\mathcal{O}(\mathcal{C}_S + \mathcal{C}_P) = \mathcal{O}(N^2)$. When A and Φ are the partial Fourier transform and wavelet transform in CS-MRI [1], it is $\mathcal{O}(N \log N)$.

6.3.3 Convergence analysis

Theorem 1. *The global optimal solution x^* of (6.11) is the global optimal solution of original problem (6.8).*

Proof. Suppose x_1^* is the global optimal solution of (6.11) and x_2^* is the global optimal solution of (6.8). Consider Q as a function corresponds to x and W. We have:

$$Q(x_1^*, W_1^*) \leq Q(x_2^*, W), \quad \forall\ W; \tag{6.18}$$

$$F(x_2^*) \leq F(x_1^*). \tag{6.19}$$

Based on the inequalities (6.9)(6.10), we have

$$F(x) \leq Q(x, W^k) \quad \forall\ x; \tag{6.20}$$

$$F(x^k) = Q(x^k, W^k). \tag{6.21}$$

Therefore,

$$F(x_2^*) \leq F(x_1^*) = Q(x_1^*, W_1^*) \leq Q(x_2^*, W_2^*) = F(x_2^*), \tag{6.22}$$

which indicates $F(x_1^*) = F(x_2^*)$. Here W_1^* and W_1^* are weights of x_1^*, x_2^* based on (6.12) and (6.13).

Theorem 2. *$F(x^k)$ is monotonically decreased by Algorithm 2, i.e. $F(x^{k+1}) \leq F(x^k)$. In particular, we have $\lim_{k\to\infty}(F(x^k) - F(x^{k+1})) = 0$.*

Proof. With the property (6.20), we have

$$F(x^{k+1}) \leq Q(x^{k+1}, W^k). \tag{6.23}$$

To balance the cost and accuracy when solving (15), we apply the PCG method to decrease $Q(x, W^k)$ and efficiently obtain the solution x^{k+1}. Because x^k is the initial guess for x^{k+1}, based on the property of PCG we have:

$$Q(x^{k+1}, W^k) \leq Q(x^k, W^k). \tag{6.24}$$

And we finally get:

$$F(x^{k+1}) \leq Q(x^{k+1}, W^k) \leq Q(x^k, W^k) = F(x^k). \tag{6.25}$$

$F(x)$ is convex and bounded. Due to the monotone convergence theorem, we have:

$$\lim_{k\to\infty} (F(x^k) - F(x^{k+1})) = 0. \tag{6.26}$$

Theorem 3. *Any accumulation point of* $\{x^k\}$ *is a stationary point of problem (6.11).*

Proof. When we have any accumulation point $x^k = x^{k+1}$ for $k \to \infty$, it demonstrates the inner PCG loop has converged for problem (6.15). Therefore, it indicates

$$(A^T A + \lambda \Phi^T G^T W^k G \Phi) x^k - A^T b = 0. \tag{6.27}$$

Consider Q as a function corresponding to x and W. We have

$$\frac{\partial Q(x^k, W^k)}{\partial x} = 0. \tag{6.28}$$

In addition,

$$\begin{aligned}
&\frac{\partial Q(x^k, W^k)}{\partial W} \\
=&\partial\{\frac{\lambda}{2}(x^k)^T \Phi^T G^T W G \Phi x^k + \frac{\lambda}{2}\sum_{i=1}^{s} \frac{1}{W_{g_i}^k}\}/\partial W \\
=&\frac{\lambda}{2}[(x^k)^T \Phi^T G^T G \Phi x^k - \sum_{i=1}^{s} (W_{g_i}^k)^{-2}].
\end{aligned} \tag{6.29}$$

Based on (6.12) and (6.13), it can be rewritten as:

$$\frac{\lambda}{2}\sum_{i=1}^{s}[||(G\Phi x^k)_{g_i}||_2^2 - (||(G\Phi x^k)_{g_i}||_2^2 + \epsilon)] \qquad = \lambda s\epsilon/2. \tag{6.30}$$

Note that ϵ is negligible and we finally have

$$\frac{\partial Q(x^k, W^k)}{\partial x} = \frac{\partial Q(x^k, W^k)}{\partial W} \approx 0. \tag{6.31}$$

Hence x^k is a stationary point of (6.11) when $k \to \infty$.

It indicates that the algorithm converges to a local minimum of the problem. In our experiments, if we let the initial guess $x^0 = A^T b$, an accurate solution can be always obtained. Note that Theorems 1, 2 and 3 always hold no matter how many inner PCG iterations are used.

6.4 FIRLS for Total Variation

We have presented an efficient algorithm for overlapping group sparsity under an orthogonal sparse basis Φ. In image reconstruction problems, another widely used sparsity regularizer is the TV. Due to the non-orthogonality of the TV semi-norm, the FIRLS_OG algorithm can not be applied to solve the TV problem. In this section, we will present an efficient algorithm for TV based image reconstruction. For brevity and clarity, we first present the algorithm for single channel image reconstruction and then extended it to multi-channel reconstruction [115].

6.4.1 An alternative formulation for total variation

TV minimization exploits the sparsity of the image in the gradient domain. For brevity, we assume the image is n by n with $n \times n = N$. Let D_1, D_2 be two N-by-N two first-order finite difference matrices in vertical and horizontal directions.

$$D_1 = \begin{bmatrix} 1 & & & & & & \\ -1 & 1 & & & & & \\ & -1 & 1 & & & & \\ & & \ldots & \ldots & & & \\ & & & \ldots & \ldots & & \\ & & & & -1 & 1 & \\ & & & & & -1 & 1 \end{bmatrix}, \tag{6.32}$$

$$D_2 = \begin{bmatrix} 1 & & & & & & \\ \ldots & \ldots & & & & & \\ -1 & \ldots & 1 & & & & \\ & -1 & \ldots & 1 & & & \\ & & \ldots & & \ldots & & \\ & & & -1 & \ldots & 1 & \\ & & & & -1 & \ldots & 1 \end{bmatrix}. \tag{6.33}$$

The main diagonal elements of D_1 and D_2 are all ones. The first diagonal elements below the main diagonal are all minus ones in D_1, while in D_2 n-th diagonal elements below the main diagonal are all minus ones. With these notations, the ℓ_1 and isotropic TV based image reconstruction can be reformulated as:

$$\min_x \{\frac{1}{2}||Ax - b||_2^2 + \lambda||D_1 x||_1 + \lambda||D_2 x||_1\}, \tag{6.34}$$

and

$$\min_x \{F(x) = \frac{1}{2}||Ax - b||_2^2 + \lambda|||[D_1 x, D_2 x]|||_{2,1}\}. \tag{6.35}$$

Here, the $\ell_{2,1}$ norm is the summation of the ℓ_2 norm of each row, which is a special case of (6.2). Here and later, we denote [,] as the concatenating of the matrices horizontally. To avoid repetition, all the following derivations only consider isotropic TV function (6.35). ℓ_1-based TV function can be derived in the same way.

Considering the Young's inequality in (6.9), we majorize (6.35) by the MM method [162]:

$$\begin{aligned} Q(x, W^k) = \frac{1}{2}||Ax - b||_2^2 + \frac{\lambda}{2} x^T D_1^T W^k D_1 x \\ + \frac{\lambda}{2} x^T D_2^T W^k D_2 x + \frac{\lambda}{2} \mathrm{Tr}((W^k)^{-1}), \end{aligned} \tag{6.36}$$

where Tr() denotes the trace. W^k is a diagonal weight matrix in the k-th iteration:

$$W_i^k = 1/\sqrt{(\nabla_1 x_i^k)^2 + (\nabla_2 x_i^k)^2 + \epsilon}, \ i = 1, 2, ..., N, \tag{6.37}$$

and

$$W^k = \begin{Bmatrix} W_1^k & & & \\ & W_2^k & & \\ & & ... & \\ & & & W_N^k \end{Bmatrix}. \tag{6.38}$$

When $D_1 = D_2 = I$, it is identical to the ℓ_1 norm minimization as in the conventional IRLS methods [159, 154, 17].

6.4.2 Accelerating with PCG and incomplete LU decomposition

After the weight matrix is updated by (6.37) and (6.38), the problem is to update x. With the same rule as that in the overlapping group sparsity regularization, we have

$$(A^TA + \lambda D_1^T W^k D_1 + \lambda D_2^T W^k D_2)x = A^T b. \tag{6.39}$$

Similar to (6.15), the system matrix here is in large scale. We have discussed that the system matrix is not dense but follows some special structure in image reconstruction. A good solver should consider such special structure of the problem. In TV based image deblurring problems, by observing that A has a circulant structure (under periodic boundary conditions), many efficient algorithms have been proposed to accelerate the minimization [165, 166, 167]. However, these algorithms can not be applied to the TV reconstruction problems.

Based on the diagonally dominant prior information in image reconstruction, we can obtained an accurate preconditioner like (6.17).

$$P = \overline{A^TA}I + \lambda D_1^T W^k D_1 + \lambda D_2^T W^k D_2 \tag{6.40}$$

However, the inverse can not be efficiently obtained for this preconditioner, due to the non-orthogonality of D_1 and D_2.

Fortunately, P has a sparse structure due to the forms of D_1 and D_2. We observe that preconditioner P in (6.40) is a penta-diagonal matrix [163] and is symmetric, which has the following form:

$$P = \begin{bmatrix} a_1 & b_1 & & c_1 & & & \\ b_1 & a_2 & b_2 & & c_2 & & \\ & b_2 & a_3 & \dots & & \dots & \\ & & \dots & \dots & & & c_{N-n} \\ c_1 & & & & & & \\ & c_2 & & & & \dots & \\ & & \dots & & \dots & \dots & b_{N-1} \\ & & & c_{N-n} & & b_{N-1} & a_N \end{bmatrix}. \tag{6.41}$$

Such penta-diagonal matrix has incomplete LU decomposition $P \approx LU$, where

$$L = \begin{bmatrix} 1 & & & & & & \\ \frac{b_1}{a_1} & 1 & & & & & \\ & \frac{b_2}{a_2} & 1 & & & & \\ & & ... & ... & & & \\ \frac{c_1}{a_1} & & & \frac{b_n}{a_n} & 1 & & \\ & ... & & & ... & ... & \\ & & \frac{c_{N-n}}{a_{N-n}} & ... & & \frac{b_{N-1}}{a_{N-1}} & 1 \end{bmatrix}, \tag{6.42}$$

$$U = \begin{bmatrix} a_1 & b_1 & ... & c_1 & & & \\ & a_2 & b_2 & ... & c_2 & & \\ & & ... & ... & & ... & \\ & & & a_{N-n} & b_{N-n} & ... & c_{N-n} \\ & & & & ... & ... & \\ & & & & & a_N & b_{N-1} \\ & & & & & & a_N \end{bmatrix}. \tag{6.43}$$

The decomposition is very accurate as P is diagonally dominated with $a_i \gg b_i^2, a_i \gg c_i^2$ and $a_i \gg b_i c_i$ for all i. To the best of our knowledge, this incomplete LU decomposition is first proposed for TV minimization. Due to the special structure of P, the incomplete LU decomposition only takes $\mathcal{O}(N)$ time. Therefore, the total time to obtain $P^{-1} \approx U^{-1}L^{-1}$ is $\mathcal{O}(N)$. We can conclude the proposed method for TV reconstruction in Algorithm 11. The convergence of this algorithm can be proven in a similar way as Section 6.3.3, which is omitted to avoid repetition.

Algorithm 11 FIRLS_TV

Input: A, b, x^1, λ, $k = 1$

while not meet the stopping criterion **do**

Update W^k by (6.37) and (6.38)

Update $S = A^TA + \lambda D_1^T W^k D_1 + \lambda D_2^T W^k D_2$

Update $P = \overline{A^TA}I + \lambda D_1^T W^k D_1 + \lambda D_2^T W^k D_2 \approx LU$, $P^{-1} \approx U^{-1}L^{-1}$

while not meet the PCG stopping criterion **do**

Update x^{k+1} by PCG for $Sx = A^Tb$ with preconditioner P

end while

Update $k = k + 1$

end while

6.4.3 Extension to JTV

In many multiple measurement vector problems (MMV), the image with multiple channels has the joint sparsity property. In these cases, the TV can be extended to joint total variation (JTV) [115, 78]:

$$\min_x \{\frac{1}{2}\sum_{t=1}^{T} ||A_tX_t - b_t||_2^2 + \lambda ||[D_1X, D_2X]||_{2,1}\}, \tag{6.44}$$

where $X \in \mathbb{R}^{N\times T}$ is a T-channel image with $X = [X_1, X_2, ..., X_T]$; A_t is the undersampling matrix for channel t and b_t is the measurement vector for channel t. Similar as (6.36), we have:

$$\begin{aligned} Q(X, W^k) = \frac{1}{2}\sum_{t=1}^{T} ||A_tX_t - b_t||_2^2 + \frac{\lambda}{2}[\sum_{t=1}^{T} X_t^T D_1^T W^k D_1 X_t \\ + \sum_{t=1}^{T} X_t^T D_2^T W^k D_2 X_t + \mathrm{Tr}((W^k)^{-1})], \end{aligned} \tag{6.45}$$

where

$$W_i^k = 1/\sqrt{\sum_{t=1}^{T}(\nabla_1 X_{t,i}^k)^2 + (\nabla_2 X_{t,i}^k)^2 + \epsilon}, \forall\ i, \tag{6.46}$$

and

$$W^k = \begin{Bmatrix} W_1^k & & & \\ & W_2^k & & \\ & & ... & \\ & & & W_N^k \end{Bmatrix}. \tag{6.47}$$

It indicates that the weights for X_1 to X_T are the same. Similar, X_t can be updated by solving:

$$(A_t^T A_t + \lambda D_1^T W^k D_1 + \lambda Q_2^T W^k D_2)x = A_t^T b_t. \tag{6.48}$$

It also can be solved efficiently by the PCG method with the proposed preconditioner. Again to avoid repetition, the algorithm for JTV based reconstruction is not listed here.

6.5 Experiments

6.5.1 Experiment setup

The experiments are conducted using MATLAB on a desktop computer with 3.4GHz Intel core i7 3770 CPU. We validate different versions of our method on wavelet tree sparsity based reconstruction, wavelet joint sparsity reconstruction, TV and JTV reconstruction. To avoid confusion, we denote the tree sparsity version as FIRLS_OG and non-overlapping joint sparsity version FIRLS_MT. The version for standard ℓ_1 norm minimization is denoted by FIRLS_L1. FIRLS_TV and FIRLS_JTV denotes the TV and JTV reconstruction, respectively.

Table 6.1. Computational cost comparison between FOCUSS [17] and the proposed method

	FOCUSS [17]			FIRLS_L1		
Time (seconds)	64.8	110.8	727.7	10.5	29.8	120.2
MSE	0.0485	0.0442	0.0432	0.0481	0.0440	0.0427

Note that some algorithms need a very small number of iterations to converge (higher convergence rate), while they cost more time in each iteration (higher complexity). The others take less time in each iteration; however, more iterations are required. As we are interested in fast reconstruction, an algorithm is said to be better if it can achieve higher reconstruction accuracy with less computational time.

6.5.2 The accuracy of the proposed preconditioner

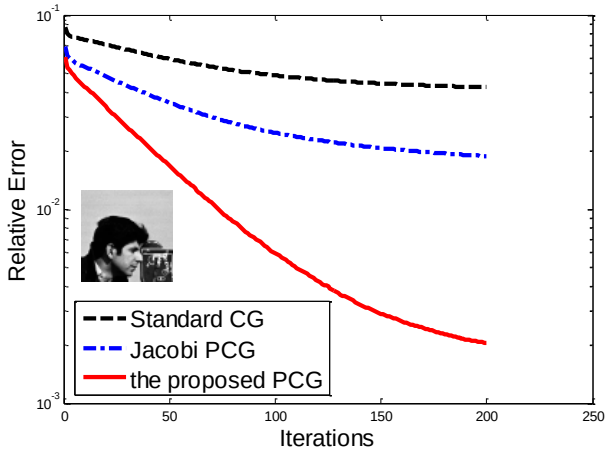


Figure 6.2. Convergence rate comparison among standard CG, Jacobi PCG and the proposed PCG for ℓ_1 norm minimization.

One of our contributions is the proposed pseudo-diagonal preconditioner for sparse recovery. First, we conduct an experiment to validate its effectiveness with the

orthogonal wavelet basis. Without loss of generality, a patch (64×64) cropped from the cameraman image is used for reconstruction, which is feasible to obtain the closed form solution of S^{-1} for evaluation. As most existing preconditioners cannot support the inverse of operators, the sampling matrix is set as the random projection and Φ is a dense matrix for wavelet basis here. Fig. 6.2 demonstrates the performance of the proposed PCG compared with Jacobi PCG and standard CG for the problem (6.15). The performance of the proposed PCG with less than 50 iterations is better than that of CG and Jacobi PCG with 200 iterations. Although Jacobi preconditioner is diagonal, it removes all the non-diagonal elements which makes the preconditioner less precise.

To validate the effectiveness of the proposed preconditioner in TV reconstruction, we take experiments on the Shepp-Logan phantom image with 64×64 pixels. The Shepp-Logan phantom image is very smooth and is an ideal example to validate TV reconstruction. The relative errors of CG, PCG Jacobi and the proposed method are shown in Fig. 6.3. It shows that only 20 iterations of PCG with the proposed preconditioner can outperform conventioanal CG with 200 iterations. Jacobi PCG requires approximately 2 times of iterations to reach the same accuracy as our method, because it discards all non-diagonal information directly and makes the preconditioning less precise. Comparing with the results in Fig. 6.2, our preconditioner seems less powerful on TV reconstruction. This is expected as we further decompose the preconditioner into two triangle matrices L and U, which introduces minor approximation error. However, it still converges much faster than the existing Jacobi PCG. These experiments demonstrate that the inner loop subproblem in our method is solved efficiently with the proposed preconditioner.

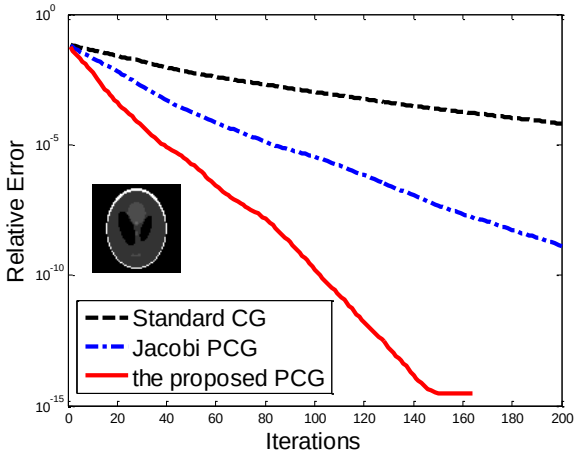


Figure 6.3. Convergence rate comparison among standard CG, Jacobi PCG and the proposed PCG for TV minimization.

6.5.3 Convergence Rate and Computational Complexity

One of the properties of the proposed FIRLS is its fast convergence rate, i.e., only a small number of iterations can achieve high reconstruction accuracy. In addition, each iteration has low computational cost. To validate its fast convergence rate, we compare it with three existing algorithms with known convergence rate. They are the IST algorithm SpaRSA [56], FISTA [42] and IRLS algorithm FOCUSS [17], with $\mathcal{O}(1/k)$, $\mathcal{O}(1/k^2)$ and exponential convergence rates, respectively. Mean squared error (MSE) is used as the evaluation metric.

The test data is a random 1D signal of length 4000, with 10% elements being non-zeros. The number of measurements are 800. Fig. 6.4 demonstrates the comparison. In each iteration, FOCUSS needs to compute the inverse of a large-scale matrix, and the proposed method uses 30 PCG iterations to approximate the inverse. Both FOCUSS and the proposed method converge within 200 iterations. FISTA tends to converge at about 800 iterations. However, SpaRSA requires much more than 800 iterations to converge. Table 6.1 lists the reconstruction results at different CPU time between FOCUSS and the proposed method. The proposed algorithm always

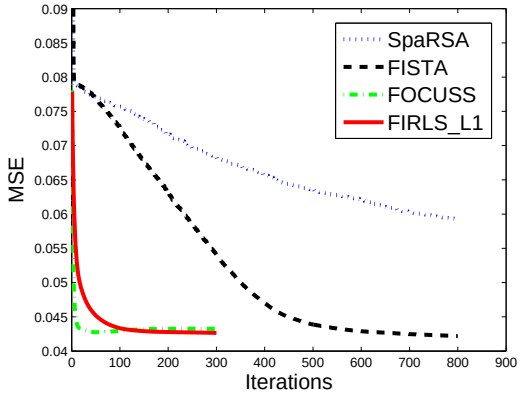


Figure 6.4. Convergence Rate Comparison among FOCUSS, FISTA and SpaRSA for ℓ_1 norm minimization .

achieves more accurate result in much less time. After convergence, the 0.0005 difference in terms of MSE may be caused by approximation or rounding errors. With the size of the data becomes larger, the time cost of FOCUSS will increase at a cubic speed. More importantly, it is not known how to solve the overlapping group sparsity problem with FOCUSS.

6.5.4 Application: compressive sensing MRI

Compressive sensing MRI (CS-MRI) [1] is one of the most successful applications of compressive sensing and sparsity regularization. There are various sparsity patterns on MR images. Therefore, we validate the performance of different versions of our method on CS-MRI as a concrete reconstruction instance. Partial but not full k-space data is acquired and the final MR image can be reconstructed by exploiting the sparsity of the image. With little information loss, this scheme can significantly accelerate MRI acquisition. In CS-MRI, $A = RF$ is an undersampled Fourier operator, where F is the Fourier transform and $R \in \mathbb{R}^{M \times N}$ is a selection matrix containing M rows of the identity matrix. Therefore, $A^T A = F^T R^T RF$ is diagonally dominant

as $R^T R$ is diagonal. Based on (6.16), $\overline{A^T A}$ is identical to the sampling ratio (a fixed scalar).

Following previous works, Signal-to-Noise Ratio (SNR) are used as metric for result evaluation:

$$SNR = 10 \log_{10}(V_s/V_n), \tag{6.49}$$

where V_n is the Mean Square Error between the original image x_0 and the reconstructed x; $V_s = var(x_0)$ denotes the variance of the values in x_0.

6.5.5 CS-MRI

6.5.5.1 CS-MRI with wavelet tree sparsity

The MR images are often piecewise smooth, which are widely assumed to be sparse under the wavelet basis or in the gradient domain [1, 2, 3, 4]. Furthermore, the wavelet coefficients of a natural image yield a quadtree. If a coefficient is zero or nonzero, its parent coefficient also tends to be zero or nonzero. This wavelet tree structure has already been successfully utilized in MR image reconstruction, approximated by the overlapping group sparsity [14, 88]. Tree-structured CS-MRI method [14, 88] has been shown to be superior to standard CS-MRI methods [1, 2, 4]. Therefore, we compare our algorithm with two latest and fastest tree-based algorithms, turbo AMP [13] and WaTMRI [14]. In addition, overlapping group sparsity solvers SLEP [15, 152] and YALL1 [16] are also compared. The total number of iterations is 100 except that turbo AMP only runs 10 iterations due to its higher time complexity. Followed by the previous works [2, 4, 14], four MR images with the same size 256×256 are used for testing, which are shown in Fig. 6.5. Using a similar sampling strategy, we randomly choose more Fourier coefficients from low frequency and less on high

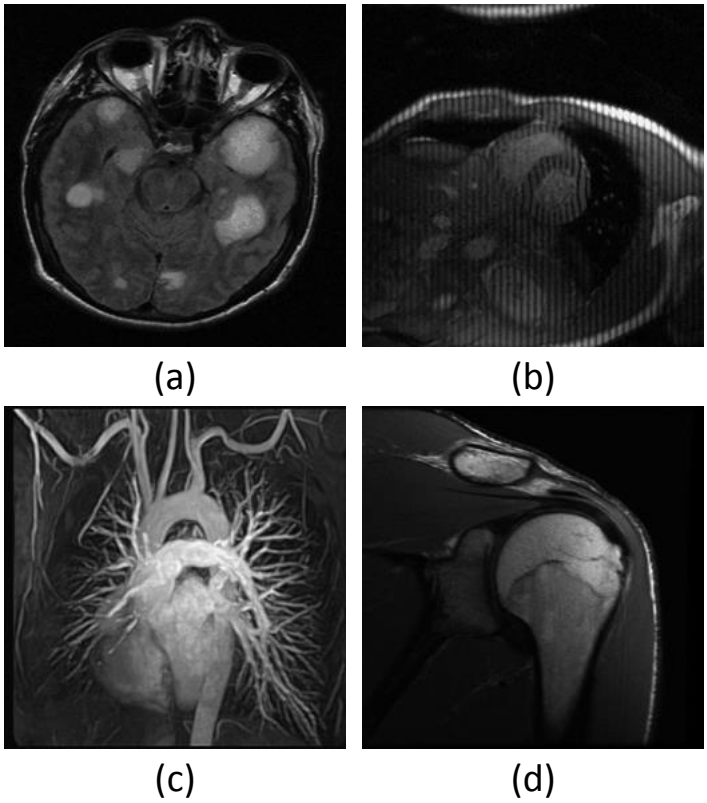


Figure 6.5. The original images: (a) Brain; (b) Cardiac; (c) Chest; (d) Shoulder.

frequency. The sampling ratio is defined as the number of sampled measurements divided by the total size of the signal/image.

A visual comparison on the Brain image is shown in Fig. 6.6, with a 25% sampling ratio. Visible artifacts can be found on the results by YALL1 [16]. The image reconstructed by the AMP [13] tends to be blurry when compared with the original. The image recovered by SLEP [15] is noisy. Our method and WaTMRI [14] produce the most accurate results in terms of SNR. Note that WaTMRI has more parameters required to be tuned due to its variable splitting strategy. Besides SNR, we also compare the mean structural similarity [121] (MSSIM) of different images, which mimics the human visual system. The MSSIM for the images recovered by

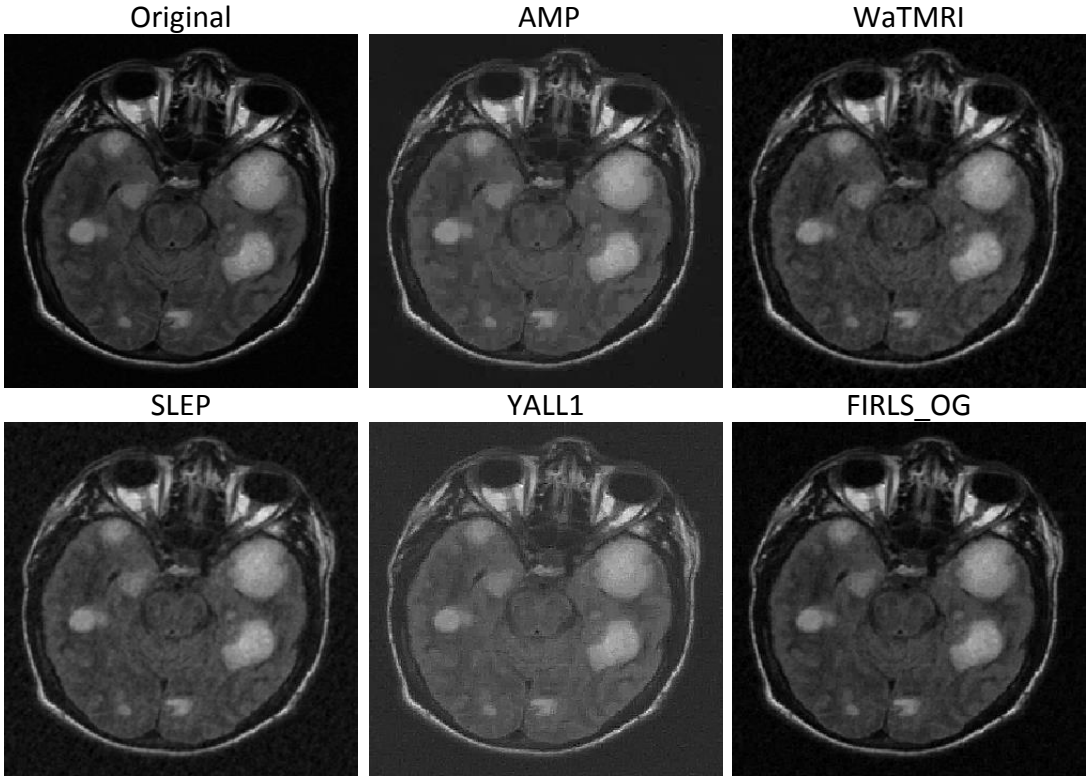


Figure 6.6. Visual comparison on the Brain image with 25% sampling. The SNRs of AMP [13], WaTMRI [14], SLEP [15], YALL1 [16] and the proposed method are 15.91, 16.72, 16.49, 12.86 and 18.39, respectively.

AMP [13], WaTMRI [14], SLEP [15], YALL1 [16] and the proposed method are 0.8890, 0.8654, 0.8561, 0.7857 and 0.9009. In terms of MSSIM, our method still has the best performance, which is consistent with the observation in terms of SNR.

The corresponding convergence speed of the this experiment is presented in Fig. 6.7. From SNR versus outer loop iterations, the proposed algorithm far exceeds that of all other algorithms, which is due to the fast convergence rate of IRLS. However, there is no known convergence rate better than $\mathcal{O}(1/k^2)$ for WaTMRI and SLEP, and $\mathcal{O}(1/k)$ for YALL1, respectively. These results are consistent with that in previous work [14]. For the same number of total iterations, the computational cost of our method is comparable to the fastest one YALL1, and it significantly outperforms

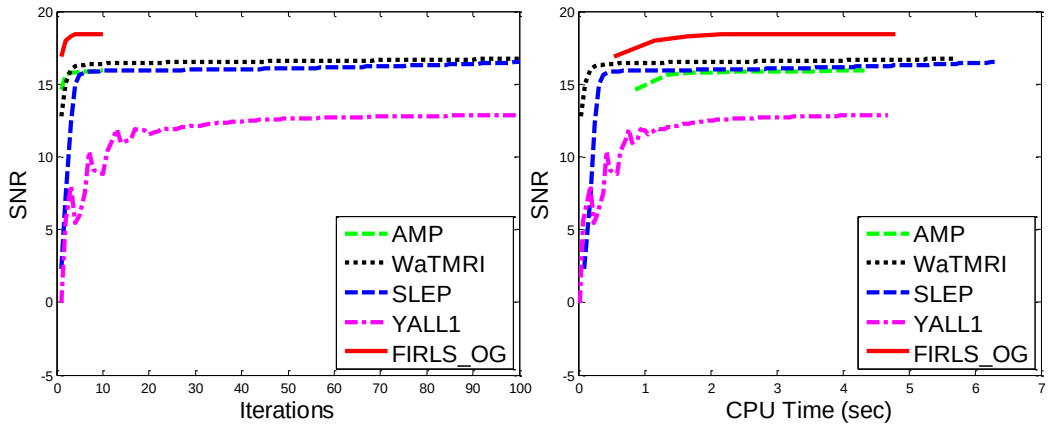


Figure 6.7. Convergence speed comparison on the Brain image with 25% sampling. Left: SNR vs outer loop iterations. Right: SNR vs CPU time. The SNRs of reconstructed images with these algorithms are 15.91, 16.72, 16.49, 12.86 and 18.39 respectively. The time costs are 4.34 s, 5.73 s, 6.28 s, 4.71 s and 4.80 s, respectively.

YALL1 in terms of reconstruction accuracy. SLEP has the same formulation as ours. To reach our result in this experiment, it requires around 500 iterations with about 43 seconds. Similar results can be obtained on the other testing images. The results on the four images with different sampling ratios are listed in Table 2. Our results are consistently more accurate.

Table 6.2. Average SNR (dB) comparisons on the four MR images with wavelet tree sparsity.

Sampling Ratio	20%	23%	25%	28%	30%
AMP [13]	11.64	15.7	16.43	17.08	17.44
WaTMRI [14]	15.56	17.43	18.23	19.22	20.45
SLEP [15]	11.59	16.51	17.36	18.51	20.07
YALL1 [16]	12.13	13.29	14.12	15.29	16.07
FIRLS_OG	**15.67**	**18.78**	**19.43**	**20.53**	**21.52**

6.5.5.2 CS-MRI by TV reconstruction

TV is another popular regularizer for MRI reconstruction and the images recovered by TV tend to be less noisy [1]. For TV based reconstruction, we compare our method with classical method CG [1] and the fastest ones TVCMRI [2], RecPF [3], FCSA [4] and SALSA [156].

The convergence speed of different algorithms on the Chest image is presented in Fig. 6.8. It is worthwhile to note that no closed form solutions exist for the subproblems of these algorithms. Therefore, the subproblems in these algorithms are often solved in an approximate way. Therefore, it is important to evaluate the accuracies of these algorithms. From the figure, the final results of our method and TVCMRI are almost the same while the others converges to different results. We further found that only TVCMRI has analyzed their global convergence (in Section 2.3 of [2]), while the accuracy of all the other methods [1, 3, 4, 156] has not been discussed in details. For the four MR images, the average SNRs of CG [1], TVCMRI [2], RecPF [3], FCSA [4], SALSA [156] and the proposed algorithm are 19.45, 21.78, 21.70, 21.53 21.95 and 23.07, respectively.

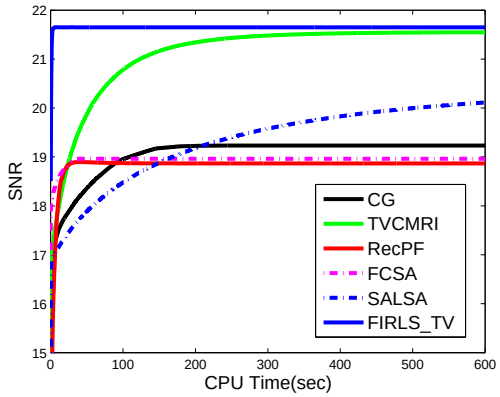


Figure 6.8. Convergence rate comparison for TV minimization on the Chest image with 25% sampling.

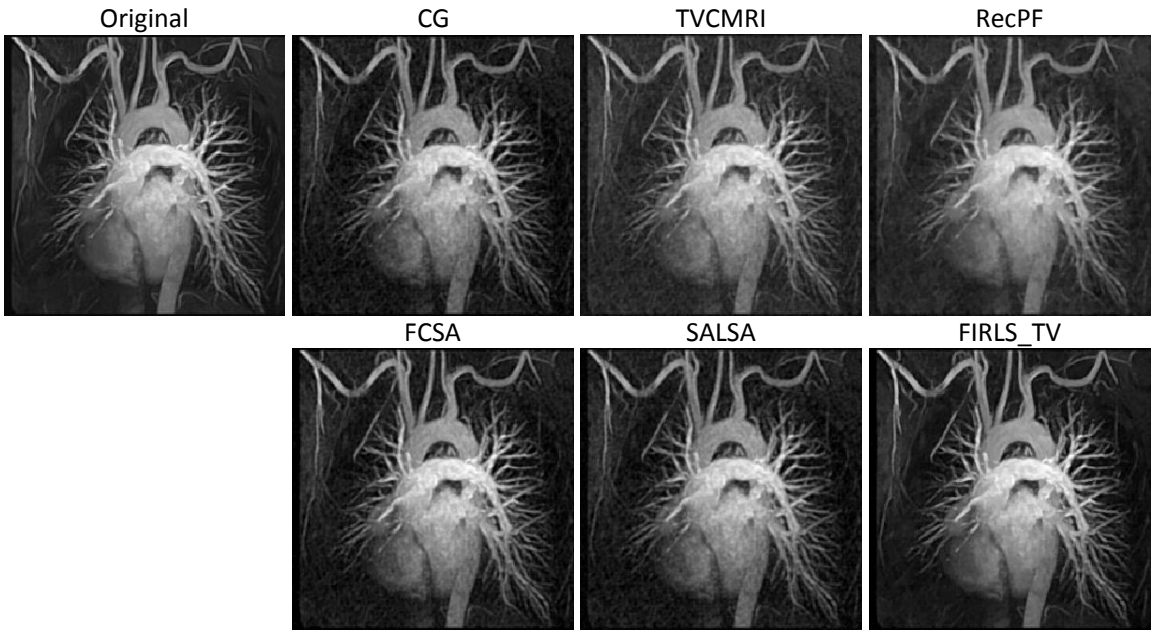


Figure 6.9. Chest MR image reconstruction from 25% sampling. All methods terminate after 4 s. The SNRs for CG, TVCMRI, RecPF, FCSA, SALSA and the proposed are 17.13, 17.32, 16.18, 18.28, 16.96 and 21.63, respectively.

We then terminate each algorithm after a fixed toleration is reached, e.g. 10^{-3} of the relative solution change. The final SNR and convergence speed of different methods are listed in Table 6.3. To produce a similar result of TVCMRI, our method only requires about its 1/70 computational time. These convergence performances are not surprising. FIRLS is converges exponentially fast (as shown in Fig. 6.4) and require the fewest iterations. FCSA is a FISTA based algorithm, which has $\mathcal{O}(1/k^2)$ convergence rate. It converges with the second fewest iterations. For the rest algorithms, there is no known convergence rate better than $\mathcal{O}(1/k)$.

Due to the relatively slower convergence speed, we note that previous methods [1, 2, 3, 4] often terminate after a fixed number of iterations (e.g. 200) in practice. This is because the exactly convergence is time consuming that may not be feasible

Table 6.3. Quantitative comparison of convergence speed on the Chest image by TV regularization with 25% sampling.

	Iterations	CPU time (sec)	SNR (dB)
CG [1]	3181	397.8	19.23
TVCMRI [2]	21392	495.1	21.54
RecPF [3]	7974	163.4	18.86
FCSA [4]	1971	39.6	18.96
SALSA [156]	9646	882.4	20.13
FIRLS_TV	**29**	**6.9**	**21.65**

for clinic applications. Following by this scheme, we run TVCMRI 200 iterations. All the other algorithms terminate after the same running time of TVCMRI (i.e. around 4 seconds). The reconstruction results on the Chest MR image are shown in Fig. 6.9. A close look shows that our method preserve highest organ-to-background contrast without contaminated by reconstruction noise. Such results are expected if we take a review on Figure 6.8. Similar results can be obtained on the Brain, Cardiac and Artery images.

6.5.6 Multi-contrast MRI

6.5.6.1 Multi-contrast MRI with wavelet joint sparsity

To assist clinic diagnose, multiple MR images with different contrasts are often acquired simultaneously from the same anatomical cross section. For example, T1 and T2 weighted MR images could distinguish fat and edema better, respectively. Different from the CS-MRI for individual MR imaging, multi-contrast reconstruction for weighted MR images means the simultaneous reconstruction of multiple T1/T2-weighted MR images. Joint sparsity of the wavelet coefficients and JTV across different contrasts have been used in recent multi-contrast reconstruction methods [98, 78].

Here, the multi-contrast MR images are extracted from the SRI24 Multi-Channel Brain Atlas Data [168]. An example of the test images is shown in Fig. 6.10. We compare our method with the fastest multi-contrast MRI methods [98, 78], which use the algorithms SPGL1_MMV [40] and FCSA to solve the corresponding problems, respectively. The experiment setup is the similar as in the previous experiments, except group setting is constructed for joint sparsity (non-overlapping) case. FCSA_MT and FIRLS_MT denotes the algorithm in [78] and the proposed method in this setting.

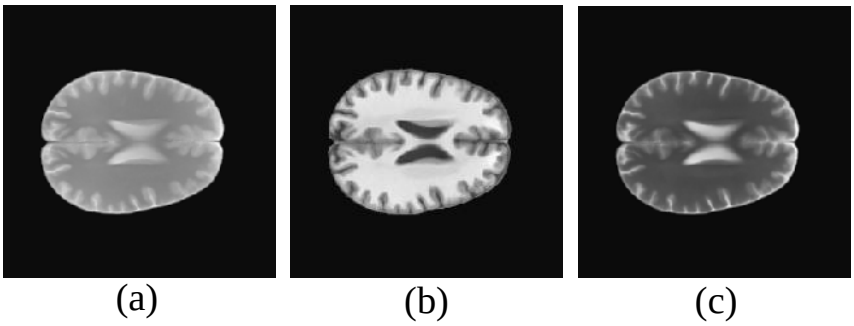
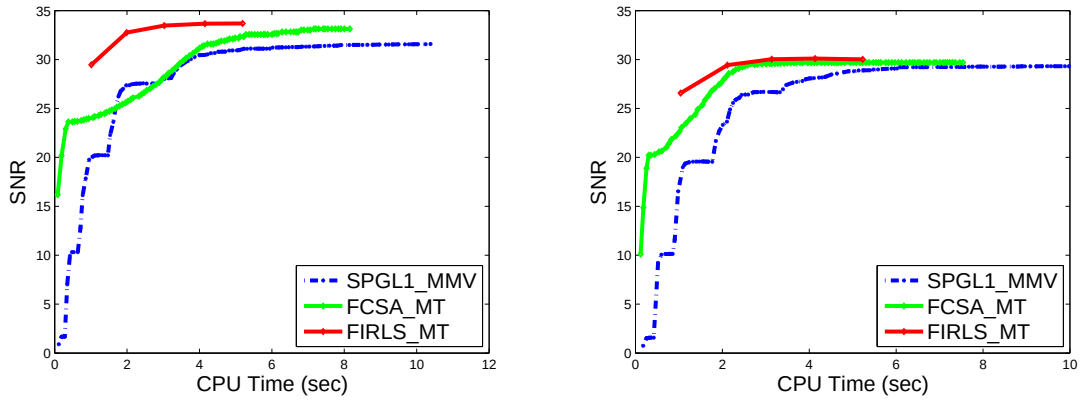


Figure 6.10. The original images for multi-contrast MRI.

Fig. 6.11 shows the performance comparisons among SPGL1_MMV [40], FCSA_MT [78] and FIRLS_MT on the example images shown in Figure 6.10. Each algorithm runs 100 iterations in total. After convergence, three algorithms achieve similar accuracy for 20% sampling and SPGL1 is only slightly worse than others for 25% sampling. From the curves, our method always ourperforms SPGL1_MMV and FCSA_MT, i.e., higher accuracy for the same reconstruction time.

To quantitatively compare the convergence speed of these three methods, we conduct experiments on 20 set images (i.e. total 60 images) that are from SRI24. Different from the tree-based CS-MRI, each algorithm for non-overlapping group sparsity converges much faster. To reduce randomness, all algorithms run 100 times and the



(a) (b)

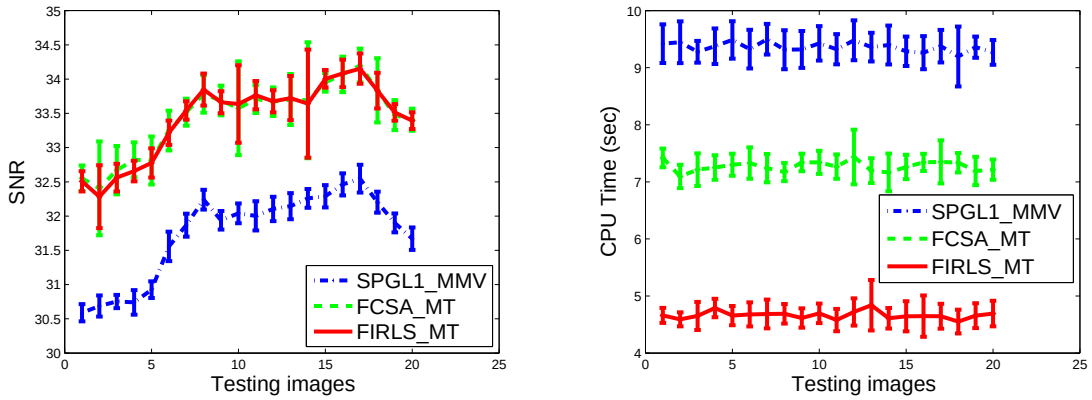
Figure 6.11. (a) Performance comparison for multi-contrast MRI with 25% sampling. The average time costs of SPGL1_MMV, FCSA_MT, and the proposed method are 10.38 s, 8.15 s, 5.19 s. Their average SNRs are 31.58, 33.12 and 33.69. (b) Performance comparison for multi-contrast MRI with 20% sampling. Their average time costs are 9.98 s, 7.54 s, 5.23 s. Their average SNRs are 29.31, 29.69 and 30.01.

reconstruction results are shown in Fig. 6.12. With 25% sampling, the accuracy of our method is almost the same as FCSA_MT, and always better than SPGL1. In the process to achieve the convergence, our method is consistently faster than the other two algorithms. These results demonstrate the efficiency of proposed method.

6.5.6.2 Multi-contrast MRI with JTV

Finally, we reconstruct multi-contrast MR images with JTV. It is unknown how to extend existing methods CG [1], TVCMRI [2], RecPF [3] and SALSA [156] to the JTV versions. Here, we only compare our method FIRLS_JTV with FCSA_JTV [78].

Fig. 6.13 shows the performance comparison on the example images (in Fig. 6.10) from 25% and 30% sampling, without setting stoping criteria. After convergence, the accuracy of our results are slightly higher than those of FCSA_JTV. Also, it is clearly that FIRLS_JTV requires much less time to converge in both cases. We



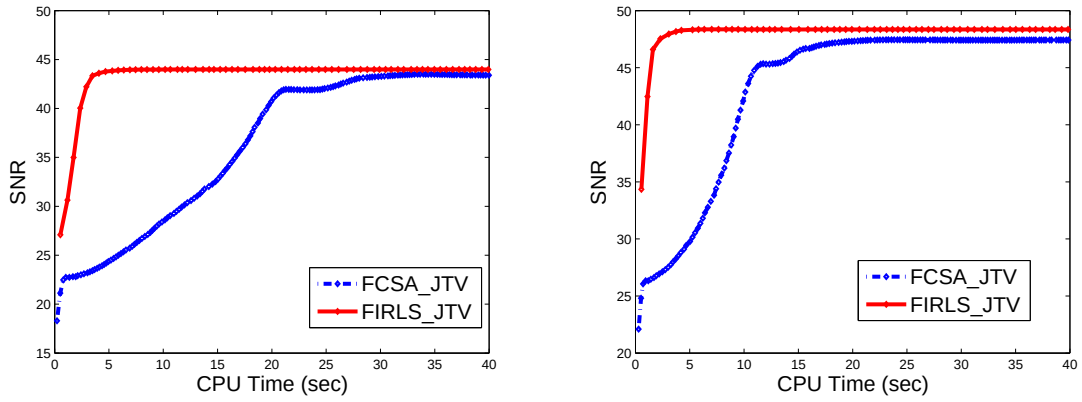
(a) (b)

Figure 6.12. Performance comparison on 60 images from SRI24 dataset with 25% sampling. (a) SNR comparison. (b) CPU time comparison. The average convergence time for SPGL1, FCSA_MT and the proposed FIRLS_MT is 9.3 s, 7.2 s, 4.6 s, respectively.

then let each algorithm terminate with the 10^{-3} tolerance. FCSA_JTV cost 35.6 seconds and 19.7 seconds to converge for the two sampling cases, while the proposed FIRLS_JTV only requires 6.7 seconds and 4.6 seconds for the two cases, respectively.

6.5.7 Discussion

The first and second experiments validate the fast convergence speed of our method due to the proposed preconditioner. The advantages over the state-of-the-arts are further validated on practical application CS-MRI with four sparsity patterns: overlapping groups with tree sparsity, non-overlapping groups with joint sparsity, TV and JTV. Although results on these problems are promising, some difference can be found. The non-overlapping group sparsity problem is often easier to solve. For example, the subproblem in FISTA has the closed form solution for joint sparsity but not for overlapping group sparsity. However, our method has similar difficulty for non-overlapping and overlapping group sparsity. That is why our method outperforms the



(a) (b)

Figure 6.13. Multi-contrast MRI with JTV reconstruction. (a) The performance comparison with 25% sampling. (b) The performance comparison with 30% sampling .

fastest methods on joint sparsity reconstruction, and significantly outperforms those for tree-sparsity reconstruction, TV and JTV reconstruction. We do not compare the performance between the wavelet transform and TV, since their performances are data-dependent. In this work, we only focus on fast minimization of the given functions.

The superior performance of the proposed preconditioner also attributes to the structure of the system matrix S, which is often diagonally dominant in reconstruction problems (e.g. A is random projection or partial Fourier transform). It can be applied to other applications where S is not diagonally dominant (e.g. image blurring), and will be still more accurate than Jacobi preconditioner as it keeps more non-diagonal information.

CHAPTER 7

Conclusion

In this thesis, we have developed the tree sparsity model for MRI and forest sparsity model for sparse learning and compressed sensing. A dynamic gradient sparsity model is developed to improve image fusion. To overcome the difficulties in image registration, we propose a new method based on the deep sparse representation of images. Finally, a set of efficient algorithms are derived to solve the sparsity regularization problems. The main contributions of this thesis are summarized as follows:

- MRI has been one of the most successful applications of sparsity techniques and compressed sensing. While most of exiting methods only exploit the sparsity of the MR images, we proposed a new model based on the wavelet tree structure for CSMRI. We developed an efficient algorithm for the tree-based MR image reconstruction. This method can be easily extended to other medical imaging problems, such as CT image reconstruction. The extensive experiments have demonstrated the effectiveness and efficiency of our method over the previous works.
- Although the wavelet tree sparsity can significantly improve the reconstruction accuracy of a single-channel image, it is not able to utilize the structure correlations in multiple/multi-channel image reconstruction. Therefore, we extended the tree sparsity to the forest sparsity on multi-channel data. Under compressive sensing assumptions, significant reduction of measurements is achieved with forest sparsity compared with standard sparsity, joint sparsity or independent

tree sparsity. The benefit of the proposed model has been theoretically proved and empirically validated in practical applications.

- The above sparse models assumes that the images themselves can be sparsely represented in some transforming domain. In some applications, the data is has strong correlations to some known patterns. More precisely, the data to be recovered can be more sparsely represented based on the information of the reference. We proposed the dynamic gradient sparsity model for such data. We have validated the proposed dynamic gradient sparsity model on image fusion of remote sensing data. In the experiments, our method is shown to impressively outperform the classical methods and the recent methods in terms of both accuracy and computational complexity.
- In the images registration problem, the images to be registered may have different intensity fields due to the various illumination conditions. In addition, the contents in these images may be slightly different, as the images may be formed in different times. To overcome such difficulty, we proposed a hierarchical sparsity model for images registration. Through multiple sparse representation layers, the model can handle intensity distortions, outliers and partial occlusions. With these advantages, this model has been shown to be superior than the existing models in a wide range of applications.
- The optimization problem by using the advanced sparsity models is often more difficult to solve. Motivated by the special data structures in image processing and medical imaging. We proposed a set of fast iterative reweighted least squares algorithms for the analysis-based sparsity reconstruction problems. The proposed method inherit the fast convergence rate of the traditional IRLS algorithms, that is, exponentially fast. Moreover, with the devised preconditioner, the computational cost for each iteration is significantly less than that

of traditional IRLS algorithms, which makes it feasible for large scale problems. Extensive results demonstrate that the proposed method achieves superior performance over 14 state-of-the-art algorithms in terms of both accuracy and computational cost.

Although each of the above models can not be applied to all the medical imaging and image processing applications, the success of our work provides new possibilities for future research. It implies that developing new methods with the suggestions of theoretical results, as well as exploiting the prior information of the data, may lead impressive results.

REFERENCES

[1] M. Lustig, D. Donoho, and J. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[2] S. Ma, W. Yin, Y. Zhang, and A. Chakraborty, "An efficient algorithm for compressed mr imaging using total variation and wavelets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[3] J. Yang, Y. Zhang, and W. Yin, "A fast alternating direction method for tvl1-l2 signal reconstruction from partial fourier data," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 288–297, 2010.

[4] J. Huang, S. Zhang, and D. Metaxas, "Efficient MR image reconstruction for compressed MR imaging," *Medical Image Analysis*, vol. 15, no. 5, pp. 670–679, 2011.

[5] S. Becker, J. Bobin, and E. Candès, "NESTA: a fast and accurate first-order method for sparse recovery," *SIAM Journal on Imaging Sciences*, vol. 4, no. 1, pp. 1–39, 2011.

[6] R. Szeliski, "Image alignment and stitching: A tutorial," *Foundations and Trends® in Computer Graphics and Vision*, vol. 2, no. 1, pp. 1–104, 2006.

[7] A. Myronenko and X. Song, "Intensity-based image registration by minimizing residual complexity," *IEEE Transactions on Medical Imaging*, vol. 29, no. 11, pp. 1882–1891, 2010.

[8] J. Kim and J. A. Fessler, "Intensity-based image registration using robust correlation coefficients," *IEEE Transactions on Medical Imaging*, vol. 23, no. 11, pp. 1430–1444, 2004.

[9] B. Cohen and I. Dinstein, "New maximum likelihood motion estimation schemes for noisy ultrasound images," *Pattern Recognition*, vol. 35, no. 2, pp. 455–463, 2002.

[10] A. Myronenko, X. Song, and D. J. Sahn, "Maximum likelihood motion estimation in 3d echocardiography through non-rigid registration in spherical coordinates," in *Functional Imaging and Modeling of the Heart*, 2009, pp. 427–436.

[11] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[12] F. Zana and J.-C. Klein, "A multimodal registration algorithm of eye fundus images using vessels detection and hough transform," *IEEE Transactions on Medical Imaging*, vol. 18, no. 5, pp. 419–428, 1999.

[13] S. Som and P. Schniter, "Compressive imaging using approximate message passing and a markov-tree prior," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3439–3448, 2012.

[14] C. Chen and J. Huang, "Compressive Sensing MRI with Wavelet Tree Sparsity," in *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1124–1132.

[15] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University, 2009. [Online]. Available: http://www.public.asu.edu/ jye02/Software/SLEP

[16] W. Deng, W. Yin, and Y. Zhang, "Group sparse optimization by alternating direction method," *TR11-06, Department of Computational and Applied Mathematics, Rice University*, 2011.

[17] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 600–616, 1997.

[18] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[19] E. Candès, "Compressive sampling," in *Proceedings of the International Congress of Mathematicians*, 2006, pp. 1433–1452.

[20] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 763–770.

[21] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[22] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[23] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2009, pp. 417–424.

[24] D. Reddy, A. Agrawal, and R. Chellappa, "Enforcing integrability by error correction using 1-minimization," in *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR). IEEE, 2009, pp. 2350–2357.

[25] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[26] C. Chen, F. Tian, H. Liu, and J. Huang, "Diffuse optical tomography enhanced by clustered sparsity for functional brain imaging." *IEEE Transactions on Medical Imaging*, vol. 33, no. 12, pp. 2323–2331, 2014.

[27] C. Chen, Y. Li, L. Axel, and J. Huang, "Real time dynamic mri with dynamic total variation," in *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).* Springer, 2014, pp. 138–145.

[28] C. Chen, Z. Peng, and J. Huang, "O (1) algorithms for overlapping group sparsity," in *Proceedings of the International Conference on Pattern Recognition (ICPR).* IEEE, 2014, pp. 1645–1650.

[29] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[30] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," *Journal of Machine Learning Research*, vol. 12, pp. 3371–3412, 2011.

[31] J. Huang, "Structured sparsity: Theorems, Algorithms and Applications," Ph.D. dissertation, Rutgers University, 2011.

[32] K. Pruessmann, M. Weiger, M. Scheidegger, P. Boesiger, *et al.*, "SENSE: sensitivity encoding for fast MRI," *Magnetic Resonance in Medicine*, vol. 42, no. 5, pp. 952–962, 1999.

[33] M. Griswold, P. Jakob, M. Nittka, J. Goldfarb, and A. Haase, "Partially parallel imaging with localized sensitivities (PILS)," *Magnetic Resonance in Medicine*, vol. 44, no. 4, pp. 602–609, 2000.

[34] D. Sodickson and W. Manning, "Simultaneous acquisition of spatial harmonics (SMASH): fast imaging with radiofrequency coil arrays," *Magnetic Resonance in Medicine*, vol. 38, no. 4, pp. 591–603, 2005.

[35] M. Griswold, P. Jakob, R. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase, "Generalized autocalibrating partially parallel acquisitions (GRAPPA)," *Magnetic Resonance in Medicine*, vol. 47, no. 6, pp. 1202–1210, 2002.

[36] M. Lustig and J. Pauly, "SPIRiT: Iterative self-consistent parallel imaging reconstruction from arbitrary k-space," *Magnetic Resonance in Medicine*, vol. 64, no. 2, pp. 457–471, 2010.

[37] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[38] D. Liang, B. Liu, J. Wang, and L. Ying, "Accelerating sense using compressed sensing," *Magnetic Resonance in Medicine*, vol. 62, no. 6, pp. 1574–1584, 2009.

[39] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2419–2434, 2009.

[40] E. Van Den Berg and M. Friedlander, "Probing the pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.

[41] J. Huang, S. Zhang, and D. Metaxas, "Efficient MR image reconstruction for compressed MR imaging," in *Proceedings of International Conference on Med-*

ical Image Computing and Computer-Assisted Intervention (MICCAI), 2010, pp. 135–142.

[42] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[43] A. Manduca and A. Said, "Wavelet compression of medical images with set partitioning in hierarchical trees," in *Proceedings of the SPIE Symposium on Medical Imaging*, 1996.

[44] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Transactions on Circuits and systems for video technology*, vol. 6, no. 3, pp. 243–250, 1996.

[45] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 886–902, 1998.

[46] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Structured sparsity through convex optimization," *Statistical Science*, vol. 27, no. 4, pp. 450–468, 2012.

[47] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal methods for hierarchical sparse coding," *Journal of Machine Learning Research*, vol. 12, pp. 2297–2334, 2011.

[48] C. La and M. Do, "Tree-based orthogonal matching pursuit algorithm for signal reconstruction," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2006, pp. 1277–1280.

[49] N. Rao, R. Nowak, S. Wright, and N. Kingsbury, "Convex approaches to model wavelet sparsity patterns," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 1917–1920.

[50] L. He and L. Carin, "Exploiting structure in wavelet-based bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 9, pp. 3488–3497, 2009.

[51] L. He, H. Chen, and L. Carin, "Tree-structured compressive sensing with variational bayesian analysis," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 233–236, 2010.

[52] L. Jacob, G. Obozinski, and J. Vert, "Group lasso with overlap and graph lasso," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2009, pp. 433–440.

[53] J. Ye, S. Tak, Y. Han, and H. Park, "Projection reconstruction MR imaging using FOCUSS," *Magnetic Resonance in Medicine*, vol. 57, no. 4, pp. 764–775, 2007.

[54] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 707–710, 2007.

[55] J. Trzasko and A. Manduca, "Highly undersampled magnetic resonance image reconstruction via homotopic ℓ_0-minimization," *IEEE Transactions on Medical Imaging*, vol. 28, no. 1, pp. 106–121, 2009.

[56] S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2479–2493, 2009.

[57] Y. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$," *Doklady AN USSR (translated as Soviet Math. Docl.)*, vol. 269, pp. 543–547, 1983.

[58] C. Chen and J. Huang, "The benefit of tree sparsity in accelerated MRI," in *MICCAI Workshop on Sparsity Techniques in Medical Imaging*, 2012.

[59] C. Chen, Y. Li, and J. Huang, "Forest sparsity for multi-channel compressive sensing," *IEEE Transactions on Signal Processing*, vol. 62, no. 11, pp. 2803–2813, 2014.

[60] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, p. 969, 2007.

[61] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[62] B. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[63] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[64] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[65] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[66] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[67] D. Needell and J. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[68] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007.

[69] K. Koh, S. Kim, and S. Boyd, "An interior-point method for large-scale l1-regularized logistic regression," *Journal of Machine Learning Research*, vol. 8, no. 8, pp. 1519–1555, 2007.

[70] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2346–2356, 2008.

[71] D. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[72] J. Meng, W. Yin, H. Li, E. Hossain, and Z. Han, "Collaborative spectrum sensing from sparse observations in cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 2, pp. 327–337, 2011.

[73] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, 1996.

[74] D. Baron, M. Wakin, M. Duarte, S. Sarvotham, and R. Baraniuk, "Distributed compressed sensing," *arXiv preprint arXiv:0901.3403*, 2005.

[75] A. Majumdar and R. Ward, "Compressive color imaging with group-sparsity on analysis prior," in *Proceedings of the 17th IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 1337–1340.

[76] C. Chen, Y. Li, W. Liu, and J. Huang, "Image fusion with local spectral consistency and dynamic gradient sparsity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2760–2765.

[77] B. Bilgic, V. Goyal, and E. Adalsteinsson, "Multi-contrast reconstruction with bayesian compressed sensing," *Magnetic Resonance in Medicine*, vol. 66, no. 6, pp. 1601–1615, 2011.

[78] J. Huang, C. Chen, and L. Axel, "Fast Multi-contrast MRI Reconstruction," in *Proceedings of the Annual International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2012, pp. 281–288.

[79] J. Huang and T. Zhang, "The benefit of group sparsity," *The Annals of Statistics*, vol. 38, no. 4, pp. 1978–2004, 2010.

[80] J. Huang, X. Huang, and D. Metaxas, "Learning with dynamic group sparsity," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 64–71.

[81] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2005.

[82] F. Bach, "Consistency of the group lasso and multiple kernel learning," *Journal of Machine Learning Research*, vol. 9, pp. 1179–1225, 2008.

[83] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, 2005.

[84] D. Wipf and B. Rao, "An empirical bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3704–3716, 2007.

[85] S. Ji, D. Dunson, and L. Carin, "Multitask compressive sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 92–106, 2009.

[86] J. Ziniel and P. Schniter, "Efficient high-dimensional inference in the multiple measurement vector problem," *arXiv preprint arXiv:1111.5272*, 2011.

[87] S. Kim and E. Xing, "Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping," *The Annals of Applied Statistics*, vol. 6, no. 3, pp. 1095–1117, 2012.

[88] C. Chen and J. Huang, "The benefit of tree sparsity in accelerated MRI," *Medical image analysis*, vol. 18, no. 6, pp. 834–842, 2014.

[89] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar, "C-hilasso: A collaborative hierarchical sparse modeling framework," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4183–4198, 2011.

[90] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Uniform uncertainty principle for bernoulli and subgaussian ensembles," *Constructive Approximation*, vol. 28, no. 3, pp. 277–289, 2008.

[91] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[92] T. Blumensath and M. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1872–1882, 2009.

[93] X. Chen, S. Kim, Q. Lin, J. G. Carbonell, and E. P. Xing, "Graph-structured multi-task regression and an efficient optimization method for general fused lasso," *arXiv preprint arXiv:1005.3579*, 2010.

[94] X. Chen, X. Shi, X. Xu, Z. Wang, R. Mills, C. Lee, and J. Xu, "A two-graph guided multi-task lasso approach for eqtl mapping," in *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012, pp. 208–217.

[95] M. Kowalski, K. Siedenburg, and M. Dörfler, "Social sparsity! neighborhood systems enrich structured shrinkage operators," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2498–2511, 2013.

[96] J. Huang, S. Zhang, H. Li, and D. Metaxas, "Composite splitting algorithms for convex optimization," *Computer Vision and Image Understanding*, vol. 115, no. 12, pp. 1610–1622, 2011.

[97] T. Rohlfing, N. Zahr, E. Sullivan, and A. Pfefferbaum, "The SRI24 multichannel atlas of normal adult human brain structure," *Human Brain Mapping*, vol. 31, no. 5, pp. 798–819, 2009.

[98] A. Majumdar and R. Ward, "Joint reconstruction of multiecho MR images using correlated sparsity," *Magnetic Resonance Imaging*, vol. 29, no. 7, pp. 899–906, 2011.

[99] C. Chen and J. Huang, "Exploiting both intra-quadtree and inter-spatial structures for multi-contrast MRI," in *Proceedings of the International Symposium on Biomedical Imaging (ISBI)*, 2014.

[100] A. Majumdar and R. Ward, "Calibration-less multi-coil MR image reconstruction," *Magnetic Resonance Imaging*, 2012.

[101] C. Chen, Y. Li, and J. Huang, "Calibrationless parallel MRI with joint total variation regularization," in *Proceedings of the Annual International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2013, pp. 106–114.

[102] J. Ma, "Single-pixel remote sensing," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 2, pp. 199–203, 2009.

[103] G. G. Lorentz, M. von Golitschek, and Y. Makovoz, *Constructive approximation: advanced problems*. Springer Berlin, 1996, vol. 304.

[104] J. Y. Park, H. L. Yap, C. J. Rozell, and M. B. Wakin, "Concentration of measure for block diagonal matrices with applications to compressive signal processing," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 5859–5875, 2011.

[105] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot, "Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1301–1312, 2008.

[106] P. S. Chavez, Jr, S. C. Sides, and J. A. Anderson, "Comparison of three different methods to merge multiresolution and multispectral data: Landsat tm and spot panchromatic," *Photogrammetric Engineering and Remote Sensing*, vol. 57, no. 3, pp. 295–303, 1991.

[107] R. Haydn, G. W. Dalke, J. Henkel, and J. E. Bare, "Application of the IHS color transform to the processing of multisensor data and image enhancement," in *Proceedings of the International Symposium on Remote Sensing of Environment*, 1982.

[108] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge landsat tm and spot panchromatic data," *International J. Remote Sensing*, vol. 19, no. 4, pp. 743–757, 1998.

[109] I. Amro, J. Mateos, M. Vega, R. Molina, and A. K. Katsaggelos, "A survey of classical methods and new trends in pansharpening of multispectral images," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 79, pp. 1–22, 2011.

[110] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. i. decorrelation and hsi contrast stretches," *Remote Sensing of Environment*, vol. 20, no. 3, pp. 209–235, 1986.

[111] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé, "A variational model for P+ XS image fusion," *International Journal of Computer Vision*, vol. 69, no. 1, pp. 43–58, 2006.

[112] M. Möller, T. Wittman, A. L. Bertozzi, and M. Burger, "A variational approach for sharpening high dimensional images," *SIAM Journal of Imaging Sciences*, vol. 5, no. 1, pp. 150–178, 2012.

[113] F. Fang, F. Li, C. Shen, and G. Zhang, "A variational approach for pan-sharpening," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2822–2834, 2013.

[114] T. Goldstein and S. Osher, "The split bregman method for l1-regularized problems," *SIAM Journal Imaging Sciences*, vol. 2, no. 2, pp. 323–343, 2009.

[115] X. Bresson and T. F. Chan, "Fast dual minimization of the vectorial total variation norm and applications to color image processing," *Inverse Problems and Imaging*, vol. 2, no. 4, pp. 455–484, 2008.

[116] V. Vijayaraj, C. G. O'Hara, and N. H. Younan, "Quality analysis of pansharpened images," in *Proc. International Geoscience and Remote Sensing Symposium*, 2004.

[117] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images," *Photogrammetric Engineering and Remote Sensing*, vol. 63, no. 6, pp. 691–699, 1997.

[118] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce, "Comparison of pansharpening algorithms: Outcome of the 2006 grs-s data-fusion contest," *IEEE Transactions on Geoscience and Remote Sensing*.

[119] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[120] M. Choi, "A new intensity-hue-saturation fusion approach to image fusion with a tradeoff parameter," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 6, pp. 1672–1682, 2006.

[121] 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, no. 4, pp. 600–612, 2004.

[122] M. Kumar and S. Dass, "A total variation-based algorithm for pixel-level image fusion," *IEEE Transactions on Image Processing*, vol. 18, no. 9, pp. 2137–2143, 2009.

[123] Y. Li, C. Chen, F. Yang, and J. Huang, "Deep sparse representation for robust image registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[124] J. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical image analysis*, vol. 2, no. 1, pp. 1–36, 1998.

[125] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[126] J. Modersitzki, *Numerical methods for image registration*. OUP Oxford, 2003.

[127] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[128] P. Blanc, L. Wald, T. Ranchin, *et al.*, "Importance and effect of co-registration quality in an example of pixel to pixel fusion process," in *Proceedings of the 2nd International Conference Fusion of Earth Data: merging point measurements, raster maps and remotely sensed images*, 1998, pp. 67–74.

[129] Y. Zheng, E. Daniel, A. A. Hunter III, R. Xiao, J. Gao, H. Li, M. G. Maguire, D. H. Brainard, and J. C. Gee, "Landmark matching based retinal image alignment by enforcing sparsity in correspondence matrix," *Medical image analysis*, vol. 18, no. 6, pp. 903–913, 2014.

[130] J. Ma, J. Zhao, J. Tian, Z. Tu, and A. L. Yuille, "Robust estimation of non-rigid transformation for point set registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2147–2154.

[131] J. Huang, X. Huang, and D. Metaxas, "Simultaneous image transformation and sparse representation recovery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[132] G. Tzimiropoulos, V. Argyriou, S. Zafeiriou, and T. Stathaki, "Robust FFT-based scale-invariant image registration with image gradients," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1899–1906, 2010.

[133] C. Studholme, C. Drapaca, B. Iordanova, and V. Cardenas, "Deformation-based mapping of volume change from serial brain MRI in the presence of local tissue contrast change," *IEEE Transactions on Medical Imaging*, vol. 25, no. 5, pp. 626–639, 2006.

[134] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2233–2246, 2012.

[135] Y. Wu, B. Shen, and H. Ling, "Online robust image alignment via iterative convex optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1808–1814.

[136] J. He, D. Zhang, L. Balzano, and T. Tao, "Iterative grassmannian optimization for robust image alignment," *Image and Vision Computing*, vol. 32, no. 10, pp. 800–813, 2014.

[137] V. Hamy, N. Dikaios, S. Punwani, A. Melbourne, A. Latifoltojar, J. Makanyanga, M. Chouhan, E. Helbren, A. Menys, S. Taylor, *et al.*, "Respiratory motion correction in dynamic mri using robust data decomposition registration–application to dce-mri," *Medical image analysis*, vol. 18, no. 2, pp. 301–313, 2014.

[138] N. Chumchob, "Vectorial total variation-based regularization for variational image registration," *IEEE Transactions on Image Processing*, vol. 22, no. 11, 2013.

[139] C. Frohn-Schauf, S. Henn, and K. Witsch, "Multigrid based total variation image registration," *Computing and Visualization in Science*, vol. 11, no. 2, pp. 101–113, 2008.

[140] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[141] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[142] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *European Conference on Computer Vision (ECCV)*, 1992, pp. 237–252.

[143] http://www.robots.ox.ac.uk/∼vgg/research/affine/.

[144] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[145] C. A. Cocosco, V. Kollokian, R. K.-S. Kwan, G. B. Pike, and A. C. Evans, "Brainweb: Online interface to a 3d MRI simulated brain database," *NeuroImage*, 1997.

[146] Y. Li, C. Chen, J. Zhou, and J. Huang, "Robust image registration in the gradient domain," in *Proceedings of the International Symposium on Biomedical Imaging (ISBI)*, 2015.

[147] Space-Imaging, "IKONOS scene po-37836," *Geoeye IKONOS Scene Data*, 2000. [Online]. Available: http://glcf.umd.edu/data/ikonos/

[148] http://www.mathworks.com/help/images/register-an-aerial-photograph-to-a-digital-orthophoto.html.

[149] F. Bach, "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2011.

[150] Y. Xiao and J. Yang, "A fast algorithm for total variation image reconstruction from random projections," *arXiv preprint arXiv:1001.1774*, 2010.

[151] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Transactions on Image Processing*, vol. 12, no. 8, pp. 882–889, 2003.

[152] L. Yuan, J. Liu, and J. Ye, "Efficient methods for overlapping group lasso," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2104–2116, 2013.

[153] S. Mosci, S. Villa, A. Verri, and L. Rosasco, "A primal-dual algorithm for group sparse regularization with overlapping groups," in *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2010, pp. 2604–2612.

[154] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Communications on Pure and Applied Mathematics*, vol. 63, no. 1, pp. 1–38, 2010.

[155] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 3869–3872.

[156] M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Transactions on Image Processing*, vol. 19, no. 9, pp. 2345–2356, 2010.

[157] P. Rodríguez and B. Wohlberg, "Efficient minimization method for a generalized total variation functional," *IEEE Transactions on Image Processing*, vol. 18, no. 2, pp. 322–332, 2009.

[158] C. Chen, J. Huang, L. He, and H. Li, "Preconditioning for accelerated iteratively reweighted least squares in structured sparsity reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2713–2720.

[159] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[160] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2010, pp. 543–550.

[161] J. Liu and J. Ye, "Moreau-yosida regularization for grouped tree structure learning." in *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, vol. 23, 2010, pp. 1459–1467.

[162] D. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, pp. 30–37, 2004.

[163] Y. Saad, *Iterative methods for sparse linear systems*. SIAM, 2003.

[164] G. Papandreou and A. Yuille, "Efficient variational inference in large-scale bayesian compressed sensing," in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011.

[165] A. B. S. Lefkimmiatis and M. Unser, "Hessianbased norm regularization for image restoration with biomedical applications," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 983–995, 2012.

[166] J. Yang, W. Yin, Y. Zhang, and Y. Wang, "A fast algorithm for edge-preserving variational multichannel image restoration," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 569–592, 2009.

[167] S. H. Chan, R. Khoshabeh, K. B. Gibson, P. E. Gill, and T. Q. Nguyen, "An augmented lagrangian method for total variation video restoration," *IEEE Transactions on Image Processing*, vol. 20, no. 11, pp. 3097–3111, 2011.

[168] T. Rohlfing, N. Z. NM, E. Sullivan, and A. Pfefferbaum, "The sri24 multichannel atlas of normal adult human brain structure," *Human Brain Mapping*, vol. 31, pp. 798–819, 2010.

BIOGRAPHICAL STATEMENT

Chen Chen was born in Jingshan, Hubei, China in 1987. He received the B.E. degree and M.S. degree both from Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2011, respectively. He has been a graduate student in the Department of Computeter Science and Engineering at the University of Texas at Arlington since 2012. His major research interests include image processing, medical imaging, computer vision and machine learning. He has published more than 20 papers during his study in he University of Texas at Arlington, including the ones in top tier conferences CVPR, NIPS, MICCAI and journals such as IEEE transactions. He is a student member of several IEEE societies.