ENTROPY CONSTRAINED DICTIONARY LEARNING FOR REMOTE SENSING IMAGE COMPRESSION

Xin Zhan, Rong Zhang, Dong Yin, Kai Chen
Department of Electronic Engineering and Information Science
University of Science and Technology of China, Hefei, China, 230027
zhanxin@mail.ustc.edu.cn, zrong@ustc.edu.cn, yindong@ustc.edu.cn

ABSTRACT
This paper presents a new compression scheme for remote sensing image, the texture of which is much richer than that of natural image. The last decade has seen a growing interest in the study of dictionary learning and sparse representation, which have been proven to perform well on image compression. Compared with DCT or DWT for compression, the advantage of sparse representation is that the coefficient matrix is sparser with less non-zero elements. However, the non-zero positions in the sparse coefficient matrix are almost random and thus not conducive to entropy coding. As the content of remote sensing image is usually very complex and the position indices cannot be quantized, the coding cost of these non-zero positions becomes the bottleneck of compression. In this paper, an entropy-constrained dictionary learning algorithm is introduced to make the coefficient matrix more structured and adapted for entropy coding. The experimental results reveal that the proposed method reduces the coding cost of the position indices significantly and achieves better results than the JPEG2000 standard.

KEY WORDS
Compression, image coding, remote sensing, sparse representation.

1. Introduction
Remote sensing images are very important data sources for widespread applications. However, the data volume of remote sensing images is so huge that it makes storage and communication very difficult, and thus compression of the image data is necessary.

Most image compression algorithms rely on appropriate signal representation, like the discrete cosine transform (DCT) for JPEG, or the discrete wavelet transform (DWT) for JPEG2000 [1]. Recent years have witnessed a growing interest in the research for sparse representation of signals [2]. Instead of using a fixed transformation based on a mathematical model, an over-complete dictionary is learned from a training set and sparse coding is applied to decompose the signal into a linear combination of a few atoms from the dictionary. These new theories have been given considerable attention in the area of image processing applications like compression, classification, and automatic target recognition [3]-[5]. The advantages and disadvantages of using learned dictionaries for image compression are analyzed in [5]. It pointed out that JPEG, SPIHT, and JPEG2000 exploit the structure in the quantized coefficients in their advanced entropy coding schemes, like the zig-zag structure and the zero-tree structure for DCT and DWT coefficients, respectively. However, the structure in the sparse coefficient matrix is lost or hidden, i.e., the non-zero positions are almost random. In this work, an entropy regularization is integrated into the sparse coding and dictionary learning procedure, which makes the indices of the sparse coefficients more structured and adapted for entropy coding.

The rest of this paper is organized as follows. In Section 2, we introduce sparse coding, dictionary learning, and how to use them for image coding. Section 3 describes the proposed entropy-constrained compression method. The experimental results are presented in Section 4 and this paper concludes in Section 5.

2. Review of Sparse Representation for Compression
Signal representation is a basic problem in the field of image compression. Sparse representation with respect to an over-complete dictionary has been studied with growing interest recently. Given a signal $b \in \mathbb{R}^N$ (typically an $8 \times 8$ image patch from pixel or wavelet domain), the sparse representation of $b$ is expressed as

$$b = Dw + e \quad (1)$$

where $D \in \mathbb{R}^{N \times K}$ is the dictionary matrix, $e$ is the approximate error, and $w \in \mathbb{R}^K$ denotes the coefficient vector. As can be seen, there are two unknowns, $w$ and $D$, that must be handled. In the following subsections, we first review the sparse coding method that find $w$ with a given dictionary $D$ and the dictionary learning method that find a proper $D$ from a training set, and then introduce the image compression scheme based on sparse representation.

2.1 Orthogonal Matching Pursuit for Sparse Coding
Suppose we have signal $b$ and dictionary $D$, sparse coding aim at seeking approximations of $b$ with as few...
atoms in $\mathbf{D}$ as possible by solving the minimization problem [2]:

$$\min_{\mathbf{w}} ||\mathbf{b} - \mathbf{Dw}||_2^2 \text{ s.t. } ||\mathbf{w}||_0 \leq n_0. \quad (2)$$

We use $||\cdot||_0$ to denote the $\ell_0$ pseudo-norm, which is the number of non-zero entries of a vector. The resulted $\mathbf{w}$ has $n_0$ or fewer non-zero entries. This problem is known to be NP-hard and many algorithms, such as orthogonal matching pursuit (OMP) [6], have been developed to solve it efficiently. In each iteration step of OMP, one atom that reduces the residual most is selected, and then the support (position indices of non-zero entries) and the projection values are updated. This process is repeated until the number of non-zero entries reaches the given upper limit.

2.2 Dictionary Learning

Compared with orthogonal transforms, the learned dictionary is well adapted to its purpose, i.e., sparse representation of a specific class of signals. Given a training set $\mathbf{B} = \{\mathbf{b}_i \in \mathbb{R}^N\}_{i=1}^M$, we need to find the optimal dictionary $\mathbf{D} \in \mathbb{R}^{N \times K}$ by solving the following optimization problem [2]:

$$\min_{\mathbf{D},\{\mathbf{b}_i\}_{i=1}^M} \frac{1}{M} \sum_{i=1}^M ||\mathbf{b}_i - \mathbf{Dw}_i||_2^2 \text{ s.t. } ||\mathbf{w}_i||_0 \leq n_0, 1 \leq i \leq M. \quad (3)$$

This problem aims to jointly find the proper representations and the dictionary. It can also be viewed as a nested minimization problem: an inner minimization of the number of non-zeros in the representation vectors $\mathbf{w}_i$ for a given fixed $\mathbf{D}$, and an outer minimization over $\mathbf{D}$. A strategy of alternating minimization thus seems natural: at the $k$-th step, we use the dictionary $\mathbf{D}^{k-1}$ from the $(k-1)$-th step and solve $M$ instances of sparse coding problem, one for each training sample $\mathbf{b}_i$. This gives us the coefficient matrix $\mathbf{W}^k = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M]$ and we then update the dictionary assuming $\mathbf{W}^k$ is fixed.

Various methods have been proposed to successively update the dictionary $\mathbf{D}^{k-1}$ when $\mathbf{W}^k$ is obtained. Method of optimal directions (MOD) [7] updates the dictionary by global least squares fit and the K-SVD algorithm [8] relies on singular value decomposition. The recursive least squares (RLS) dictionary learning algorithm provided in [9], which is employed in this work, achieves the best performance in comparison with the others. More details about RLS can be found in [9].

2.3 Image Compression using Sparse Representation

Sparse representation-based image compression scheme consists of three parts [5]: the training, sparse coding, and quantizing/entropy coding parts. The detailed descriptions are given in the following.

- Training part for training images. A dictionary learning algorithm, e.g., RLS with OMP, is applied to learn a dictionary $\mathbf{D}$ from the set of patches extracted from the training images in the pixel domain or the wavelet domain. In [5] and [9], it has been shown that training dictionaries from the wavelet domain always gives better compression performance. In this paper, we first transform the images by using a three-level two-dimensional 9/7 wavelet transformation, and then form the $8 \times 8$ training patches in the coefficient domain. Note that the dictionary will not be included in the compressed bit stream, as it is trained offline by using a training set and supposed to be a general dictionary for a particular remote sensing sensor.

- Sparse coding part for test images. The test image to be compressed is sliced into non-overlapping patches in the wavelet coefficient domain. These patches are denoted as $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^L$. For each $\mathbf{x}_i$, sparse coding method, e.g., OMP, is applied to find the corresponding $\mathbf{w}_i$ with the dictionary $\mathbf{D}$ obtained from the training part. These sparse coefficient vectors form the coefficient matrix $\mathbf{W} = \{\mathbf{w}_i\}_{i=1}^L$.

- Quantizing/entropy coding part. After sparse coding, the first row of the coefficient matrix $\mathbf{W}$ contains the DC components. These DC components are coded separately by using a uniform quantizer and DPCM predictor followed by an adaptive arithmetic coder to encode the prediction residuals. Then the rest AC coefficients are uniform quantized and made into two sequences, one for selected atom position indices $\mathbf{S}$ and the other for non-zero values $\mathbf{V}$. The position indices are DPCM predicted and arithmetic encoded while the non-zero values are entropy coded into bit stream directly.

3. Entropy Constrained Dictionary Learning for Compression

Since remote sensing images contain rich textures, to reach a realistic compression quality standard, more atoms are selected and the coding costs become much higher. The DC components and the non-zero values $\mathbf{V}$ of AC components will be quantized to reduce the entropy. However, the indices $\mathbf{S}$ that denote the selected atom positions cannot be quantized, otherwise the decoder will fail in recovering the correct image. In fact, the non-zero positions in coefficient matrix $\mathbf{W}$ seem to be random, so encoding of the indices $\mathbf{S}$ becomes the bottleneck for compression. To solve this problem, a new compression scheme using entropy constrained dictionary learning is presented in this section.

In the training stage, instead of solving problem (3), our entropy constrained dictionary learning aims to jointly minimize the representation distortion, the number of non-zeros in coefficient matrix $\mathbf{W}$, as well as the entropy of the atom position indices $\mathbf{S}$:

$$\min_{\mathbf{D},\{\mathbf{b}_i\}_{i=1}^M} \sum_{i=1}^M ||\mathbf{b}_i - \mathbf{Dw}_i||_2^2 + \lambda H(\mathbf{S}) \text{ s.t. } ||\mathbf{w}_i||_0 \leq n_0, 1 \leq i \leq M \quad (4)$$

where $\lambda$ is the Lagrangian multiplier. The entropy $H(\mathbf{S})$ is defined as
\[ H(S) = - \sum_{i=1}^{K} p_j \log_2 p_j \]  
(5)

where \( p_j \) is the selected probability of the \( j \)-th atom in \( D \).

The proposed entropy constrained dictionary learning algorithm is described in Algorithms 2. It is also an alternating direction method, i.e., keeping \( D \) fixed, find \( W \), and then keeping \( W \) fixed, update \( D \). In each iteration step, when the \((k-1)\)-th dictionary \( D^{k-1} \) is updated, the \( k \)-th \( W^k \) will be computed, and hence the selected probability vector \( P^k = \{p_1, p_2, ..., p_K\} \) can be updated by

\[ p_j = \frac{||W^k(j,:)||_0}{||W^k||_0}, j = 1, 2, ..., K. \]  
(6)

Here \( W(j,:) \) is the \( j \)-th row vector of \( W \) and its \( \ell_0 \) pseudo-norm represents the selected times of the \( j \)-th atom. \( ||W||_0 \) represents the total number of non-zero entries. The sparse coding method used here is no longer OMP. Instead, an entropy constrained OMP (ECOMP), which is presented in Algorithm 3, is applied to solve the problem

\[ \min_w ||b_i - D^{k-1}w||_2^2 + \lambda H(S^k) \text{ s.t. } ||w||_0 \leq n_0. \]  
(7)

In each iteration step of ECOMP, one atom that reduces the residual as well as the entropy most is selected. In contrast, only the residual reduction ability is considered by OMP.

**Algorithm 2** Entropy constrained dictionary learning

**Input:** The training patches \( B = \{b_i\}_{i=1}^{M} \).

**Initialization:** Initialize \( k = 0 \), and

- Initialize the dictionary \( D^0 \in \mathbb{R}^{N \times K} \) by using randomly chosen examples and normalize the columns.
- Initialize the selected probabilities for the \( K \) atoms in \( D^0 \): \( P^0 = \{p_j = 1/K\}_{j=1}^{K} \).

**Main Iteration:** Increment \( k \) by 1 and perform the following steps:

- Entropy Constrained Sparse Coding (Algorithm 3): \( \tilde{w}_i = \text{ECOMP}(b_i, D^{k-1}, P^{k-1}), i = 1, 2, ..., M. \)
- Obtaining sparse coefficient matrix \( W^k = \{\tilde{w}_i\}_{i=1}^{M} \).
- Update Dictionary: Using RLS algorithm to compute \( D^k \):
  \( D^k = \text{RLS}(B, W^k). \)
- Update selected probabilities:
  \( P^k = \{p_j = ||W^k(j,:)||_0 / ||W^k||_0\}_{j=1}^{K} \).
- Stopping Rule: If \( ||B - D^k W^k||_F^2 \) is small enough, stop. Otherwise, apply another iteration.

**Output:** The dictionary \( D = D^k \) and the probability vector \( P = P^k \).

**Algorithm 3** Entropy constrained OMP (ECOMP)

**Input:** The signal \( b \), the dictionary \( D \), the maximum number \( n_0 \) of non-zeros, the selected probabilities \( P = \{p_1, p_2, ..., p_K\} \), the Lagrangian multiplier \( \lambda \).

**Initialization:** Initialize \( k = 0 \), and set

- Solution \( w^0 = 0 \).
- Number of non-zero entries \( n^0 = 0 \).
- Residual \( r^0 = b - Dw^0 = b \).
- Solution support \( S^0 = \text{Support}(w^0) = \phi \).

**Main Iteration:** Increment \( k \) by 1 and perform the following steps:

- Sweep: Compute the errors and the entropies
  \[ \epsilon(j) = \min_{z^*_j} \|d_j z^*_j - r^{k-1}\|_2^2 - \lambda p_j \log_2 p_j \]
  for all \( j \) using the optimal choice \( z^*_j = d_j^* r^{k-1}/\|d_j\|_2^2 \).
- Update Support: Find a minimizer, \( j_0 \) of \( \epsilon(j) : \forall j \notin S^{k-1}, \epsilon(j_0) \leq \epsilon(j) \), and update \( n^k = n^{k-1} + 1 \) and \( S^k = S^{k-1} \cup \{j_0\} \).
- Update Provisional Solution: Compute \( w^k \), the minimizer of \( \|b - Dw^k\|_2^2 \) s.t. \( \text{Support}(w) = S^k \).
- Update Residual: \( r^k = b - Dw^k \).
- Stopping Rule: If \( n^k = n_t \), stop. Otherwise, apply another iteration.

**Output:** The coefficient vector \( w = w^k \).

Our ultimate goal is to compress an image. Recall the three parts for compression described in section 2.3. In our proposed scheme, the training part employs entropy constrained dictionary learning to obtain a dictionary \( D \) and a probability vector \( P \). In the sparse coding part, ECOMP is applied to find the coefficient matrix \( W = \{w_i\}_{i=1}^{M} \) for the test patches \( X = \{x_i\}_{i=1}^{M} \), i.e., \( W = \text{ECOMP}(X, D, P) \). The quantizing/entropy coding part remains the same with that in section 2.3. When compared with the original RLS and OMP, the advantage of the training and sparse coding part of our method is that the position indices of the obtained coefficient matrix \( W \) are more concentrated, and as a result, the entropy of these indices is significantly reduced.

ECOMP is used both in the learning and sparse coding stage, and a natural question is, how to decide the Lagrangian multiplier \( \lambda \) in ECOMP properly? In problem (4) and (7), \( \lambda \) keeps a balance between the representation error and the entropy of the indices. Along with the increase of \( \lambda \), the coding cost of the indices decreases, but it also causes a decline in the reconstructed image quality. In our experiments, \( \lambda \) is adjusted to make the total encoding cost reaches a given compression bit rate.
4. Results

For our experiments, the dictionary is learned from the training images depicted in Fig. 1. The training set is formed by randomly selecting 1500 overlapping patches from each of the training images, giving 9000 samples in the training set. The method is tested on four remote sensing images acquired from RADARSAT-2, each of which are $512 \times 512$ pixels large and represented with 16 bpp as shown in Fig. 2. These four test images contain both the natural scenes (mountains and forests) and manmade scenes (cities and manmade targets) that typically appear in remote sensing images. Note that the dictionary is intended to be general for the large class of data acquired from a specific imaging system. The dictionary-based compression method has been proven to perform well on natural images and many relevant results are available in [3]-[5]. In our experiments, only the compression results of remote sensing images are listed. However, our method can be extended to natural image compression easily by using a dictionary learned from a set of natural images instead of remote sensing images.

Fig. 1. The training set consists of 6 remote sensing images acquired from RADARSAT-2.1

In order to assess the effectiveness of the proposed compression scheme, four other methods are applied for comparison: JPEG, SPIHT, JPEG2000, and an approach using regular dictionary learning RLS [1], [5]. The distortion is measured by signal to noise ratio (SNR)

$$SNR = 10 \log_{10} \left( \frac{\sigma_x^2}{\sigma_e^2} \right)$$  \hspace{1cm} (8)

where $\sigma_x^2$ denotes the variance of the original image and $\sigma_e^2$ denotes the variance of error between the original and reconstructed data. The experimental results with various compression bit rates are listed in Table 1 and Table 2 for image (a) and (b), respectively. All reported bit rates are calculated from real-coded bit streams.

Table 1

<table>
<thead>
<tr>
<th></th>
<th>1bpp</th>
<th>2bpp</th>
<th>3bpp</th>
<th>4bpp</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG</td>
<td>25.70</td>
<td>28.34</td>
<td>29.98</td>
<td>33.24</td>
</tr>
<tr>
<td>SPIHT</td>
<td>27.17</td>
<td>29.16</td>
<td>32.55</td>
<td>37.15</td>
</tr>
<tr>
<td>JPEG2000</td>
<td>27.42</td>
<td>29.89</td>
<td>34.42</td>
<td>38.80</td>
</tr>
<tr>
<td>RLS</td>
<td>27.47</td>
<td>30.91</td>
<td>34.55</td>
<td>38.56</td>
</tr>
<tr>
<td>Proposed method</td>
<td>27.61</td>
<td>31.02</td>
<td>34.65</td>
<td>38.84</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th></th>
<th>1bpp</th>
<th>2bpp</th>
<th>3bpp</th>
<th>4bpp</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG</td>
<td>14.72</td>
<td>17.92</td>
<td>20.88</td>
<td>24.65</td>
</tr>
<tr>
<td>SPIHT</td>
<td>15.68</td>
<td>19.46</td>
<td>23.26</td>
<td>27.83</td>
</tr>
<tr>
<td>JPEG2000</td>
<td>15.96</td>
<td>19.92</td>
<td>23.89</td>
<td>28.10</td>
</tr>
<tr>
<td>RLS</td>
<td>16.19</td>
<td>20.11</td>
<td>23.76</td>
<td>27.66</td>
</tr>
<tr>
<td>Proposed method</td>
<td>16.34</td>
<td>20.28</td>
<td>23.86</td>
<td>27.78</td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th></th>
<th>1bpp</th>
<th>2bpp</th>
<th>3bpp</th>
<th>4bpp</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG</td>
<td>25.74</td>
<td>32.86</td>
<td>38.41</td>
<td>40.72</td>
</tr>
<tr>
<td>SPIHT</td>
<td>28.34</td>
<td>34.87</td>
<td>39.63</td>
<td>43.98</td>
</tr>
<tr>
<td>JPEG2000</td>
<td>28.98</td>
<td>35.64</td>
<td>40.65</td>
<td>45.64</td>
</tr>
<tr>
<td>RLS</td>
<td>28.95</td>
<td>35.58</td>
<td>40.71</td>
<td>45.36</td>
</tr>
<tr>
<td>Proposed method</td>
<td>29.19</td>
<td>35.74</td>
<td>40.82</td>
<td>45.66</td>
</tr>
</tbody>
</table>

---

1 RADARSAT-2 Data and Products © MacDONALD, DETTWILER AND ASSOCIATES LTD. (2008) –All Rights Reserved. RADARSAT is an official mark of the Canadian Space Agency.
Table 4
SNR (dB) for compressed remote sensing image (d)

<table>
<thead>
<tr>
<th></th>
<th>1bpp</th>
<th>2bpp</th>
<th>3bpp</th>
<th>4bpp</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG</td>
<td>26.52</td>
<td>32.63</td>
<td>36.11</td>
<td>39.21</td>
</tr>
<tr>
<td>SPIHT</td>
<td>29.61</td>
<td>34.95</td>
<td>38.09</td>
<td>41.9</td>
</tr>
<tr>
<td>JPEG2000</td>
<td>30.02</td>
<td>35.21</td>
<td>38.88</td>
<td>43.07</td>
</tr>
<tr>
<td>RLS</td>
<td>30.09</td>
<td>35.3</td>
<td>38.93</td>
<td>42.64</td>
</tr>
<tr>
<td>Proposed method</td>
<td>30.23</td>
<td>35.44</td>
<td>39.04</td>
<td>42.76</td>
</tr>
</tbody>
</table>

The results show that the proposed scheme outperforms JPEG2000 by average 0.2 to 0.4 dB and is much better than JPEG and SPIHT. Compared with the approach using regular dictionary learning RLS, our proposal has a steady performance gain by limiting the entropy in the training and sparse coding stages. We also notice that JPEG2000 benefits a lot from the advanced EBCOT coding strategy, whereas ours is relatively simple. The performance of the proposed method can be improved further if more advanced quantization and entropy coding mechanisms are applied.

The computational complexity of the entropy constrained dictionary learning algorithm is high because dozens of iterations are required. However, the dictionary is trained offline with a representative training set so that it has no effect on the encoding and decoding running time. The running time is mainly spent in the sparse coding of the test image. In the experiments we use the SPAMS software [10] for sparse coding, which has an extremely fast implementation of OMP algorithm by Cholesky decomposition. All algorithms are tested on a regular PC (Pentium® Dual-Core CPU, 2.70GHz, 2GByte RAM). For a 16 bpp test image, 512×512, Fig. 2 (a), ECOMP runs 0.218 s and the quantization and entropy coding runs 0.160 s. The total encoding time of the proposed method is 0.378 s. Meanwhile, JPEG takes 0.28 s and JPEG2000 0.21 s. The execution time for the proposed decoder is only 0.12 s.

5. Conclusion

In this paper, we have designed a new entropy constrained dictionary learning algorithm to improve the compression performance for remote sensing images. The experimental results have revealed that the proposed method is better than the still image compression standards JPEG and JPEG2000. This method can be improved further in two aspects, 1) quantize and code the coefficients adaptively and 2) make the trained dictionary shape-adaptive. Our experiments prove that dictionary learning and sparse representation have great potential in remote sensing image compression. It is hoped that this paper will motivate future researches in this field.

Acknowledgements

We would like to thank the anonymous reviewers and editors, the support of the National Grand Fundamental Research 973 Programs of China under Grant No.2010CB731904, and the support of National Nature Science Foundation of China under Grant No.61172154.

References