DICTIONARY TRANSFER FOR IMAGE DENOISING VIA DOMAIN ADAPTATION

Gang Chen *Caiming Xiong* *Jason J. Corso*

Department of Computer Science, State University of New York at Buffalo
{gangchen,cxiong,jcorso}@buffalo.edu

ABSTRACT

The idea of using overcomplete dictionaries with prototype signal atoms for sparse representation has found many applications, among which image denoising is considered as an active research topic. However, the standard process to train a new dictionary for image denoising requires the whole image (or most parts) as input, which is costly; training the dictionary on just a few patches would result in overfitting. We instead propose a dictionary learning approach for image denoising via transfer learning. We transfer the source domain dictionary to a target domain for image denoising via a dictionary-regularization term in the energy function. Thus, we have a new dictionary that is trained from only a few patches of the target noisy image. We measure the performance on various corrupted images, and show that our method is fast and comparable to the state of the art. We also demonstrate cross-domain transfer (photo to medical image).

***Index Terms*—** Dictionary learning, image denoising, sparse representations, domain adaptation, transfer learning

1. INTRODUCTION

Compressive sensing [1] involves decomposing a signal into a linear combination of a few elements from a basis set, called a *dictionary* [2], such that only a very few samples are sufficient to reconstruct the signal [3, 4]. Formally, let $y \in \mathbb{R}^n$ be a signal and $D = [d_1, d_2, ..., d_k]$ be a dictionary in $\mathbb{R}^{n \times k}$ with k atoms ($k > n$, for an overcomplete dictionary). It is assumed that y can be represented as a sparse linear combination of these atoms; i.e., y may either be exactly $y = Dx$, or approximate, $y \approx Dx$, s.t. $||y - Dx||_2 \leq \epsilon$. In this setting, sparse coding with an ℓ_0 regularization amounts to computing

$$\min_x ||x||_0 \quad \text{subject to} \quad ||y - Dx||_2 \leq \epsilon \tag{1}$$

where $||\cdot||$ is the ℓ_0 norm, the number of the non-zero entries of a vector, and $\epsilon > 0$ models potentially noisy measurements. A natural approach to solving this problem is to alternate between the two variables (x and D), minimizing over one while keeping the other one fixed, see [4, 5].

This work was supported in part by NSF CAREER IIS-0845282, DARPA CSSG D11AP00245, and NIH 1 R21 CA160825-01.

Fig. 1. Sample images. The three images are different, such as different texture, shape and intensity. By dictionary transfer, we can denoise different kinds of images.

Exact determination of sparsest representations proves to be an NP-hard problem [6]. Thus, approximate solutions are considered instead: e.g., Matching Pursuit (MP) [7] and Orthogonal Matching Pursuit (OMP) algorithms [8]. As for dictionary learning, an overcomplete dictionary D can either be chosen as a prespecified set of functions, such as wavelets, DCT, steerable filters, etc., or designed by adapting its content to fit a given set of signal examples. Although denoising with a prespecified dictionary is simple and fast, it results in low reconstruction accuracy in most cases [3]. Moreover, a global learned dictionary does not transfer well to different types of images, see Fig. (1). Recent methods, such as K-SVD, use an adaptive strategy to learn a new dictionary, and show improvements in signal-to-noise ratios.

However, these adaptive methods, such as K-SVD [3] and MOD [9], are designed to work with overlapping patches (one per-pixel) of whole image. A 512×512 image will generate about 250000 patches (8×8). Despite fast algorithms [5], training with this many patches from a single image is time-consuming. Conversely, training with a few patches leads to overfitting. We hence face a dilemma on how to quickly learn the dictionary and yet achieve higher accuracy.

To that end, we propose a domain adaptation approach for dictionary learning. Our goal is to balance the trade off between learning speed and accuracy by transferring an existing dictionary trained from other images. Even though other dictionary learning approaches has been studied recently [10, 11, 12], we are aware of little work on dictionary transfer for image denoising. Domain adaptation [13] is the problem that arises when the data distribution in our test domain is different

from that in our training domain. In other words, we cannot directly use predefined dictionary for the target image denoising. Thus, we need to transfer the predefined dictionary to the target domain with a few training samples from the new image. Compared to the recent methods, such as K-SVD [3], our method speeds up the training process, maintains or improves accuracy and avoids overfitting.

2. OUR APPROACH FOR DICTIONARY LEARNING

In this section, we propose a dictionary learning through domain adaptation. Consider a few training patches $Y = \{y_i, i = 1, 2, ..., m\}$ in $\mathbb{R}^{n\times m}$, which are randomly sampled from a new corrupted image that we hope to reconstruct. Denote the corresponding coefficient $X = \{x_i, i = 1, 2, ..., m\}$ in $\mathbb{R}^{k\times m}$, s.t. $y_i \approx Dx_i$, where $D \in \mathbb{R}^{n\times k}$.

2.1. Dictionary learning with domain adaptation

Suppose we have a learned dictionary $D_0 \in \mathbb{R}^{n\times k}$ available, and we want to learn a new dictionary D given just a few image patches y_i by transfer learning. Thus, we propose to minimize the following equation:

$$L(X, D) = \min_{x,D} \sum_{i}^{m} ||y_i - Dx_i||_2^2 + \lambda_1 ||D - D_0||_F^2 \qquad (2)$$
$$\text{subject to} \quad \forall i, ||x_i||_0 \le L$$
$$\text{and} \quad \forall j = 1, 2, ..., k, d_j^T d_j \le 1$$

where y_i is an image patch written as a column vector, x_i is the corresponding coefficient, D in $\mathbb{R}^{n\times k}$ is a dictionary to be learned, L is the number of nonzero elements in each coefficient vector, and $||\cdot||_F$ is the Frobenius norm. To prevent D from being arbitrarily large (which would lead to arbitrarily small values of x_i), it is common to constrain its columns $(d_j)_{j=1}^k$ to have an ℓ_2 norm less than or equal to one. The first term in Eq. (2) is the data-dependent loss. Compared to K-SVD [3] and on-line dictionary learning [14], in the second term, we add D_0 as a regularizer to control the complexity of the target dictionary D—this term allows the dictionary transfer and is the main technical innovation of our paper. λ_1 is the weight for measuring the relevance between the source domain and the target domain. If two domains are relevant, D should be close to D_0. Note that the second regularization term in Eq. (2) is vital for our approach: (a) controls the complexity of the target domain; (b) avoids overfitting for the target dictionary D; (c) makes it possible with a few training data for dictionary learning, meanwhile getting high denoising accuracy. To minimize Eq. (2), we can use the standard alternating strategy between X and D, which we now explain.

2.2. Algorithm Outline

Our algorithm is summarized in Algorithm 1. It alternates the classical sparse coding steps with a fixed dictionary D, and

Algorithm 1

Input: Input image patches y_i, $D_0 \in \mathbb{R}^{n\times k}$, block size, λ_1, k (number of atoms) and T(number of iterarations).
Output: Dictionary D and coefficient $X = \{x_i, i = 1, 2, ..., m\}$.
Method:

1: initial $D = D_0$
2: **for** $t = 1 \to T$ **do**
3: *Sparse Coding Stage:* Use any pursuit algorithm to compute x_i for i =1,2,..., m
  • $\min_{x_i} ||y_i - Dx_i||_2^2$ subject to $||x_i||_0 \le L$
4: $A \leftarrow 0$, $B \leftarrow 0$,
5: for all training patches y_i and its corresponding coefficient x_i, $i = 1, 2, ...m$
  $A \leftarrow A + x_i x_i^T$
  $B \leftarrow B + y_i x_i^T$
6: *Dictionary Update Stage:*
  • $\min_D \sum_{i=1}^m ||y_i - Dx_i||_2^2 + \lambda_1 ||D - D_0||_F^2$
  • normalize D
7: **end for**

the dictionary update steps where X is fixed. As for sparse coding with D fixed, we compute the decompositions x_i by minimizing Eq. (1) for each patch y_i. In this paper, we use orthogonal match pursuit (OMP) method to calculate coefficient X (from the current D). Then, for updating the dictionary, there are two approaches: (1) using stochastic gradient descent [14]. This method (on-line learning) is effective when training dataset is very large. The strength of stochastic gradient descent methods (SGD) is that they are easy to implement and effective for large training data. However, they are subject to manual tuning of parameters (learning rate and convergence criteria) and may not be robust. (2) By setting gradient $\frac{\partial L(X,D)}{\partial D} = 0$, we can compute new dictionary D. In our work, since we consider a few training data for dictionary learning, method (2) is preferred.

2.3. Dictionary Update

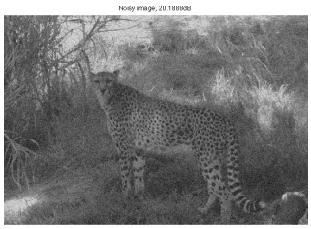
With X fixed, we minimize

$$\min_D \sum_{i=1}^{m} ||y_i - Dx_i||_2^2 + \lambda_1 ||D - D_0||_F^2$$
$$= \min_D \left(\mathbf{tr}(D^T D A) - 2\mathbf{tr}(D^T B) + \lambda_1 ||D - D_0||_F^2\right), \quad (3)$$

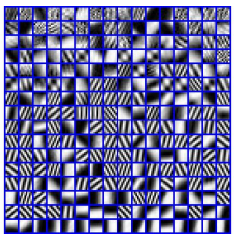
which has the following analytical solution:

$$D = (B + \lambda_1 D_0)(A + \lambda_1 \mathbf{I})^{-1} \ . \qquad (4)$$

Note that we need to normalize D when we compute it according to above formula. The parameter λ_1 is required to yield an invertible $(A + \lambda_1 \mathbf{I})$. When we have smaller training patches, larger λ_1 is preferred in order to avoid overfitting,

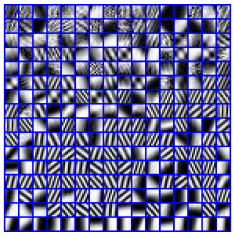


(a) Noise image (440×640)



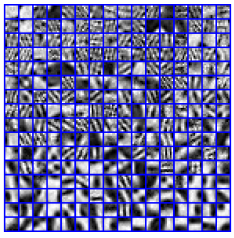
(b) Initial dictionary D_0

(c) Result (20 patches, 26.10 dB)



(d) Dictionary (20 patches)

(e) Result (26.45 dB)



(f) Dictionary (65000 patches)

Fig. 2. The first row is the noisy image and the pre-learned dictionary by K-means from Barbara (see Fig. 1). The second row is the denoising result and dictionary learned with 20 training patches. The last row is the denoising result and dictionary learned with 65000 training patches from corrupted cheetah image.

whereas when we have larger training patches, smaller λ_1 is set so that each atom in D can effectively reconstructed the original image.

In this paper, we use a dynamic strategy to set λ_1. We set $\lambda_1 = 2.0 \times 10^6 \times (1 - \frac{m}{col \times row})$, where m is the number of training patches, col and row are respectively the height and width of the noisy image; the constants were determined empirically and used for all experiments. Our dynamic strategy to set λ_1 can effectively avoid overfitting when m is small, and keep more weight on fidelity term so that the learned dictionary can describe the content of the target image. We do not observe any problem with inverting Eq. (4) either.

3. EXPERIMENTS

We have carried out several experiments on both natural and medical images to show the practicality of the proposed algorithm: our earlier claim is that the dictionary transfer would yield comparable or superior denoised images more efficiently than existing methods. We evaluate this claim and find it substantiated.

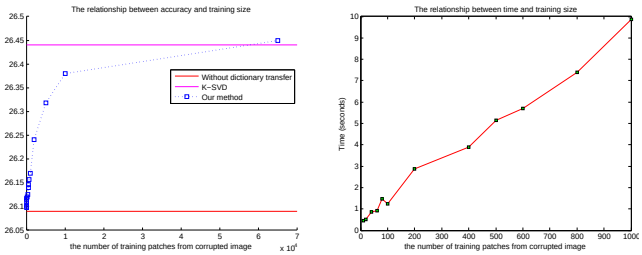
First, we learn the initial dictionary D_0 from image Barbara using K-means, and transferred it to Cheetah and MRI Brain, from Fig. (1). We use two points of comparison: a baseline without dictionary transfer (using D_0 for denoising directly) and the state of the art K-SVD [3]. In all experiments, we take the same parameters setting as K-SVD. For example, the dictionaries in use are of size 64×256, $T = 10$, i.e., the number of iteration is 10. Using the OMP, atoms were accumulated till the average error passed the threshold (chosen empirically to 1.15σ) or the number of nonzero coefficient larger than L (setting $L = 6$).

In experiment 1, we fix Gaussian noise $\sigma = 25$ and vary training size (number of patches). The training data are sampled patches (8×8 pixels) randomly from the target corrupted image (adding noise $\sigma = 25$ to Cheetah and MRI Brain). Fig. 2 demonstrates different dictionaries learned with different training size and Fig. 3 further describes the relationship between accuracy and training size. It shows that dictionary learning with transfer improves signal-to-noise when compared to no transfer. We also demonstrate the relationship between time and training size: our method can finish dictionary learning in few seconds for a thousand patches (and scales linearly). In contrast, K-SVD requires several minutes to train an adaptive dictionary for image size 512×512. We do not include the 10 patches result of K-SVD in Table 1, because it has an overfitting problem (10 patches are less than dictionary size 256). Table 1 and Fig. 3 show that the accuracy of our method increases gradually with more training patches.

In experiment 2, we compare the three methods by varying σ on the Cheetah image. We compute the peak SNR value (refer to the software package `http://www.cs.technion.ac.il/~elad/software/`). The results indicate a trade-off in performance between K-SVD and our method depending on the level of corruption in the image; see Figs. (4) and (5) for more details. In all cases, our method yields results comparable or superior to the selected state of the art and yet is more efficient.

4. CONCLUSIONS

In this paper, we have studied the dictionary transfer learning model for trade-off between computational cost and accuracy. The approach taken is based on sparse and redundant representations in overcomplete dictionaries learned. With only a few training samples, our method can speed up the learning process via domain adaptation. Furthermore, the domain adaptation allows us to circumvent the overfitting problem effectively. Our experiments have demonstrated that our method yields comparable or superior denoising more efficiently than the state of the art K-SVD method.



(a) (b)

Fig. 3. (a) Denoising accuracy ($\sigma = 25$) with varying number of training patches. (b) Computational time changing with the number of training patches. Few patches will definitely speed up dictionary training process.

Table 1. Accuracy with different training size ($\sigma = 25$)

Acc. / Img	#(patches) 10	500	1000	Methods
Cheetah	26.09	26.09	26.09	No transfer
	-	22.24	26.06	K-SVD
	26.10	26.17	26.22	Our method
MRI Brain	27.70	27.70	27.70	No transfer
	-	25.99	28.10	K-SVD
	27.74	28.80	28.82	Our method

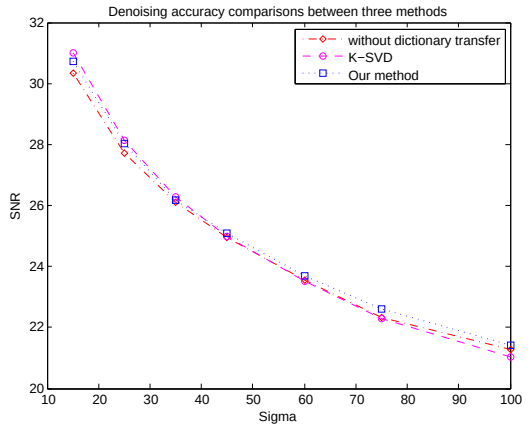


Fig. 4. Accuracy comparisons between the three methods by varying σ. It shows that our method outperforms K-SVD when Gaussian noise increasing.

5. REFERENCES

[1] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision Research*, vol. 37, 1997.

[2] D.L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] M. Aharon, M. Elad, and A.M. Bruckstein, "K-svd: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. On Signal Processing*, vol. 54, no. 11, pp. 4311–5322, 2005.

[4] Michael Elad and Michal Aharon, "Image denoising via learned dictionaries and sparse representation," in *CVPR*, 2006, pp. 17–22.

[5] H. Lee, A. Battle, R. Raina, and A. Y Ng, "Efficient sparse coding algorithms," in *NIPS*, 2007, pp. 801–808.

[6] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Journal of Constructive Approximation*, vol. 13, pp. 57–98, 1997.

[7] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[8] J.A. Tropp, "Greed is good: Algorithmic results for sparse approximation.," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[9] K. Engan, S. O. Aase, and J. H. Husoy, "Frame based signal compression using method of optimal directions (mod)," 1999.

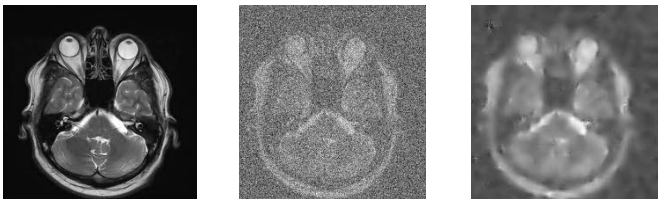
[10] Duc-Son Pham and Svetha Venkatesh, "Joint learning and dictionary construction for pattern recognition," pp. 1–8, 2008.

[11] Ron Rubinstein, Michael Zibulevsky, and Michael Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1553–1564, 2010.

[12] Ignacio Ramrez and Guillermo Sapiro, "Sparse coding and dictionary learning based on the mdl principle.," in *ICASSP*. 2011, pp. 2160–2163, IEEE.

[13] Hal Daumé, III and Daniel Marcu, "Domain adaptation for statistical classifiers," *J. Artif. Int. Res.*, vol. 26, pp. 101–126, May 2006.

[14] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online dictionary learning for sparse coding," in *ICML '09*, 2009, pp. 689–696.



(a) original image (b) corrupted image (c) our result

Fig. 5. Sample denoising result with Gaussian noise $\sigma = 100$.