CHAPTER

Dictionary learning for medical image denoising, reconstruction, and segmentation

6

T. Tong, J. Caballero, K. Bhatia, D. Rueckert
Imperial College London, London, United Kingdom

CHAPTER OUTLINE

6.1 INTRODUCTION

Throughout the past decades, different forms of representation have emerged, leading in recent years to dictionary learning (DL). DL is the term given to the search for optimal sparse signal transforms which are obtained through a training stage, which is a radically different approach to signal modeling compared to hand-crafted

Machine Learning and Medical Imaging. http://dx.doi.org/10.1016/B978-0-12-804076-8.00006-2

signal models such as wavelets. It has been successfully applied to numerous image processing tasks such as denoising, super-resolution, and segmentation, and led to state-of-the-art results in recent years. In what follows we will present a brief history of signal transforms leading to DL and a summary of some of the DL algorithms available. Further details can be found in two excellent reviews (Rubinstein et al., 2008; Tosic and Frossard, 2011).

The choice of a transform for signal representation is crucial and involves a number of compromises. The use of orthogonal or bi-orthogonal transforms has long been favored because transform coefficients are given by a simple inner product between the signal and the transform or the transform inverse, respectively. However, the use of complete bases has limitations in representation flexibility as some signals may not be well encompassed by their modeling. The desire for greater flexibility at the expense of mathematical complexity drove the switch from complete transform bases to overcomplete dictionaries, and from transform functions to dictionary atoms.

6.1.1 THE CONVENIENCE OF ORTHOGONAL TRANSFORMS

One of the most recurrent signal analysis tools is the Fourier transform, which was greatly popularized in the 1960s with the emergence of the fast Fourier transform proposed in Cooley and Tukey (1965). The decomposition of a signal into its global frequency content can sparsely represent uniformly smooth signals, but is very inefficient for capturing discontinuities given that their energy is spread among several frequency coefficients. Sharp discontinuities are rare in natural signals, but the periodic assumption of finite signals for the computation of its transform artificially creates them at the signal boundary. This naturally led to the use of the discrete cosine transform (DCT), which avoids this phenomenon by assuming odd periodicity and is the core ingredient of the JPEG image compression standard discussed in Wallace (1992).

The following decades of the 1970s and 1980s centered the search of data simplicity on the data itself. Statistical tools such as principal component analysis as shown in Jolliffe (2005) and most notably the Karhunen-Loève transform as revisited by Mallat (1999) gained interest as they reduced the complexity of the signal on a low-dimensional subspace with minimum l_2-norm error. Using the first few eigenvectors of the eigenvalue decomposition of a signal's covariance matrix, it can be seen as a low-dimensional Gaussian data fit. Although it is more powerful as a data sparsifier than the Fourier transform, it is considerably more complex given its data-driven nature.

During the 1980s, it became clear that the search for simpler, sparser representations required the departure from restrictive linear transforms leading to the design of nonlinear transforms, where the support of nonzero coefficients is signal specific. Two major concepts are at the origin of wavelet design, emerging with the specific purpose of nonlinear sparse coding for natural signals: localization and multiresolution.

The Fourier transform allows the identification of the different frequency content of a signal, but it does not reveal where in time or space this content can be found. This lack of localization hinders compact signal representation, and results of this realization were the short-time Fourier transform and the use of Gabor filters as originally proposed by Gabor (1946). Multiresolution analysis was the consequence of noticing how natural signals exhibit fractal-like patterns, which repeat at different scales. Multiscale wavelet analysis was introduced in Grossmann and Morlet (1984) as the scaling and translation of a single function of finite support which could be designed to form an orthogonal basis. In Mallat (1999) this concept was later extended for optimal 1D multiresolution signal analysis, and most importantly, fast algorithms for wavelet decomposition were proposed enabling their practical use. Even though at higher dimensions wavelet analysis loses its optimality, these advances were adopted in the newer JPEG2000 image compression mechanism as described in Skodras et al. (2001).

Wavelet analysis is also limited in the lack of adaptability and geometric invariance. The orthogonality condition limits the range of temporal or spatial support of the functions which, if violated, allows for greater flexibility in representation. Wavelet packets were suggested by Coifman et al. (1992) as an extension to wavelets which, given a signal, could be reduced to the optimal orthogonal subset, gaining adaptability but keeping the attractive properties of orthogonal wavelets. Translation and rotation sensitivity are further drawbacks of standard wavelet transforms which were deemed unavoidable by orthogonal transforms in Simoncelli et al. (1992). Thus began work on overcomplete transforms with early examples such as the stationary wavelet transform in Beylkin (1992) seeking geometric invariance.

6.1.2 THE FLEXIBILITY OF OVERCOMPLETE DICTIONARIES

With the development in the 1990s of greedy algorithms for sparse solutions and the influential discovery that this problem could be approximated by the tractable l_1 relaxation, the use of overcomplete frames adopting the name of dictionaries was popularized. Allowing multiple representations of the same signal in a dictionary of atoms opened new perspectives in coding design, which could now be driven by a cost function, and markedly separated the task of dictionary design from signal coding. Simple concatenations of bases could overcome what used to be fundamental limitations. For instance, a Fourier transform was unable to compactly represent discontinuities, but concatenating it with a Dirac basis could immediately solve this problem.

Abandoning orthogonality paved the way for creative dictionary design. Two trends can currently be identified: the design of analytic dictionaries and data-driven adaptive dictionaries (Aharon et al., 2006; Mairal et al., 2009). The former approach relies on a mathematical model of the data to generate the dictionary and is usually characterized by efficient mechanisms to implicitly compute transform coefficients, as well as robust theoretical guarantees for signal approximation. Some examples of this category are curvelets as proposed in Candès and Donoho (1999), contourlets from Do and Vetterli (2002) and bandelets discussed in Le Pennec and Mallat (2005). Data-driven dictionary design is more recent, and draws from example observations of a signal to obtain an optimal representation. Adaptive dictionaries are powerful as there is no reason to believe a single dictionary should be optimal for all kinds of signals, but come at the price of increased processing complexity and weaker theoretical guarantees. The search for optimal dictionaries for a specific set of training signals is known as the DL problem.

6.2 SPARSE CODING AND DICTIONARY LEARNING

This section introduces some of the most important DL algorithms. In addition, sparse coding techniques are described before the introduction of DL techniques as they form an essential part of DL.

6.2.1 SPARSE CODING

The technique of finding a representation for a given signal with a small number of significant coefficients is referred to as sparse coding. To describe signal sparse coding we can rely on the sparse synthesis model, which is characterized by two properties: (1) synthesis: signals are assumed to be a linear combination of basis functions and (2) sparsity: the coding vector defining the synthesis of a signal from basis functions is sparse. Assume that the set of basis functions $\mathbf{D} \in \mathbb{R}^{M\times K}$ is given, which is called a dictionary. $\mathbf{D}$ can be predefined (eg, overcomplete wavelets) or learnt. Each column in $\mathbf{D}$ is called an atom. Here, the dictionary $\mathbf{D}$ has K atoms and each atom has M elements. A signal $\mathbf{y} \in \mathbb{R}^{M\times 1}$ can then be represented as a linear combination of atoms in $\mathbf{D}$, which is formulated as

$$\mathbf{y} = \mathbf{D}\boldsymbol{\gamma}. \tag{6.1}$$

To seek a sparse solution, different regularization schemes have been proposed to impose prior information over the coding coefficients $\boldsymbol{\gamma}$. One commonly used regularization scheme adds the minimum l_0-norm constraint on the coefficients, minimizing the number of nonzero entries in $\boldsymbol{\gamma}$. The linear model is then formulated as

$$\min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{D}\boldsymbol{\gamma}\|_2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}\|_0 \leq S \tag{6.2}$$

or

$$\min_{\boldsymbol{\gamma}} \|\boldsymbol{\gamma}\|_0 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{D}\boldsymbol{\gamma}\|_2 \leq \varepsilon, \tag{6.3}$$

which represent the sparsity constrained and the error constrained l_0 norm problems, respectively. Both problems are equivalent in that it is possible to select an error ε that will provide the same solution as a maximum sparsity S. The use of the l_0-norm as a measure of sparsity makes the problem nonconvex and finding the exact solution for $\boldsymbol{\gamma}$ in the above equations is NP-hard (Elad, 2010).

Algorithms for finding approximate solutions have been extensively investigated. A very well-known algorithm is orthogonal matching pursuit (OMP) (Mallat and Zhang, 1993). OMP is a greedy approach that proposes to iteratively seek the locally optimal choice in the hope of approximating the global minimum. This is a reasonable compromise for the sparse approximation problem given that the global solution to the NP-hard problem is practically unreachable, but sequentially deciding the entries in $\boldsymbol{\gamma}$ that will minimize the approximation error is computationally very cheap. Various extensions such as compressive sampling OMP (CoSaMP) (Needell and Tropp, 2009), regularized OMP (ROMP) (Needell and Vershynin, 2010), and stagewise OMP (StOMP) (Donoho et al., 2012) have been proposed to accelerate the convergence of OMP.

A major inconvenience for solving Eqs. (6.2) and (6.3) is that they are nonconvex problems. An alternative is to look for the solution to the closest problem that is convex, for which the vast literature on convex optimization would immediately apply. This was proposed in Chen et al. (1998) with the relaxation of the l_0-norm by the l_1-norm, targeting the solutions to

$$\min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{D}\boldsymbol{\gamma}\|_2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}\|_1 \leq S \tag{6.4}$$

or

$$\min_{\boldsymbol{\gamma}} \|\boldsymbol{\gamma}\|_1 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{D}\boldsymbol{\gamma}\|_2 \leq \varepsilon, \tag{6.5}$$

which is known as the basis pursuit (BP) problem. This problem, in constrained form or as an l_1 regularized least-squared problem, can be efficiently solved with different approaches, such as homotopy methods (Tibshirani, 1996), coordinate-wise descent methods (Friedman et al., 2007), Bregman iterative methods (Osher et al., 2005), and iterative shrinkage methods (Bioucas-Dias and Figueiredo, 2008).

6.2.2 DICTIONARY LEARNING PROBLEM

In sparse coding, it is assumed that the overcomplete dictionary $\mathbf{D}$ is given or known a priori. The dictionary can be directly chosen as a set of training signals or a prespecified basis such as overcomplete wavelets, curvelets, contourlets, and short-time Fourier transforms. Recent research has focused on learning an overcomplete dictionary based on a set of training signals rather than choosing a prespecified dictionary. Given a set of training signals $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N] \in \mathbb{R}^{M\times N}$, it is assumed that there exists a dictionary $\mathbf{D}$ that can sparsely represent each signal in $\mathbf{Y}$. The process of DL is then formulated as

$$\min_{\mathbf{D},\boldsymbol{\Gamma}} \|\mathbf{Y} - \mathbf{D}\boldsymbol{\Gamma}\|_F^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}_i\|_0 \leq S \ \ \forall i. \tag{6.6}$$

Here each column in $\mathbf{\Gamma} \in \mathbb{R}^{K \times N}$ contains the coding coefficients corresponding to each training signal $\mathbf{y}_i$. It is a nonconvex problem to optimize Eq. (6.6) over the dictionary $\mathbf{D}$ and sparse coefficients $\mathbf{\Gamma}$ jointly (Aharon et al., 2006). This problem is intricately related to sparse coding, but has the additional difficulty that, on top of finding a sparse code, the dictionary for sparse representation has to be simultaneously estimated. Commonly, the problem is simplified by solving for the sparse code and the dictionary separately, and iteratively alternating their solutions until convergence. The method of optimal directions (MOD) (Engan et al., 1999) and the K-SVD (Aharon et al., 2006) are two efficient algorithms to learn dictionaries which utilize variants of this iterative optimization strategy. In practice, it has been observed that K-SVD converges with fewer iterations than MOD (Aharon et al., 2006). In the next section, we will give a detailed introduction to the K-SVD algorithm.

6.2.3 K-SVD DICTIONARY LEARNING

The K-SVD algorithm is inspired from the k-means clustering algorithm, which is also an NP-hard problem. The aim of k-means clustering is to partition all the signals into K clusters, in which each training signal belongs to the cluster with the nearest mean. It employs an iterative approach to find the solution of K clusters and there are two steps at each iteration: In the first step, each training signal is assigned to its nearest cluster; in the second step, the K clusters are updated as the centroids of their assigned training signals. The K-SVD follows a similar iterative two-step process to learn dictionary atoms and find sparse solutions of the training signals using those atoms. The dictionary is first initialized with a traditional data sparsifier, such as an overcomplete DCT dictionary. We can then look for the sparse coding matrix $\mathbf{\Gamma}$ by keeping the dictionary fixed and sparsely coding each training signal independently with an OMP coding stage. Then, the K atoms in the dictionary are updated separately in a dictionary update stage with $\mathbf{\Gamma}$ fixed. The K-SVD algorithm decomposes the penalty term by questioning one atom $\mathbf{d}_k$ and its associated sparse codes given in row k of $\mathbf{\Gamma}$. Denoting this row vector as $\boldsymbol{\gamma}_t^k$, the penalty can be rewritten as

$$\begin{aligned} \|\mathbf{Y} - \mathbf{D}\mathbf{\Gamma}\|_F^2 &= \left\| \mathbf{Y} - \sum_{i=1}^{K} \mathbf{d}_k \boldsymbol{\gamma}_t^k \right\|_F^2 \\ &= \left\| \left(\mathbf{Y} - \sum_{j \neq k} \mathbf{d}_j \boldsymbol{\gamma}_t^j \right) - \mathbf{d}_k \boldsymbol{\gamma}_t^k \right\|_F^2 \\ &= \|\mathbf{E}^k - \mathbf{d}_k \boldsymbol{\gamma}_t^k\|_F^2, \end{aligned} \tag{6.7}$$

where $\mathbf{E}^k$ would be the approximation error if atom $\mathbf{d}_k$ were to be removed from the dictionary.

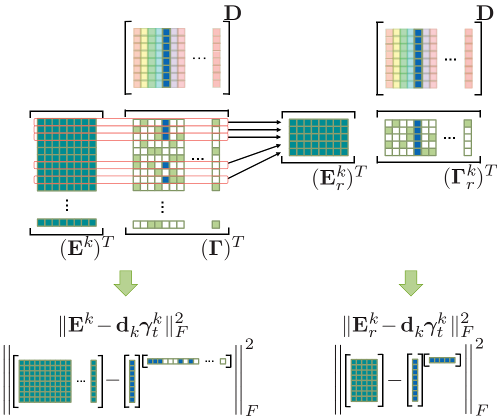
$\mathbf{D}$
$\mathbf{D}$
$(\mathbf{E}^k)^T$
$(\mathbf{\Gamma})^T$
$(\mathbf{E}_r^k)^T$
$(\mathbf{\Gamma}_r^k)^T$
$\|\mathbf{E}^k - \mathbf{d}_k\gamma_t^k\|_F^2$
$\|\mathbf{E}_r^k - \mathbf{d}_k\gamma_t^k\|_F^2$
2
F
2
F

FIG. 6.1

K-SVD dictionary update step. The left-hand side of the figure shows the decomposition of the approximation error in $\mathbf{E}^k$ and the contribution from $\mathbf{d}_k$ and $\boldsymbol{\gamma}_t^k$ through a rank-1 matrix. Updating $\mathbf{d}_k$ and $\boldsymbol{\gamma}_t^k$ directly with an SVD decomposition of $\mathbf{E}^k$ does not guarantee the maintenance of sparsity in $\boldsymbol{\gamma}_t^k$. Instead, shrinking the matrices as shown on the right-hand side of the figure solves this problem as only the active support of $\boldsymbol{\gamma}_r^k$ is updated.

With this separation, the approximation $\mathbf{D\Gamma}$ has been divided into K rank-1 matrices, only one of which is being questioned for update. Finding the rank-1 approximation of $\mathbf{E}_k$ through singular value decomposition (SVD) and using it for the update of $\mathbf{d}_k$ and $\mathbf{y}_t^k$ would be the optimal update step, but this is likely to fill row $\mathbf{y}_t^k$, which we would like to keep as a sparse vector. A simple solution is to only consider the indices $\mathbf{w}_k$ of nonzero entries in $\mathbf{y}_t^k$, and define the shrunken vector $\mathbf{y}_r^k$ and matrix $\mathbf{E}_r^k$. The rank-1 approximation of this new error matrix $\mathbf{E}_r^k$ provides then the optimal update of $\mathbf{d}_k$ and $\mathbf{y}_r^k$ while the sparse coding support is either unchanged or reduced. This shrinkage operation is illustrated in Fig. 6.1, and the full K-SVD algorithm is summarized in Algorithm 6.1. Although OMP is used for the sparse coding stage, the K-SVD algorithm is flexible and can work with other sparse coding methods (Aharon et al., 2006).

ALGORITHM 6.1 THE K-SVD DICTIONARY LEARNING ALGORITHM

```
Require: A set of training signals $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n] \in \mathbb{R}^{M \times N}$.
Ensure: An overcomplete dictionary $\mathbf{D} \in \mathbb{R}^{M \times K}$ and sparse coding coefficients $\mathbf{\Gamma} \in \mathbb{R}^{K \times N}$.
Initialize the dictionary $\mathbf{D}$ with $K$ randomly selected training signals
while converged do
    Sparse Coding:
    for each training signal $\mathbf{y}_i \in \mathbf{Y}$, use OMP to compute the corresponding coding
    coefficients $\boldsymbol{\gamma}_i$: do
```

```
            $\min_{\boldsymbol{\gamma}_i} \|\boldsymbol{\gamma}_i\|_0 \quad$ s.t. $\mathbf{y}_i = \mathbf{D}\boldsymbol{\gamma}_i, i = 1, \ldots, N$
    end for
    Dictionary Update:
    for $k = 1, \ldots, K$, update the $k$th atom $\mathbf{d}_k$ of $\mathbf{D}$ and the $k$th row $\boldsymbol{\gamma}_t^k$ of the coding coefficients $\boldsymbol{\Gamma}$: do
        Find the groups that use $\mathbf{d}_k$: $\mathbf{w}_k = \left\{i \in \{1, \ldots, N\} : \boldsymbol{\gamma}_t^k(i) \neq 0\right\}$, and $\boldsymbol{\gamma}_r^k$ is obtained by discarding zero entries in $\boldsymbol{\gamma}_t^k$.
        Compute representation error matrix: $\mathbf{E}^k = \mathbf{Y} - \sum_{i \neq k} \mathbf{d}_i \boldsymbol{\gamma}_t^k$.
        Obtain $\mathbf{E}_r^k$ by selecting the columns of $\mathbf{E}^k$ corresponding to $\mathbf{w}_k$.
        Apply SVD decomposition $\mathbf{E}_r^k = \mathbf{U}\Sigma\mathbf{V}^t$, update the atom $\mathbf{d}_k$ with the first column of $\mathbf{U}$, and update $\boldsymbol{\gamma}_r^k$ with the first column of $\mathbf{V}$ multiplied by $\Sigma(1, 1)$.
    end for
end while
```

Despite the fact that the K-SVD algorithm converges quickly, it is still computationally expensive at each iteration as an SVD decomposition must be calculated K times and all the N training signals are used for sparse coding at each iteration. This task is computationally expensive and relies on high memory use, especially when the set of training signals is large. Some modifications to the original algorithm have been presented that can alleviate computational cost, such as the use of Batch-OMP for K-SVD as presented in Rubinstein et al. (2008), and other strategies differing from K-SVD have also emerged. In the next section, we will introduce an online DL approach that can effectively solve this large-scale learning problem.

ALGORITHM 6.2 ONLINE DICTIONARY LEARNING ALGORITHM

```
Require: A set of training signals $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N] \in \mathbb{R}^{M \times N} \sim \rho(\mathbf{y})$, sparsity weight $\lambda$, $T$ (number of iterations).
Ensure: An overcomplete dictionary $\mathbf{D} \in \mathbb{R}^{M \times K}$.
Initialize the dictionary $\mathbf{D}_0$ with $K$ randomly selected training signals, $\mathbf{A}_0 \leftarrow 0$, $\mathbf{B}_0 \leftarrow 0$.
for $t = 1$ to $T$ do
    Draw $\mathbf{y}_t$ from $\rho(\mathbf{y})$.
    Sparse Coding (Lasso):
    $\hat{\boldsymbol{\gamma}}_t = \arg\min_{\boldsymbol{\gamma}_t} \|\mathbf{y}_t - \mathbf{D}_{t-1}\boldsymbol{\gamma}_t\|_2^2 + \lambda\|\boldsymbol{\gamma}_t\|_1$.
    $\mathbf{A}_t \leftarrow \mathbf{A}_{t-1} + \hat{\boldsymbol{\gamma}}_t\hat{\boldsymbol{\gamma}}_t^T$, $\mathbf{B}_T \leftarrow \mathbf{B}_{t-1} + \mathbf{y}_t\hat{\boldsymbol{\gamma}}_t^T$.
    Dictionary Update: compute $\mathbf{D}_t$ with $\mathbf{D}_{t-1}$ as initialization
    $\mathbf{D}_t = \arg\min_{\mathbf{D}} \frac{1}{t}\sum_{i=1}^{t}\left(\|\mathbf{y}_i - \mathbf{D}\boldsymbol{\gamma}_i\|_2^2 + \lambda\|\boldsymbol{\gamma}_i\|_1\right)$
        $= \arg\min_{\mathbf{D}} \frac{1}{t}\left(\frac{1}{2}Tr\left(\mathbf{D}^T\mathbf{D}\mathbf{A}_t\right) - Tr\left(\mathbf{D}^T\mathbf{B}_t\right)\right)$
end for
```

6.2.4 ONLINE DICTIONARY LEARNING

A stochastic online learning algorithm was proposed in Mairal et al. (2009) in order to learn dictionaries for a large set of training signals. A relaxed version of the objective function for DL using the l_1-norm is formulated as

$$\begin{aligned}\left\langle \hat{\mathbf{D}}, \hat{\mathbf{\Gamma}} \right\rangle &= \underset{\mathbf{D},\mathbf{\Gamma}}{\arg\min} \|\mathbf{Y} - \mathbf{D}\mathbf{\Gamma}\|_2^2 + \lambda \|\mathbf{\Gamma}\|_1 \\ &= \underset{\mathbf{D},\mathbf{\Gamma}}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \left(\|\mathbf{y}_i - \mathbf{D}\boldsymbol{\gamma}_i\|_2^2 + \lambda \left\|\boldsymbol{\gamma}_i\right\|_1 \right).\end{aligned} \tag{6.8}$$

This is not jointly convex over $\mathbf{D}$ and $\mathbf{\Gamma}$. In order to find the optimized solution, a stochastic gradient descent approach was utilized in Mairal et al. (2009) to update $\mathbf{D}$ sequentially. Instead of using the full training set at each iteration as in the K-SVD algorithm, the online DL algorithm updates the dictionary atoms by accessing one training signal at a time. Assuming that the set of training signals are independent and identically distributed (i.i.d.), one signal is drawn for updating $\mathbf{D}$ at each iteration as in the stochastic gradient descent. The online optimization process is summarized in Algorithm 6.2. It follows classic DL algorithms and alternates the sparse coding step with the DL step. However, at the current iteration, the new dictionary $\mathbf{D}_t$ uses the previous dictionary $\mathbf{D}_{t-1}$ as a warm restart, which is different from other DL algorithms. The new dictionary $\mathbf{D}_t$ is updated by minimizing the following function (Mairal et al., 2009):

$$\mathbf{D}_t = \underset{\mathbf{D}}{\arg\min} \frac{1}{t} \sum_{i=1}^{t} \left(\|\mathbf{y}_i - \mathbf{D}\boldsymbol{\gamma}_i\|_2^2 + \lambda \left\|\boldsymbol{\gamma}_i\right\|_1 \right). \tag{6.9}$$

The coding coefficients $\hat{\boldsymbol{\gamma}}_i$ computed during the previous iterations aggregate past information. The information from past coefficients $\hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2, \ldots, \hat{\boldsymbol{\gamma}}_t$ is carried forward in matrices:

$$\mathbf{A}_t \leftarrow \mathbf{A}_{t-1} + \hat{\boldsymbol{\gamma}}_t \hat{\boldsymbol{\gamma}}_t^T \quad \text{and} \quad \mathbf{B}_T \leftarrow \mathbf{B}_{t-1} + \mathbf{y}_t \hat{\boldsymbol{\gamma}}_t^T. \tag{6.10}$$

This enables updating dictionaries based on past information without accessing the past training samples again. The new dictionary $\mathbf{D}_t$ can then be optimized by using these matrices and the previous dictionary $\mathbf{D}_{t-1}$ as initialization. This optimization strategy leads to faster convergence performance and better dictionaries than classical batch algorithms, scaling up gracefully to large datasets even with millions of training samples (Mairal et al., 2009).

6.3 PATCH-BASED DICTIONARY SPARSE CODING

In this section we look at practical considerations of sparse recovery problems using dictionaries. Specifically, we present the implications for sparse coding brought on by overcompleteness, redundancy and adaptability. The experiments below use the Batch-OMP implementation described in Rubinstein et al. (2008).

6.3.1 OVERCOMPLETENESS

Consider the sparse recovery problem

$$\min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{D}\boldsymbol{\gamma}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}\|_0 \leq S, \tag{6.11}$$

where $\mathbf{D} \in \mathbb{C}^{M\times K}$, $M \leq K$, is a dictionary. The solution to this problem is trivial in the complete, orthonormal case ($M = K$, $\mathbf{D}^H\mathbf{D} = \mathbf{I}$), given that the generalized Parseval theorem holds between $\mathbf{y}$ in the signal domain and $\boldsymbol{\gamma}$ in the sparsity domain. Therefore energy is preserved upon basis transformation, implying that $\|\mathbf{y} - \mathbf{D}\boldsymbol{\gamma}\|_2^2 = \|\mathbf{D}^H\mathbf{y} - \boldsymbol{\gamma}\|_2^2$, and so the best S sparse representation is trivially given by the S largest coefficients of the transform $\mathbf{D}^H\mathbf{y}$. Energy preservation between domains is, however, violated as soon as $M < K$, which adds considerable flexibility to the sparse recovery problem given that it brings about a regime where one signal can have multiple dictionary representations.

To illustrate this, let us assume we extract an $M = 8 \times 8$ patch from Fig. 6.2A. We then use OMP to solve Eq. (6.11) with a sparsity index $S = 6$. Lastly, we look at the mean squared error (MSE) $\frac{1}{M}\|\mathbf{y} - \mathbf{D}\boldsymbol{\gamma}\|_2^2$ of the result produced using DCT dictionaries of different sizes. This experiment is repeated with 10^4 different

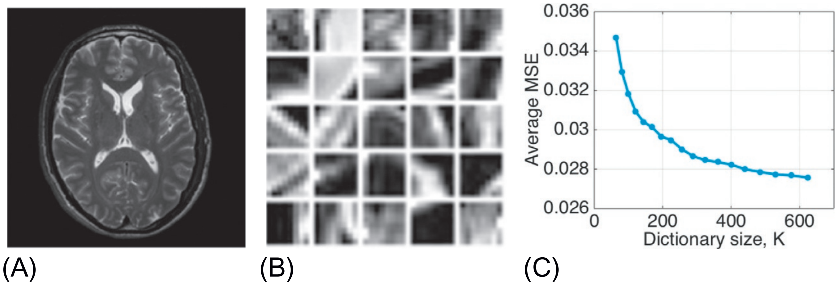

FIG. 6.2

Dictionary overcompleteness translates into increased representation sparsity. The plot in (C) shows the average representation error of 10^4 patches of size 8×8 from image (A) using a DCT dictionary of different sizes and a sparsity index $S = 6$. (A) Brain MR image. (B) Sample patches. (C) Reconstruction error.

patches and average results are plotted in Fig. 6.2C. Examples of the test patches are displayed in Fig. 6.2B. Despite the advantage of domain transformation that an orthonormal dictionary provides, the dictionary representation becomes more accurate with increasing overcompleteness.

6.3.2 REDUNDANCY

Redundant dictionaries are typically highly coherent, which immediately poses a problem for sparse recovery given that multiple representations are plausible for the same signal. If the focus, however, is not on the recovery of the particular sparse code that generated the signal but on any sparse code that will approximate it, redundant dictionaries can be useful. We analyze this statement by considering an overcomplete DCT dictionary with $K = 225$ atoms of size $M = 64$. Given a sparsity degree $S = 5$, we synthesize an 8×8 patch by linearly combining S randomly chosen atoms with random weights, and we then try to recover this sparse code with OMP solving

$$\min_{\boldsymbol{\gamma}} \|\boldsymbol{\gamma}\|_0 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{D}\boldsymbol{\gamma}\|_2 \leq \epsilon, \tag{6.12}$$

for $\epsilon = 10^{-5}\sqrt{M}$.

The greedy approach of OMP to sparse recovery is sometimes able to perfectly find the sequence of sparse coefficients that make up the signal. However, the coherence of the dictionary can sometimes make OMP fail dramatically in finding the original support of the sparse code, but the redundancy in the dictionary makes it possible to find an alternative sparse configuration that still achieves the data consistency level required in the signal domain. This behavior is analyzed for different degrees of sparsity S in Fig. 6.3. The plots show average results for the same experiment with 10^4 different patches. Despite the correct support recovery decreasing quickly for $S \geq 2$ (Fig. 6.3A), the signal domain reconstruction accuracy can be maintained below the predefined threshold (Fig. 6.3B) at the expense of a denser representation $\boldsymbol{\gamma}$ (Fig. 6.3C).[1]

6.3.3 ADAPTABILITY

One of the main advantages of adaptive dictionaries over structured dictionaries is a sparser representation for a predefined set of signals. This comes at the cost of a computationally intensive training process and the loss of structure, meaning that implicit and efficient dictionary transforms are not available and theoretical guarantees of the dictionary are more difficult to derive.

In this section we compare the sparse representations of a structured dictionary with one that is trained using the K-SVD algorithm from Aharon et al. (2006). For the

[1] The sparsity of the result was measured as the number of nonzero coefficients accounting for 99.9% of the energy of the sparse code $\boldsymbol{\gamma}$.

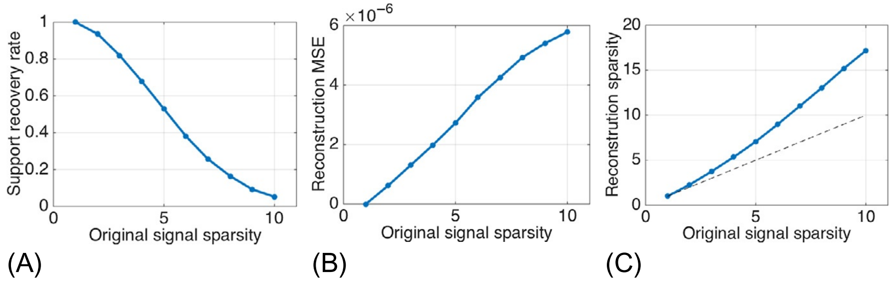

FIG. 6.3

Average empirical OMP recovery performance using an overcomplete DCT dictionary. The support recovery rate falls almost to zero for $S/M = 10/64 = 0.15$ sparsity (A). This is due to the high degree of redundancy in the dictionary, which makes OMP choose sparse coding configurations that are not the ones used to generate the original patches. Nevertheless, OMP is able to maintain the predefined data consistency tolerance (B) by using a few additional atoms relative to the original signal (C). (A) Support recovery rate. (B) Average MSE. (C) Average sparsity.

comparison we use the magnitude brain magnetic resonance (MR) image shown in Fig. 6.2A, of size $N = 256 \times 256$. Breaking the image down into $M = 8 \times 8$ overlapping patches and assuming patches wrap around the boundaries of the image, we have a total of N signals to be coded arranged as column vectors in $\mathbf{Y} \in \mathbb{R}^{M\times N}$. Extracting a subset of 2×10^4 training patches from a regular grid on the image, we analyze the first 30 iterations of the K-SVD algorithm with $S = 5$ for the training of a $K = 196$ atom dictionary. The initial dictionary is chosen to be a DCT dictionary.

The effects of the DL algorithm are shown in Fig. 6.4. It is clear how the MSE of the representation cost function decreases through the iterations of the K-SVD algorithm. This empirically confirms that the alternating strategy between a sparse coding stage and a dictionary update stage is effectively converging toward at least a local optimum. Furthermore, in order to achieve this, the dictionary is changing the shape of its atoms, and moves from the initially structured DCT dictionary toward one that incorporates new patterns.

We now focus on the implications that this adaptability has for the sparse approximation of the entire image. Assuming that the training set of patches, which is approximately a third of the full set, is a representative collection of the patches in the image, we should see the same improvement in the representation error when comparing the coding errors of both dictionaries. To recover a coded image from patches we average the contribution of overlapping patches and plot the MSE obtained from DCT coding and K-SVD coding with 30 iterations. Results are plotted in Fig. 6.5 for a range of sparsity indices $1 \le S \le 8$. We also show the accuracy of

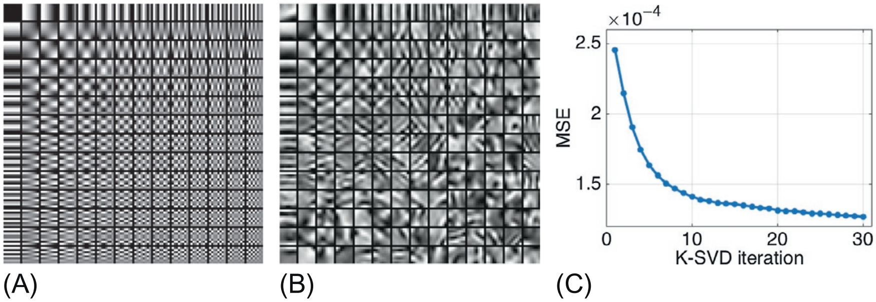

FIG. 6.4

Effect of K-SVD training on dictionary and on training dataset. (A) Initial dictionary. (B) Trained dictionary. (C) Cost function MSE.

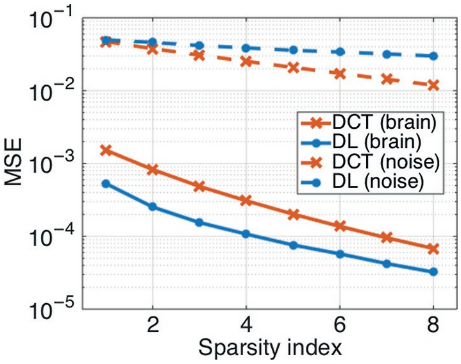

FIG. 6.5

Sparse coding reconstruction MSE of brain MR image. Adapting a dictionary to the brain image reduces the representation error with respect to it, while increasing the error with respect to data of different nature, such as a Gaussian noise image.

recovering an image of only random noise with both dictionaries, to highlight how the lack of structure in random features cannot be well captured by sparse coding.

The gap in representation accuracy represents the gain that can be achieved through DL relative to the initial structured dictionary. This gap is also visible in the error maps of the approximated images for $S = 3$, shown in Fig. 6.6. Notice how most of the representation error concentrates on edges and fine details of the image. This is expected given that those features are precisely the ones that will not conform to the sparsity criterion, and are therefore the first ones to be penalized with the assumption of sparsity.

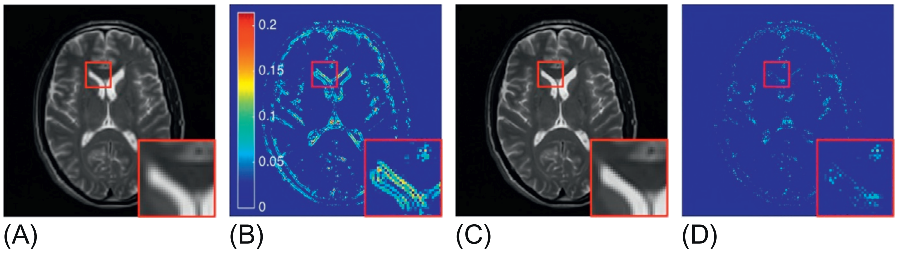

FIG. 6.6

Sparse coded approximations of a brain MR image using a DCT dictionary (A, B) and a trained dictionary (C, D) with a sparsity index $S = 3$. Error maps show absolute value differences with respect to the original image. The region of interest (ROI) in the red rectangle (dark gray in print versions) is enlarged for better visualization. (A) DCT result. (B) DCT error map. (C) DL result. (D) DL error map.

6.4 APPLICATION OF DICTIONARY LEARNING IN MEDICAL IMAGING

6.4.1 DENOISING

The quality of medical images plays an important role in the accuracy of clinical diagnosis. Denoising is a crucial step for improving image quality, which also enhances the ability of subsequent imaging analysis such as detection, segmentation and registration. The main challenge in image denoising is to remove noise corruption while keeping the integrity of relevant image information. One of the most representative methods is a nonlocal means (NLM) filter (Coupé et al., 2008). This method exploits the redundancy of image patterns for noise removal. In particular, similar patches are selected within an image and the restored intensities are the weighted average of the selected patches. Another well-known category of denoising methods utilizes sparse representations of image patterns (Bao et al., 2013). In such methods, a clean signal is assumed to be sparsely represented by a few bases such as sine or cosine functions in FFT or DCT transformations. The noise is then reduced by removing the noise-related coefficients in the transform domain.

More recently, DL has been used for image denoising (Elad and Aharon, 2006). Rather than using standard bases of DCT or FFT transforms, these techniques learn a set of bases from the image to form a dictionary. The learnt dictionary can be used to reduce the noise and restore the clean image. Given a noise image $\mathbf{y}$, the denoising process tries to remove the noise from $\mathbf{y}$ to provide an approximation of the original image $\mathbf{x}$, which can be formulated as

$$\min_{\mathbf{x},\boldsymbol{\Gamma},\mathbf{D}} \quad \|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda\|\mathbf{R}_i\mathbf{x}-\mathbf{D}\boldsymbol{\gamma}_i\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}_i\|_0 \le S \quad \forall i. \tag{6.13}$$

Here $\mathbf{R}_i$ represents a patch extractor at pixel location i and $\boldsymbol{\Gamma}$ collects as columns $\boldsymbol{\gamma}_i$ the sparse coding of image patches $\mathbf{R}_i\mathbf{x}$. The first term in Eq. (6.13) enforces the consistency between image $\mathbf{y}$ and the denoised image $\mathbf{x}$. The second and third terms ensure that every patch $\mathbf{R}_i\mathbf{x}$ in the denoised image can be represented by a linear combination of a few atoms in $\mathbf{D}$. The solution to this equation is nonconvex and can be approximated using an iterative alternating strategy (Elad and Aharon, 2006) by repeating three steps—sparse coding, dictionary update, and the estimation of $\mathbf{x}$. The sparse coding and the dictionary update steps are the same as in the K-SVD algorithms. The denoised image $\mathbf{x}$ is estimated in the third step as

$$\hat{\mathbf{x}} = \frac{\lambda\mathbf{y} + \mathbf{R}_i^T\mathbf{D}\boldsymbol{\gamma}_i}{\lambda\mathbf{I} + \mathbf{R}_i^T\mathbf{R}_i}. \tag{6.14}$$

This denoising process has been successfully applied to remove the Gaussian noise from natural images (Elad and Aharon, 2006) and recently from medical images (Li et al., 2012; Bao et al., 2013). For example, the K-SVD method was used in Patel et al. (2011) to remove the noise from high-angular resolution diffusion images. Another study (Li et al., 2012) performed image denoising by exploiting the local geometrical structure of atoms of the learnt dictionaries, which showed improved performance compared to the K-SVD method. Overall, the advantage of DL methods over standard transforms such as DCT or FFT is that the learnt bases can be better adapted to the images, enabling a sparser representation and a better separation between clean signals and noise. However, it should be mentioned that noise in MR images follows a Rician distribution rather than a Gaussian distribution (Manjón et al., 2010c) and may have spatially varying noise levels (Manjón et al., 2010c). Thus it would be interesting to develop more advanced denoising algorithms based on DL to solve these problems in future work.

6.4.2 RECONSTRUCTION

Reconstruction refers to the processing required to turn raw acquisition measurements into an image. Magnetic resonance imaging (MRI) acquisition is performed in k-space, which is the Fourier description of the MR image. Traditionally, MR acquisition and reconstruction have been governed by Shannon's sampling criterion (Shannon, 1949), where samples acquired at a sampling frequency at least twice as large as the maximum signal frequency content are linearly reconstructed. MR physics impose that samples need to be drawn sequentially and at a limited maximum rate, and trying to satisfy Shannon's criterion will often lead to lengthy acquisition times. This constraint is particularly problematic in dynamic MR such as cardiac cine imaging, where spatial and temporal sampling need to be traded off against each other. This has motivated the exploration of sampling techniques that undersample k-space, introducing aliasing in the image domain, as illustrated in Fig. 6.7.

Assuming $\mathbf{x} \in \mathbb{C}^N$ to be the image of interest and $\mathbf{y} \in \mathbb{C}^M$ its k-space acquisition, an undersampled acquisition can be described as $\mathbf{y} = \mathbf{MFx}$, where

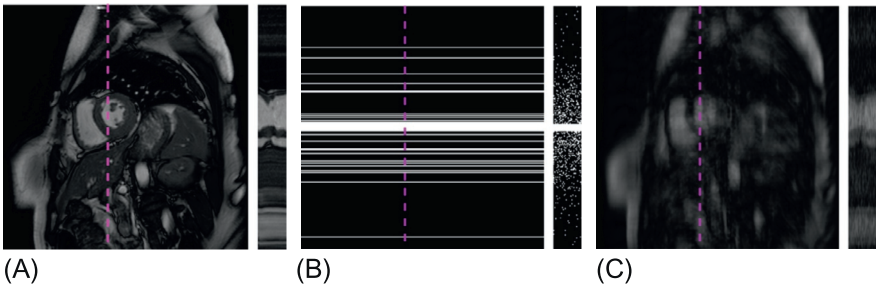

FIG. 6.7

Example of a magnitude temporal frame from a cardiac cine dataset (A). The undersampling mask (B) applied in k-space reduces acquisition time but introduces aliasing in image space (C). All figures show a 2D frame on the left-hand side and the temporal profile across the dashed line on the right-hand side. (A) Fully sampled. (B) Undersampling mask. (C) Zero-filled result.

$\mathbf{M} \in \mathbb{R}^{M\times N}$, $M < N$, is an undersampling mask equivalent to an identity matrix whose rows corresponding to nonacquired k-space locations are missing, and $\mathbf{F}$ is a discrete Fourier transform matrix. This underdetermined system of equations makes it impossible to reconstruct $\mathbf{x}$ from $\mathbf{y}$ unless some prior knowledge about the image is introduced. Compressed sensing (CS) reconstruction methods impose the condition of sparsity in image $\mathbf{x}$, and assuming that image information is redundant they look for a reconstructed image that can be sparsely represented in a transform domain. Assuming $\mathcal{S}(\mathbf{x})$ to be a sparse representation of $\mathbf{x}$, where we enforce a sparsity of S, the global CS MRI problem can be posed as the following combination of a data consistency and a sparsity term:

$$\min_{\mathbf{x}} \quad \|\mathbf{y} - \mathbf{MFx}\|_2^2 \quad \text{s.t.} \quad \|\mathcal{S}(\mathbf{x})\|_0 \leq S. \tag{6.15}$$

CS theory states that the recovery of $\mathbf{x}$ from $\mathbf{y}$ becomes theoretically possible if (a) the undersampling strategy complies with some incoherence criteria, (b) a nonlinear reconstruction method is employed to solve the nonconvex optimization problem, and (c) the image is sufficiently sparse in the transform domain (Lustig et al., 2007). A random subsampling of the Fourier domain sampling of MR can be shown to provide sufficient incoherence for CS reconstruction, while the solution to the optimization problem can be approximated either by greedy methods (Tropp, 2004) or by linear programming after convex relaxation (Candès and Tao, 2005).

A key observation in choosing the sparsity transform is to note that sparser descriptions of the data can better condition the CS problem, hence having better chances of providing an accurate reconstruction. The use of sparsifying transforms usually exploited in compression such as wavelet transforms or transforms assuming piece-wise constant structures like total variation have been proposed many times for CS reconstruction. Some examples of these approaches are presented in Lustig et al. (2007) and Ma and Yin (2008). A common sparse transform for cardiac cine data has been the Fourier transform along the temporal dimension studied, for instance, in Jung et al. (2007) and Gamper et al. (2008), given that it collapses information from slowly varying structures in a few low-frequency coefficients.

Global complete and nonadaptive transforms sometimes do not adhere well to the data and can become crude sparsity models. More recently, the use of redundant and adaptive dictionaries has been proposed to seek better, more flexible sparse models. Assuming patches overlap and the extraction operation wraps around image boundaries, the patch-based dictionary CS problem can be cast as

$$\min_{\mathbf{x},\boldsymbol{\Gamma},\mathbf{D}} \quad \|\mathbf{y} - \mathbf{MFx}\|_2^2 + \lambda\|\mathbf{R}_i\mathbf{x} - \mathbf{D}\boldsymbol{\gamma}_i\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}_i\|_0 \leq S \quad \forall i. \tag{6.16}$$

The solution to this problem looks for an image **x** weighting consistency with the original acquisitions **y** with a sparse and accurate representation of the image provided by the patch-based dictionary **D**.

The content of dictionary atoms can take different forms depending on the data they relate to. They can describe spatial information in the form of 2D or 3D patches, as proposed respectively in Ravishankar and Bresler (2011a) and Song et al. (2013). Spatio-temporal dictionaries have also been explored in Caballero et al. (2014) and Wang and Ying (2014) for the representation of cardiac cine data. One-dimensional atoms have been proposed for the description of temporal magnitude data in Awate and Dibella (2012) as well as quantitative parameter mapping data as discussed in Doneva et al. (2010).

The solution to Eq. (6.16) is nontrivial because it is highly nonconvex. It can nevertheless be approximated by keeping two variables fixed at a time and alternating the solution to the dictionary **D**, the sparse representation **Γ**, and the image **x**. As shown in Ravishankar and Bresler (2011a); Caballero et al. (2014), solving for **D** boils down to the global DL problem, and the solution to **Γ** is given as the solution to the sparse coding problem. Finally, solving for **x** can be posed as a least-squares problem that weights in k-space the original acquisition and the sparse coding image approximation.

Results are reported in Ravishankar and Bresler (2011a) where the dictionary-based approach produces more accurate reconstructions from MR undersampled acquisitions for 2D structural images than the pioneering wavelet-based method from Lustig et al. (2007). Similarly, Fig. 6.8 compares results on cardiac cine data from the dictionary-based method in Caballero et al. (2014) to the method k-t FOCUSS presented in Jung et al. (2007), which exploits sparsity in the temporal Fourier space.

6.4.3 SUPER-RESOLUTION

Resolution enhancement, or super-resolution, is an important tool in medical imaging, particularly in dynamic imaging where movement precludes the acquisition of high-resolution images. For example, in MR imaging of the lungs or heart,

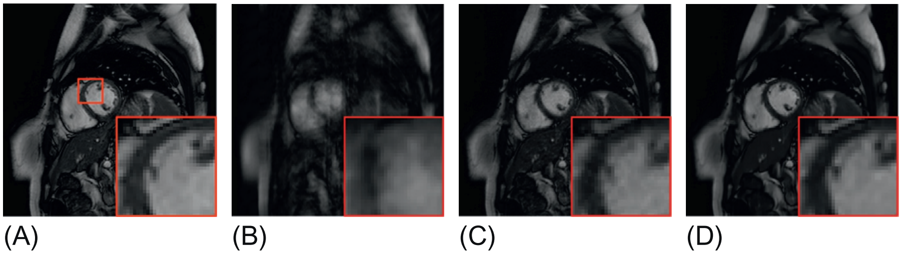

FIG. 6.8

Reconstruction from an eightfold undersampled MRI acquisition. A fully sampled magnitude frame (A), its undersampled by 8 zero-filled version (B), and reconstructions using k-t FOCUSS (C), and the dictionary learning method in Caballero et al. (2014) (D). The ROI in the red rectangle (dark gray in print versions) is enlarged for better visualization. (A) Original. (B) Zero-filled. (C) Temporal Fourier. (D) Dictionary learning.

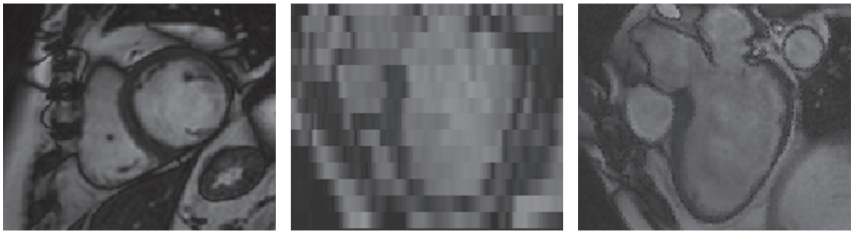

FIG. 6.9

Typical cardiac image acquisitions. Left to right: High-resolution short-axis slice, long-axis stack of short-axis slices showing low resolution in slice-select direction, frame of high-resolution long-axis slice sequence.

motion due to the respiratory or cardiac cycles forces a compromise between spatial resolution, temporal resolution, and volumetric coverage. A typical cardiac acquisition consists of imaging a number of anisotropic 2D slabs over time (Fig. 6.9). Improving the resolution of these acquired sequences is strongly motivated both by clinical visualization and analysis.

The super-resolution problem can be formulated as an inverse problem where it is assumed that the low-resolution (LR) image is a blurred and downsampled version of the original high-resolution (HR) image. Given an underlying unknown HR image $\mathbf{y}_H$, the acquired LR image $\mathbf{y}_L$ is modeled as

$$\mathbf{y}_L = (\mathbf{y}_H * \mathbf{B}) \downarrow s + \eta, \tag{6.17}$$

where $\mathbf{B}$ represents a blur operator, $\downarrow s$ is a downsampling operator that decreases the resolution by a factor of s, and η represents an additive noise term. Recovering the high-resolution image $\mathbf{y}_H$ from $\mathbf{y}_L$ is underdetermined and requires regularization or prior information on the nature of $\mathbf{y}_H$. Conventional super-resolution techniques provide this by fusing several views of the same object aligned with subpixel (or voxel in 3D) accuracy to constrain the solution. While this works well for objects which are deformed only by rigid or affine motion (such as the fetal brain (Gholipour et al., 2010)), and where accurate deformation can therefore be easily recovered, when the motion is nonrigid, such as that due to cardiac or respiratory dynamics, obtaining subpixel alignment becomes difficult.

This motivates the use of *example-based* or *hallucination* super-resolution methods. These avoid the need for accurate alignment by moving the analysis to a small patch scale. The aim is to upsample LR image patches using knowledge of the relationship between HR and LR features gained from example training data as in Manjón et al. (2010a,b), and Rousseau (2010). Such a relationship can be encapsulated using a pair of *correlated* (or *coupled*) HR and LR dictionaries. Central to this is how to train these dictionaries such that corresponding HR and LR patches have the same sparse representation with respect to their dictionary in both cases. The use of DL effectively enforces the prior on the reconstruction that signals such as image patches can be represented by a sparse combination of dictionary atoms.

The construction of correlated dictionaries for single-image super-resolution was proposed by Zeyde et al. (2012) and applied directly to single-image brain MRI enhancement in Rueda et al. (2013). However, a real benefit in medical applications is to improve resolution by combining several low-resolution views of a structure. This has been developed in Bhatia et al. (2014) using the same dictionary construction method. It is described in the following, with particular adaptation to cardiac image sequence enhancement.

Stack sequences such as those in Fig. 6.1 can be acquired; while HR is available only in two dimensions, orthogonal HR planes can be acquired and the information combined for super-resolution. For simplicity, we describe an upsampling of a stack which is HR in the cardiac short-axis (SA) only, using training data from a stack which is HR in the cardiac long-axis (LA) only. Training data consists of frames from an HR LA sequence $Y_H = \{\mathbf{y}_H\}$, and corresponding LR sequences obtained through blurring and downsampling according to Eq. (6.17). A Gaussian kernel, with full width at half maximum (FWHM) equal to the slice thickness, is used to blur in the slice-select direction only, in order to simulate the acquisition process (Greenspan, 2009). The resulting LR images are then upsampled using bicubic interpolation back to the original size, to avoid complications arising from differently sized patches, giving $Y_L = \{\mathbf{y}_L\}$. Patches at corresponding locations $P = \{\mathbf{p}_k^H, \mathbf{p}_k^L\}_k$ are extracted from these images: $\mathbf{p}_k = \mathbf{R}_k\mathbf{y}$, where $\mathbf{R}_k$ is an operator to extract a patch of size of size $n \times m$ from location k. These are used to co-train the HR and LR dictionaries, as shown diagrammatically in Fig. 6.10.

The aim of constructing correlated dictionaries is to ensure that an HR patch and its LR counterpart have the same sparse representations with respect to their individual dictionaries. It is also necessary to ensure that the LR patches can

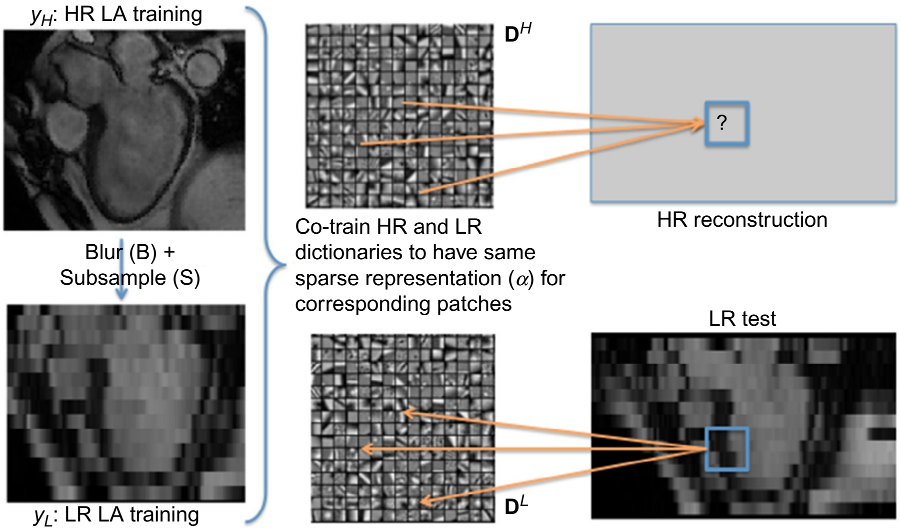

FIG. 6.10

High-resolution and low-resolution dictionaries can be co-trained such that corresponding high-resolution and low-resolution patches have the same sparse representation in each of their respective dictionaries.

be encoded sparsely *and in the same way* for both train and test data. When reconstructing an LR test patch, we find the sparse representation of that patch in terms of the LR dictionary only. The LR dictionary $\mathbf{D}^L$ is therefore also constructed using LR patches only:

$$\mathbf{D}^L, \boldsymbol{\gamma} = \arg \min_{\mathbf{D}^L, \boldsymbol{\gamma}_k} \sum_k \|\mathbf{p}_k^L - \mathbf{D}^L \boldsymbol{\gamma}_k\|_2^2 \text{ s.t.} \|\boldsymbol{\gamma}\|_0 < S, \tag{6.18}$$

where S denotes the desired sparsity of the reconstruction weights vector $\boldsymbol{\gamma}$. This is a standard DL equation which is solved sequentially for $\mathbf{D}^L$ and $\boldsymbol{\gamma}$ using K-SVD and OMP respectively. The resulting sparse code $\boldsymbol{\gamma}_k$ for each patch k is then used to solve for the HR dictionary by minimizing the reconstruction error of the HR training patches:

$$\mathbf{D}^H = \arg \min_{D_H} \sum_k \|\mathbf{p}_k^H - \mathbf{D}^H \boldsymbol{\gamma}_k\|_2^2 = \arg \min_{\mathbf{D}^H} \|\mathbf{P}^H - \mathbf{D}^H \mathbf{A}\|_F^2, \tag{6.19}$$

where the columns of $\mathbf{P}$ are formed by the high-resolution training patches, $\mathbf{p}_k^H$ and the columns of $\mathbf{A}$ are formed by the atoms $\boldsymbol{\gamma}$. The solution is given by $\mathbf{D}^H = \mathbf{P}^H A^+ = \mathbf{P}^H \mathbf{A}^T (\mathbf{A}\mathbf{A}^T)^{-1}$ where $\mathbf{A}^+$ is the Pseudo-Inverse, noting that $\mathbf{A}$ has full row rank. This results in two correlated dictionaries, $\mathbf{D}^H$ and $\mathbf{D}^L$, which can be used to reconstruct HR patches from LR input.

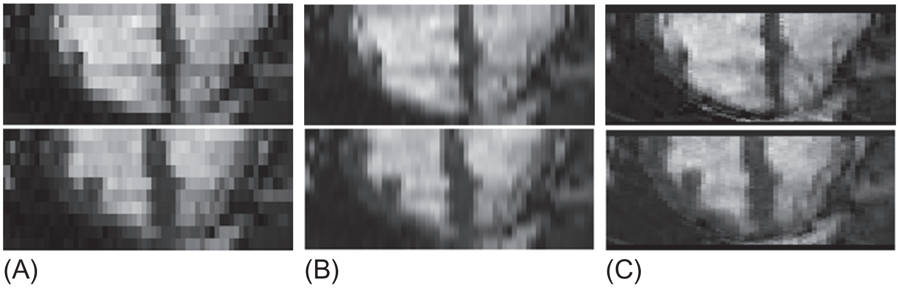

FIG. 6.11

Adult cardiac data; reconstruction using 30 high-resolution long-axis frames. (A) Original. (B) Bicubic. (C) Dictionary learning.

At the reconstruction stage, patches from an LR test image are extracted in the same way as for the LR training images. The sparse code for each patch with respect to the LR dictionary, $\mathbf{D}_L$, is found by solving:

$$\boldsymbol{\gamma}_k = \arg\min_{\boldsymbol{\gamma}_k} \sum_k \|p_k^L - \mathbf{D}^L \boldsymbol{\gamma}_k\|_2^2 \text{ s.t.} \|\boldsymbol{\gamma}_k\|_0 < S \tag{6.20}$$

using OMP. Crucially, this is the same sparse coding equation as in the training phase. The reconstruction weights vectors $\boldsymbol{\gamma}_k$ for each test patch are used to approximate high-resolution patches by $\{\tilde{\mathbf{p}}_k^H\}_k = \{\mathbf{D}^H \boldsymbol{\gamma}_k\}_k$. To create a smooth overall reconstruction, overlapping patches can be used, and the upsampled image given by the average of the overlapping reconstructed patches.

Example reconstructions for two adult subjects are shown in Fig. 6.11. Here, 30-frame sSA stack acquisitions are reconstructed to isotropic using an orthogonal HR LA stack sequence to train the dictionaries. This represents an upsampling factor of 8. Patch sizes of 12 × 24 pixels (6 × 12 pixel overlap) were used in both training and reconstruction. Training took less than 8],s, while reconstruction of all SA stack frames took under 2 s, using MATLAB on a 2.4-GHz Intel Core i5 machine in both cases. It can be clearly seen that in both cases the patch-based reconstruction algorithm produces sharper images than the standard interpolation.

6.4.4 SEGMENTATION

In previous sections, we have shown that DL is well suited for restoration tasks such as reconstruction and super-resolution. However, they cannot be directly used for image classification tasks because the learnt dictionaries only have the reconstructive power. There have been multiple attempts to learn discriminative dictionaries by using the class labels of training samples. One straightforward way is to use the training samples themselves as the dictionary without learning, in which case each atom in the dictionary has a training label. A testing sample can be represented by the training samples of all classes and be classified to the group that leads to the minimal reconstruction error. The so-called sparse representation-based classification (SRC) scheme has been successfully applied to face recognition (Wright et al., 2009) and also in medical image segmentation (Tong et al., 2013; Wang et al., 2014). For example, in Wang et al. (2014), the class labels of training patches were propagated to test patches via SRC. Specifically, all the training patches were first formed as a dictionary. The test patches can then be represented by this predefined dictionary. Since the class labels of atoms in the predefined dictionary are available, the class labels of test patches are estimated by using the coding coefficients as weights in a label fusion process. However, the noisy information in the training patches may make the segmentation less effective and the complexity of sparse coding can be very high when there are a large number of training patches. This problem can be addressed by properly learning a discriminative dictionary from the original training patches.

There are two main categories of discriminative DL methods. In the first category, dictionaries are learnt for different classes separately. These class-specific dictionaries are used to represent the test patches. The representation residual associated with each class can then be used to do classification. For example, Huang et al. (2014) proposed to learn a pair of dictionaries for the foreground and background training patches and used the representation residuals as new features for segmentation. Another category of discriminative DL methods learns a shared dictionary by all classes while training a classifier over the representation coefficients at the same time. In Tong et al. (2013) this type of discriminative DL methods has been successfully applied to the segmentation of the hippocampus in brain MR images. In the following we will describe this segmentation method, which is known as discriminative dictionary learning for segmentation (DDLS) and is illustrated in Fig. 6.12.

Given a set of training patches $\boldsymbol{P_L} = [p_1, p_2, \ldots, p_n] \in \mathbb{R}^{M \times N}$ with class labels, the segmentation process is to assign a class label for each test patch p_t. A reconstructive dictionary $\boldsymbol{D}$ with K atoms can be learnt from the training patches by solving the following problem:

$$\min_{\boldsymbol{D},\boldsymbol{\Gamma}} \|\boldsymbol{P_L} - \boldsymbol{D}\boldsymbol{\Gamma}\|_F^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}_i\|_0 \leq S \ \ \forall i. \tag{6.21}$$

Here the objective function includes the reconstruction error term and the sparsity constraint term without considering the discriminative power. Thus the learnt dictionary is not suitable for the segmentation task. To address this problem, a linear classifier was added to the objective function for learning dictionaries with both reconstructive and discriminative power. The objective function is then formulated as

$$\min_{\boldsymbol{D},\boldsymbol{W},\boldsymbol{\Gamma}} \|\boldsymbol{P_L} - \boldsymbol{D}\boldsymbol{\Gamma}\|_F^2 + \beta \|\boldsymbol{H} - \boldsymbol{W}\boldsymbol{\Gamma}\|_F^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}_i\|_0 \leq S \ \ \forall i. \tag{6.22}$$

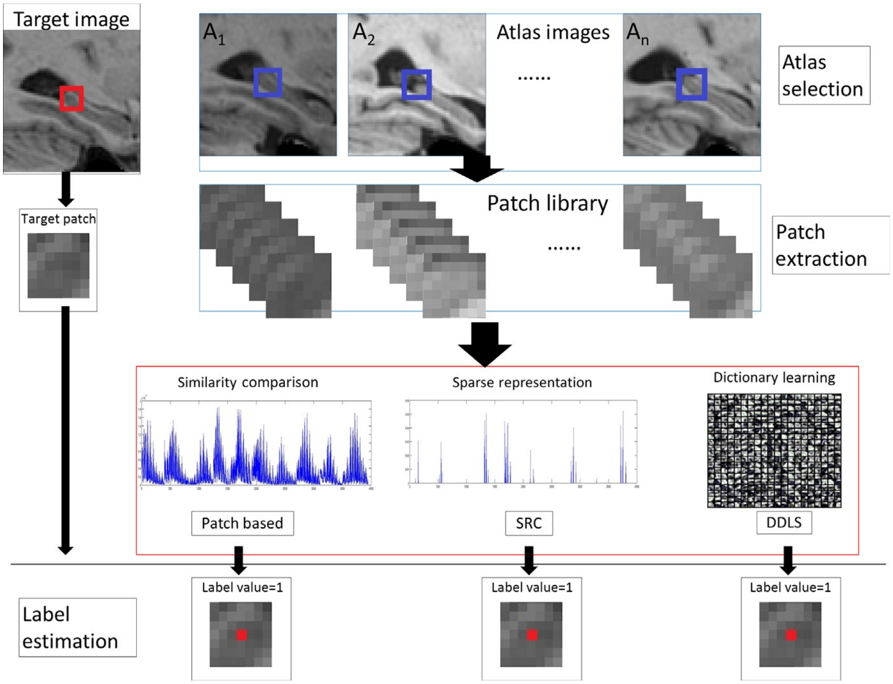

FIG. 6.12

Flow chart of labeling one target voxel by three different methods: Patch-based Labeling, Sparse Representation Classification (SRC) and Discriminative Dictionary Learning for Segmentation (DDLS). The red box (dark gray in print versions) in the target image represents the target patch. The blue boxes in atlas images represent the search volume area for extracting template patches.

Here the classification error term $\|\boldsymbol{H} - \boldsymbol{W\Gamma}\|_F^2$ is added to Eq. (6.21). $\boldsymbol{H}$ represents the labels of the central voxels of the patches in $\boldsymbol{P_L}$. Each column of $\boldsymbol{H}$ is a label vector corresponding to a training patch. Each label vector is defined as $h_i = [0, 0 \ldots 1 \ldots 0, 0]$, where the nonzero entry position indicates the label of the center voxel of the corresponding patch. $\boldsymbol{W}$ denotes the linear classifier parameters and β controls the trade-off between the reconstruction error term and the classification error term. In this testing stage, the learnt dictionary $\boldsymbol{D}$ can be used to represent the test patch while the learnt classifier $\boldsymbol{W}$ can be used for estimating a class label for the test patch.

The DDLS segmentation method has been successfully applied to MR hippocampal segmentation and multiorgan segmentation in abdominal CT images (Tong et al., 2015). Overall, the DDLS segmentation approach not only yields competitive segmentation accuracy but also can be implemented very efficiently.

However, the method was only evaluated on the segmentation of healthy structures. Another recent study in Weiss et al. (2013) utilized DL for the segmentation of multiple sclerosis lesions. The dictionary was learnt using patches from healthy regions, which is expected to reconstruct badly in lesion regions and yield large reconstruction errors. The final segmentation of the lesion was achieved by thresholding the reconstruction error map.

6.5 FUTURE DIRECTIONS

Although DL has been successfully applied in a large number of image processing tasks, there is still room for many improvements. Some immediate open questions are how to choose from the available DL algorithms and the selection of hyperparameters such as the number of atoms, the sparsity level, and the number of iterations for guaranteeing convergence. In general, there is no definite answer for choosing the optimal number of atoms or the best sparsity level in DL. The choices of these parameters are often made heuristically. In addition to learning the best dictionary for a specific application, there are several other aspects which have not yet been fully exploited for applications in medical imaging.

First, most of the DL techniques use patches at a single scale for learning dictionaries. However, in some applications, it is helpful to use patches at multiple scales for learning more efficient dictionaries. Local patches at coarser scales can provide global structural information (also called anatomical patterns in medical imaging) while patches at finer scales present important local appearance information such as intensity patterns. Learning dictionaries at multiple scales can fully utilize both the global anatomical information and local appearance information, which has been shown to be useful in image segmentations (Huang et al., 2014). However, multiscale DL approaches proposed, such as those by Mairal et al. (2007) and Ophir et al. (2011) have not been fully investigated in medical imaging. One example showing that multiscale DL could be beneficial is shown in Ravishankar and Bresler (2011b), where it is used for MR image reconstruction.

Another interesting research avenue is multimodal DL. The applications described in this chapter are based on images from a single modality, but it is possible to learn information about medical images across modalities using DL, as shown in Cao et al. (2013). For instance, learning a dictionary in a given modality to improve the reconstruction quality of a different modality, where data acquisition could be challenging, can be viable, as demonstrated in Tosic et al. (2010).

Exploiting structures in data is crucial for the success of many machine learning techniques. By adopting l_0-norm or l_1-norm for regularization, sparsity is achieved by treating each atom individually in DL techniques, regardless of its position in the dictionary or relations with other atoms. Therefore the relations and the structures between the atoms are ignored by just using the l_0-norm or l_1-norm constraint. However, this type of information could be helpful in many applications. Structured sparse coding (Elhamifar and Vidal, 2011) and structured DL (Szabó et al., 2011) have recently been introduced to encode these specific relations and structures in machine learning, which could be useful in many applications of medical image analysis.

To conclude, the description of DL throughout the chapter has been based on the sparse synthesis model, although there exist other types of signal modeling encompassed by DL such as analysis modeling (Rubinstein et al., 2013). Moreover, much research is still needed to better understand the low-dimensional spaces where medical images could potentially be efficiently represented. Although sparse coding and DL have been successfully applied in many applications, the performance of a linear model to represent medical images is limited for signals where nonlinearities can occur, such as in medical acquisition mechanisms.

6.6 CONCLUSION

In this chapter, we have provided an overview of DL and its application in medical imaging including image denoising, reconstruction, super-resolution, and segmentation. DL has been demonstrated to be very effective in these applications, showing its powerful ability in medical image analysis. In future, it would be very interesting to investigate the use of DL in new applications in medical imaging. However, it should be mentioned that identifying the appropriate applications is important as DL might not be able to provide a good solution to some specific medical image analysis problems.

REFERENCES

Aharon, M., Elad, M., Bruckstein, A., 2006. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. 54 (11), 4311–4322.

Awate, S.P., Dibella, E.V.R., 2012. Spatiotemporal dictionary learning for undersampled dynamic MRI reconstruction via joint frame-based and dictionary-based sparsity. In: Proceedings of the IEEE International Symposium on Biomedical Imaging, pp. 318–321.

Bao, L., Robini, M., Liu, W., Zhu, Y., 2013. Structure-adaptive sparse denoising for diffusion-tensor MRI. Med. Image Anal. 17 (4), 442–457.

Beylkin, G., 1992. On the representation of operators in bases of compactly supported wavelets. SIAM J. Numer. Anal. 29 (6), 1716–1740.

Bhatia, K.K., Price, A.N., Shi, W., Hajnal, J.V., Rueckert, D., 2014. Super-resolution reconstruction of cardiac MRI using coupled dictionary learning. In: IEEE International Symposium on Biomedical Imaging. IEEE, Piscataway, NJ, pp. 947–950.

Bioucas-Dias, J.M., Figueiredo, M.A., 2008. An iterative algorithm for linear inverse problems with compound regularizers. In: International Conference on Image Processing, IEEE, Piscataway, NJ, pp. 685–688.

Caballero, J., Price, A.N., Rueckert, D., Hajnal, J.V., 2014. Dictionary learning and time sparsity for dynamic MR data reconstruction. IEEE Trans. Med. Imaging 33 (4), 979–994.

Candès, E.J., Donoho, D.L., 1999. Curvelets: a surprisingly effective nonadaptive representation for objects with edges. In: International Conference on Curves and Surfaces, vol. 2, pp. 105–120.

Candès, E.J., Tao, T., 2005. Decoding by linear programming. IEEE Trans. Inform. Theory 51 (12), 4203–4215.

Cao, T., Jojic, V., Modla, S., Powell, D., Czymmek, K., Niethammer, M., 2013. Robust multimodal dictionary learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, New York, pp. 259–266.

Chen, S., Donoho, D., Saunders, M., 1998. Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. 20 (1), 33–61.

Coifman, R.R., Meyer, Y., Wickerhauser, V., 1992. Wavelet analysis and signal processing. Wavelets Their Appl. 153–178.

Cooley, J.W., Tukey, J.W., 1965. An algorithm for the machine calculation of complex Fourier series. Math. Comput. 19 (90), 297–301.

Coupé, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., Barillot, C., 2008. An optimized blockwise nonlocal means denoising filter for 3D magnetic resonance images. IEEE Trans. Med. Imaging 27 (4), 425–441.

Do, M.N., Vetterli, M., 2002. Contourlets : a directional multiresolution image representation. Image Process. 1, 357–360.

Doneva, M., Börnert, P., Eggers, H., Stehning, C., Sénégas, J., Mertins, A., 2010. Compressed sensing reconstruction for magnetic resonance parameter mapping. Magn. Reson. Med. 64 (4), 1114–1120.

Donoho, D.L., Tsaig, Y., Drori, I., Starck, J.L., 2012. Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. IEEE Trans. Inform. Theory 58 (2), 1094–1121.

Elad, M., 2010. Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer, New York.

Elad, M., Aharon, M., 2006. Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans. Image Process. 15 (12), 3736–3745.

Elhamifar, E., Vidal, R., 2011. Robust classification using structured sparse representation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1873–1879.

Engan, K., Aase, S.O., Husoy, J., 1999. Frame based signal compression using method of optimal directions (MOD). In: IEEE International Symposium on Circuits and Systems, vol. 4, pp. 1–4.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al., 2007. Pathwise coordinate optimization. Ann. Appl. Stat. 1 (2), 302–332.

Gabor, D., 1946. Theory of communication. J. Inst. Elect. Eng. 93 (26), 429–441.

Gamper, U., Boesiger, P., Kozerke, S., 2008. Compressed sensing in dynamic MRI. Magn. Reson. Med. 59 (2), 365–373.

Gholipour, A., Estroff, J.A., Warfield, S.K., 2010. Robust super-resolution volume reconstruction from slice-acquisitions: application to fetal brain MRI. IEEE Trans. Med. Imaging 29 (10), 1739–1758.

Greenspan, H., 2009. Super-resolution in medical imaging. Comput. J. 52(1), 43–63.

Grossmann, A., Morlet, J., 1984. Decomposition of Hardy functions into square integrable wavelets of constant shape. SIAM J. Math. Anal. 15 (4), 723–736.

Huang, X., Dione, D.P., Compas, C.B., Papademetris, X., Lin, B.A., Bregasi, A., Sinusas, A.J., Staib, L.H., Duncan, J.S., 2014. Contour tracking in echocardiographic sequences via sparse representation and dictionary learning. Med. Image Anal. 18 (2), 253–271.

Jolliffe, I., 2005. Principal Component Analysis. Wiley, Chichester.

Jung, H., Ye, J.C., Kim, E.Y., 2007. Improved k-t BLAST and k-t SENSE using FOCUSS. Phys. Med. Biol. 52 (11), 3201–3226.

Le Pennec, E., Mallat, S., 2005. Sparse geometric image representations with bandelets. IEEE Trans. Image Process. 14 (4), 423–438.

Li, S., Yin, H., Fang, L., 2012. Group-sparse representation with dictionary learning for medical image denoising and fusion. IEEE Trans. Biomed. Eng. 59 (12), 3450–3459.

Lustig, M., Donoho, D.L., Pauly, J.M., 2007. Sparse MRI: the application of compressed sensing for rapid MR imaging. Magn. Reson. Med. 58 (6), 1182–1195.

Ma, S., Yin, W., 2008. An efficient algorithm for compressed MR imaging using total variation and wavelets. In: IEEE Conference on Computer Vision and Pattern Recognition 2008, pp. 1–8.

Mairal, J., Sapiro, G., Elad, M., 2007. Learning multiscale sparse representations for image and video restoration. Technical Report, DTIC Document.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., 2009. Online dictionary learning for sparse coding. In: Proceedings of the 26th Annual International Conference on Machine Learning. ACM, pp. 689–696.

Mallat, S.G., 1999. A Wavelet Tour of Signal Processing. Academic Press, San Diego, CA.

Mallat, S.G., Zhang, Z., 1993. Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process. 41 (12), 3397–3415.

Manjón, J.V., Coupé, P., Buades, A., Collins, D.L., Robles, M., 2010a. MRI superresolution using self-similarity and image priors. J. Biomed. Imaging 2010 (17), 1–11.

Manjón, J.V., Coupé, P., Buades, A., Fonov, V., Collins, D.L., Robles, M., 2010b. Non-local MRI upsampling. Med. Image Anal. 14 (6), 784–792.

Manjón, J.V., Coupé, P., Martí-Bonmatí, L., Collins, D.L., Robles, M., 2010c. Adaptive non-local means denoising of MR images with spatially varying noise levels. J. Mag. Reson. Imaging 31 (1), 192–203.

Needell, D., Tropp, J.A., 2009. CoSaMP: iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmon. Anal. 26 (3), 301–321.

Needell, D., Vershynin, R., 2010. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. IEEE J. Sel. Top. Signal Process. 4 (2), 310–316.

Ophir, B., Lustig, M., Elad, M., 2011. Multi-scale dictionary learning using wavelets. IEEE J. Sel. Top. Signal Process. 5 (5), 1014–1024.

Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W., 2005. An iterative regularization method for total variation-based image restoration. Multiscale Model. Simul. 4 (2), 460–489.

Patel, V., Shi, Y., Thompson, P.M., Toga, A.W., 2011. K-SVD for Hardi denoising. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1805–1808.

Ravishankar, S., Bresler, Y., 2011a. MR image reconstruction from highly undersampled k-space data by dictionary learning. IEEE Trans. Med. Imaging 30 (5), 1028–41.

Ravishankar, S., Bresler, Y., 2011b. Multiscale dictionary learning for MRI. In: Proceedings of ISMRM, p. 2830.

Rousseau, F., 2010. A non-local approach for image super-resolution using intermodality priors. Med. Image Anal. 14, 594–605.

Rubinstein, R., Zibulevsky, M., Elad, M., 2008. Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit. CS Technion 40 (8), 1–15.

Rubinstein, R., Peleg, T., Elad, M., 2013. Analysis K-SVD: a dictionary-learning algorithm for the analysis sparse model. IEEE Trans. Signal Process. 61 (3), 661–677.

Rueda, A., Malpica, N., Romero, E., 2013. Single-image super-resolution of brain MR images using overcomplete dictionaries. Med. Image Anal. 17(1),113–132.

Shannon, C.E., 1949. Communication in the presence noise. Proc. IRE 37 (1), 10–21.

Simoncelli, E.P., Freeman, W.T., Adelson, E.H., Heeger, D.J., 1992. Shiftable multiscale transforms. IEEE Trans. Inform. Theory 38 (2), 587–607.

Skodras, A., Christopoulos, C., Ebrahimi, T., 2001. The JPEG 2000 still image compression standard. IEEE Signal Process. Mag. 18 (5), 36–58.

Song, Y., Zhu, Z., Lu, Y., Liu, Q., Zhao, J., 2013. Reconstruction of magnetic resonance imaging by three-dimensional dual-dictionary learning. Magn. Reson. Med. 71 (3), 1285–1298.

Szabó, Z., Póczos, B., Lorincz, A., 2011. Online group-structured dictionary learning. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Piscataway, NJ, pp. 2865–2872.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B 267–288.

Tong, T., Wolz, R., Coupé, P., Hajnal, J.V., Rueckert, D., Initiative, A.D.N., et al., 2013. Segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. NeuroImage 76, 11–23.

Tong, T., Wolz, R., Wang, Z., Gao, Q., Misawa, K., Fujiwara, M., Mori, K., Hajnal, J.V., Rueckert, D., 2015. Discriminative dictionary learning for abdominal multi-organ segmentation. Med. Image Anal. 23 (1), 92–104.

Tosic, I., Frossard, P., 2011. Dictionary learning. IEEE Signal Process. Mag. 28 (2), 27–38.

Tosic, I., Jovanovic, I., Frossard, P., Vetterli, M., Duric, N., 2010. Ultrasound tomography with learned dictionaries. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 5502–5505.

Tropp, J., 2004. Greed is good: algorithmic results for sparse approximation. IEEE Trans. Inform. Theory 50 (10), 2231–2242.

Wallace, G., 1992. The JPEG still picture compression standard. IEEE Trans. Consumer Electron. 38 (1), xviii–xxxiv.

Wang, Y., Ying, L., 2014. Compressed sensing dynamic cardiac cine MRI using learned spatiotemporal dictionary. IEEE Trans. Biomed. Eng. 61 (4), 1109–1120.

Wang, L., Shi, F., Gao, Y., Li, G., Gilmore, J.H., Lin, W., Shen, D., 2014. Integration of sparse multi-modality representation and anatomical constraint for isointense infant brain MR image segmentation. NeuroImage 89, 152–164.

Weiss, N., Rueckert, D., Rao, A., 2013. Multiple sclerosis lesion segmentation using dictionary learning and sparse coding. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, New York, pp. 735–742.

Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y., 2009. Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 31 (2), 210–227.

Zeyde, R., Elad, M., Protter, M., 2012. On single image scale-up using sparse-representations. In: International Conference on Curves and Survaces. Springer, New York, pp. 711–730.

GLOSSARY

CS	compressed sensing
DCT	discrete cosine transform
DL	dictionary learning
MR	magnetic resonance
MRI	magnetic resonance imaging
MSE	mean squared error
OMP	orthogonal matching pursuit