BOOSTED DICTIONARIES FOR IMAGE RESTORATION BASED ON SPARSE REPRESENTATIONS

Karthikeyan Natesan Ramamurthy, Jayaraman J. Thiagarajan, Andreas Spanias and Prasanna Sattigeri

SenSIP Center, School of ECEE, Arizona State University, Tempe, AZ 85287-5706 USA.
Email: {knatesan, jjayaram, spanias, psattige}@asu.edu

ABSTRACT

Sparse representations using learned dictionaries have been successful in several image processing applications. However, using a single dictionary model in inverse problems may lead to instability in estimation. In this paper, we propose to perform image restoration using an ensemble of *weak* dictionaries that incorporate prior knowledge about the form of linear corruption. The dictionary learned in each round of the training procedure is optimized for the training examples having high reconstruction error in the previous round. The weak dictionaries are either obtained using a weighted K-Means or an example-selection approach. The final restored data is computed as a convex combination of data restored in individual rounds. Results with compressed recovery of standard images show that the proposed dictionaries result in a better performance compared to using a single dictionary obtained with a traditional alternating minimization approach.

Index Terms— Dictionary learning, Boosting, Sparse representations, Image restoration

1. INTRODUCTION

Natural signals and images exhibit statistics that allow them to be efficiently represented using a sparse linear combination of elementary patterns [1]. The local regions of natural images, referred to as patches, can be represented using a sparse linear combination of columns from a dictionary matrix. The generative model for sparse coding is hence given as $\mathbf{x} = \mathbf{D}\mathbf{a}$, where $\mathbf{x} \in \mathbb{R}^M$ is the data sample, $\mathbf{D} \in \mathbb{R}^{M \times K}$ is the dictionary matrix with K columns, and $\mathbf{a} \in \mathbb{R}^K$ is the sparse coefficient vector. The dictionary can be either pre-defined or learned from the training data itself. Learned dictionaries have been shown to provide improved performance for restoring degraded data in applications such as denoising, inpainting, deblurring, superresolution, and compressive sensing [2, 3], and also in machine learning applications such as classification and clustering [4, 5, 6].

Assuming that the training data $\mathbf{x}$ is obtained from a probability space, the dictionary learning problem can be expressed as minimizing the objective [7]

$$g(\mathbf{D}) = \mathbf{E}_{\mathbf{x}}[h(\mathbf{x}, \mathbf{D})], \tag{1}$$

where the columns of $\mathbf{D}$, referred to as dictionary atoms, are constrained as $\|\mathbf{d}_j\|_2 \leq 1, \forall j$. The sparse coding cost is

$$h(\mathbf{x}, \mathbf{D}) = \min_{\mathbf{a}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1. \tag{2}$$

If the continuous probability distribution is unknown and we only have L training samples $\{\mathbf{x}_i\}_{i=1}^L$, each with probability mass $p(\mathbf{x}_i)$, (1) can be modified as the empirical cost function,

$$\hat{g}(\mathbf{D}) = \sum_{i=1}^{L} h(\mathbf{x}_i, \mathbf{D}) p(\mathbf{x}_i). \tag{3}$$

Typically dictionary learning algorithms solve for the sparse codes using (2) [8, 9], and obtain the dictionary by minimizing $\hat{g}(\mathbf{D})$, repeating the steps until convergence. We refer to this baseline algorithm as *Alt-Opt*. Since this is an alternating minimization process, it is important to provide a good initial dictionary and this is performed by setting the atoms to normalized cluster centers of the data [10].

In this paper, we propose to restore degraded images by learning an ensemble of *weak* dictionaries, each of which need not be highly optimized with respect to the training data. The dictionaries are obtained using a simplified procedure, taking into account the corruption operation. The representation computed using the individual dictionaries will be combined together to obtain the final reconstructed image. The proposed algorithm, illustrated in Figure 1, incorporates the *boosting* procedure which is a well-known machine learning technique [11] used to improve the accuracy of learning algorithms, using multiple weak hypotheses instead of a single strong hypothesis. We show that the proposed boosted dictionaries resulted in an improved performance, for random-projection based compressive recovery, when compared to dictionaries obtained using the *Alt-Opt* method. It is important to note that the performance of boosted dictionaries is independent of the actual corruption, as long as the proper form of corruption is used during training.

2. IMAGE RESTORATION AND BOOSTING

In restoration applications, it is necessary to solve an inverse problem, in order to estimate the data $\mathbf{x}$ from

$$\mathbf{z} = \mathbf{\Phi}(\mathbf{x}) + \mathbf{n}, \tag{4}$$

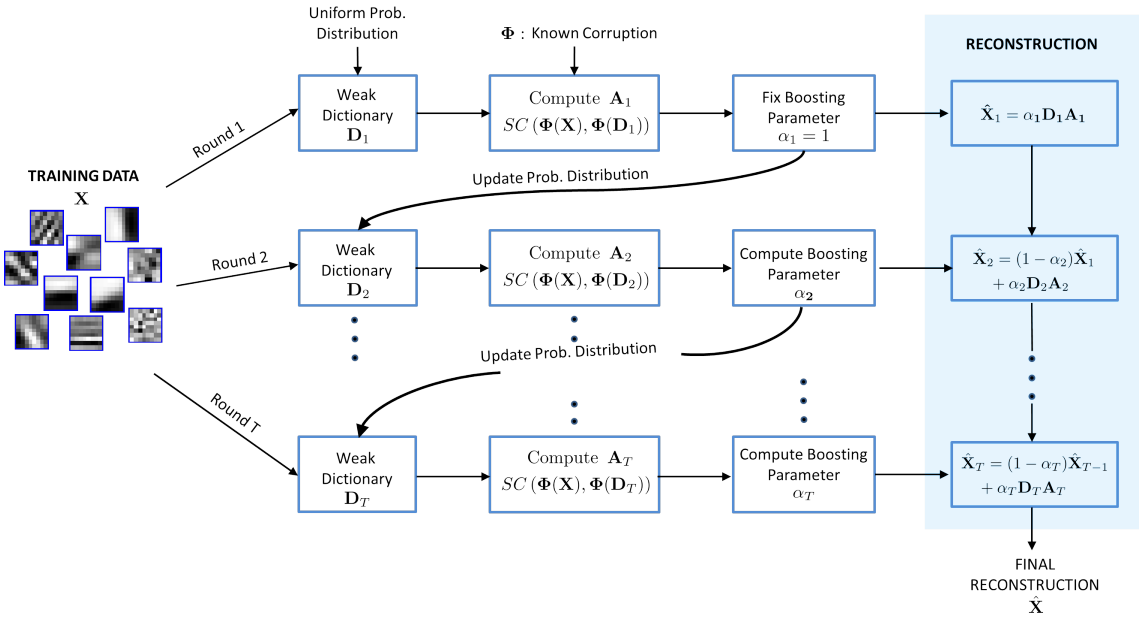


Fig. 1. Illustration of the proposed boosted dictionary learning for image restoration. SC denotes sparse coding using (2).

where $\mathbf{\Phi}(.)$ is the corruption operator and $\mathbf{n}$ is the additive noise. If the operator $\mathbf{\Phi}(.)$ is linear, which is the case in restoration applications mentioned earlier, we can represent it using the matrix $\mathbf{\Phi}$. With the prior knowledge that $\mathbf{x}$ is sparsely representable in a dictionary $\mathbf{D}$, (4) can be expressed as $\mathbf{z} = \mathbf{\Phi D a} + \mathbf{n}$. Restoring $\mathbf{x}$ now reduces to computing $\mathbf{a}$ by solving $h(\mathbf{z}, \mathbf{\Phi D})$, and estimating $\mathbf{x} = \mathbf{Da}$ [2]. Such traditional sparse models using a single learned dictionary suffer from the following drawbacks: (a) if the correlation between the atoms in $\mathbf{D}$ is not constrained to be low, the inverse problem estimation may become unstable, (b) even if the atoms in $\mathbf{D}$ are decorrelated, the atoms of the degraded dictionary $\mathbf{\Phi D}$ can be correlated, or have an ℓ_2 norm close to zero, both of which can lead to instability in inverse problems. The inversion can be stabilized and performance can be improved by including additional regularization in sparse modeling. In [12, 13], the authors propose to learn dictionaries and perform restoration by performing simultaneous sparse coding on a set of similar image patches, leading to an improved performance.

In this work, we adopt an approach of stabilizing the inversion by sequentially learning an ensemble of weak dictionaries using training data and obtaining the cumulative representation as a weighted average of the individual representations, as illustrated in Figure 1. Note that, boosting has been used with bag-of-words approach for updating codebooks in classification [14] and medical image retrieval [15]. However, it has not been used so far in sparsity based image restoration problems. The proposed algorithm starts by assuming a uniform probability distribution on the training data, and learns a weak dictionary. The restoration error of the degraded training data are computed, and the probability mass of the data samples that have higher restoration error are upweighted for the next round. The boosting weights for combining the reconstructions are optimized such that the cumulative reconstruction is in the convex hull of the individual reconstructions. We propose two methods for learning weak dictionaries - weighted K-Means method (*KM-Boost*) and example-selection method (*EX-Boost*).

3. PROPOSED ALGORITHM

In typical dictionary learning algorithms, such as the *Alt-Opt* method, a single *strong* dictionary $\mathbf{D}$ is learned by optimizing (3), assuming a uniform distribution on the training samples. As shown in Figure 1, we propose to learn T weak dictionaries by taking into account the corruption $\mathbf{\Phi}$, thereby stabilizing the inverse problem of estimating $\mathbf{x}$ from $\mathbf{z}$.

In any round t of the proposed method, the training data $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_L]$ is degraded using the known corruption $\mathbf{\Phi}$. The sparse codes $\mathbf{A}_t$ of the degraded data are computed using the weak dictionary $\mathbf{D}_t$ and the data is restored as $\mathbf{D}_t\mathbf{A}_t$. The boosting parameter α_t ensures that the estimated data for round t given by $\hat{\mathbf{X}}_t$ is a convex combination of the restoration $\mathbf{D}_t\mathbf{A}_t$ and the estimate of previous round $\hat{\mathbf{X}}_{t-1}$. This parameter α_t can be obtained as

$$\min_{\alpha} \left\| \mathbf{X} - \left[(1-\alpha)\hat{\mathbf{X}}_{t-1} + \alpha \mathbf{D}_t \mathbf{A}_t \right] \right\|_F^2 \text{ subj. to } 0 \leq \alpha \leq 1, \tag{5}$$

which ensures that the cumulative reconstruction is close to the training data. If α_t is high, it means that the data has been restored well using the dictionary $\mathbf{D}_t$. The squared error for

Image	Number of Measurements (N)								
	N = 8			N = 16			N = 32		
	Alt-Opt	KM-Boost	EX-Boost	Alt-Opt	KM-Boost	EX-Boost	Alt-Opt	KM-Boost	EX-Boost
Barbara	21.39	21.82	**22.05**	22.65	23.37	**23.76**	24.94	25.64	**26.34**
Boat	23.35	**23.91**	23.82	25.72	26.38	**26.42**	28.54	28.83	**29.32**
Couple	23.41	**24.04**	23.94	25.74	**26.51**	**26.51**	28.62	29.18	**29.60**
Fingerprint	18.40	**19.16**	18.83	21.52	**22.70**	22.34	24.97	**26.17**	26.16
House	24.84	**25.52**	25.29	27.73	**28.62**	28.44	30.97	31.66	**31.97**
Lena	25.51	**26.16**	25.94	28.12	**28.86**	28.82	30.99	31.50	**31.85**
Man	24.18	**24.75**	24.69	26.43	27.10	**27.23**	29.27	29.74	**30.24**
Peppers	21.54	**22.11**	21.99	24.08	24.60	**24.66**	27.20	27.34	**27.89**

Table 1. Compressed recovery of standard images: PSNR (dB) obtained using Alternating Dictionary Optimization (*Alt-Opt*), K-Means Boost (*KM-Boost*), and Example Boost (*EX-Boost*) methods, for different values of N. The results reported were obtained by averaging over 10 iterations with different random measurement matrices. In each, the higher PSNR is given in bold font.

i^{th} training sample in the cumulative restoration is

$$e_t(i) = \|\mathbf{x}_i - [(1-\alpha_t)\hat{\mathbf{x}}_{t-1,i} + \alpha_t \mathbf{D}_t \mathbf{a}_{t,i}]\,\|_2^2, \qquad (6)$$

and the maximum squared error across all training samples in round t is given by $e_{t,max}$. The probability mass of each training sample is updated as

$$p_{t+1}(\mathbf{x}_i) \leftarrow p_t(\mathbf{x}_i) \exp\left(\alpha_t e_t(i)/e_{t,max}\right), \qquad (7)$$

and normalized such that it sums to one. The samples that incur higher error in the current round are provided higher importance in the next round. This ensures that the weak dictionaries learned in the next round will provide a good restoration for these samples. Note that the distribution update procedure is similar to the standard procedure used in boosting for classification [11], with the modification that the cost function that measures performance is different for restoration.

3.1. Weak Dictionary Learning

The weak dictionaries obtained using the training samples and the probability distribution in each round make it possible for us to derive an ensemble model, as a combination of several weak models. Given a training set $\{\mathbf{x}_i\}_{i=1}^L$, and its probability masses $\{p(\mathbf{x_i})\}_{i=1}^L$, we will propose two simple approaches for learning weak dictionaries.

3.1.1. KM-Boost

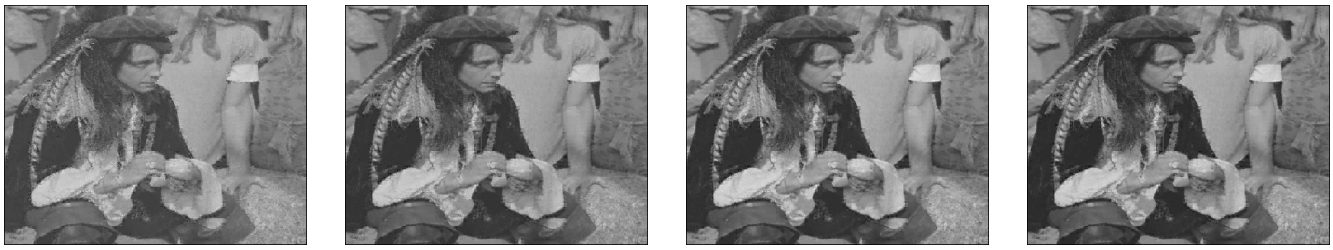
When the sparse code for each training example is constrained to take one only one non-zero coefficient of value 1, and the norms of the dictionary atoms are unconstrained, the dictionary learning problem (3) can be shown to reduce to K-Means clustering. Hence, computing a set of K-Means cluster centers and normalizing them to unit ℓ_2 norm constitutes a weak dictionary. However, since the distribution on the data could be non-uniform in our case, we need to alter the clustering scheme to incorporate the weight distribution. Denoting the cluster centers to be $\{\mathbf{c}_k\}_{k=1}^K$, the cluster membership sets to be $\{\mathcal{C}_k\}_{k=1}^K$, the weighted K-Means objective is denoted as

$$\min_{\{\mathbf{c}_k\}_{k=1}^K, \{\mathcal{C}_k\}_{k=1}^K} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} p(\mathbf{x}_i) \|\mathbf{x}_i - \mathbf{c}_k\|_2^2. \qquad (8)$$

The weighted K-Means procedure is implemented by modifying the K-Means++ algorithm [16], that provides a method for careful initialization leading to improved speed and accuracy in clustering. Let us denote δ_i as the shortest distance of the i^{th} training sample to the cluster center already chosen. The weighted K-Means algorithm proceeds as:

1. Pick first center $\mathbf{c}_1$ from the training set based on the distribution $\{p(\mathbf{x}_i)\}_{i=1}^L$.
2. $\mathbf{x}_i$ is chosen as the next center $\mathbf{c}_j$ with the probability $\frac{p(\mathbf{x}_i)\delta_i^2}{\sum_{l=1}^L p(\mathbf{x}_l)\delta_l^2}$.
3. Repeat step 2 until we have chosen K cluster centers.
4. *Cluster Assignment*: For each cluster $k = \{1, \ldots, K\}$, set the membership set $\mathcal{C}_k$ to contain training samples closer to $\mathbf{c}_k$ compared to other centers $\mathbf{c}_j, \forall j \neq k$.
5. *Cluster Update*: For each cluster $k = \{1, \ldots, K\}$, set $\mathbf{c}_k$ to be the weighted mean of all points in $\mathcal{C}_k$, $\mathbf{c}_k = \frac{\sum_{i \in \mathcal{C}_k} \mathbf{x}_i p(\mathbf{x}_i)}{\sum_{i \in \mathcal{C}_k} p(\mathbf{x}_i)}$.
6. Repeat steps 4 and 5 until convergence.

Note that the steps 1, 2, and 3 are used to compute the initial cluster centers giving preference to samples with higher probability mass. Finally, each dictionary atom $\mathbf{d}_k$ is set as the normalized cluster center $\frac{\mathbf{c}_k}{\|\mathbf{c}_k\|_2}$.



R1 (26.19 dB) **R5 (27.24 dB)** **R20 (27.55 dB)** **R50 (27.62 dB)**

Fig. 2. Compressed recovery of *Man* image using *EX-Boost* dictionaries. The reconstructed images along with their corresponding PSNR are shown for the rounds $\{1, 5, 20, 50\}$, when 25% random measurements are used.

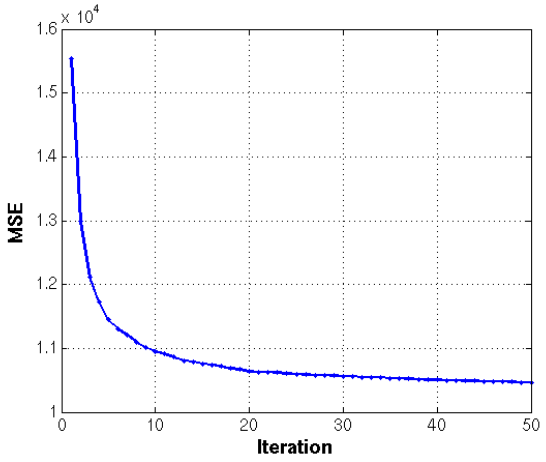


Fig. 3. Training MSE obtained after every round of dictionary training in the *EX-Boost* algorithm.

3.1.2. EX-Boost

From the dictionary update equation in (3), it is clear that the learned dictionary atoms are close to training samples that have higher probabilities. Therefore, in the *EX-Boost* method, the dictionary for round t is updated by choosing K data samples based on the non-uniform weight distribution, and normalizing them. This scheme will ensure that those samples with high restoration errors in the previous round, will be better approximated in the current round. Since the learning approach is simpler, *EX-Boost* typically requires more rounds of dictionary training compared to the *KM-Boost* procedure.

4. EXAMPLE APPLICATION: COMPRESSED RECOVERY

In compressed sensing, the $N-$dimensional observation $\mathbf{z}$ is obtained by projecting the $M-$dimensional data $\mathbf{y}$ onto a random linear subspace, where $N \ll M$ [17]. In this case, the entries of the corruption matrix $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$ are obtained as i.i.d. realizations of a Gaussian or Bernoulli random variable. In order to restore the test data from the observations, we adopt the traditional sparse coding approach, as well as the proposed approaches. The training data is obtained by randomly sampling $50,000$ patches of size 8×8 from the Berkeley Segmentation Dataset [18]. We compare the proposed *KM-Boost* and *EX-Boost* learning procedures with the alternating minimization dictionary learning algorithm (*Alt-Opt*), that learns a single dictionary. In all the cases, the sparsity penalty λ is set as 0.1. *KM-Boost* is repeated for 20 rounds, *EX-Boost* for 50 rounds, and ℓ_1-based learning algorithm is repeated for 100 iterations. The training MSE obtained with the proposed *EX-Boost* algorithm by progressively learning multiple rounds of dictionaries is shown in Figure 3.

When restoring test observations, the reconstruction procedure given in Figure 1 is used with the dictionaries $\{\mathbf{D}_t\}_{t=1}^{T}$, and the boosting coefficients $\{\alpha_t\}_{t=1}^{T}$, obtained during training. Compressed recovery was performed on 8 standard images with $N = \{8, 16, 32\}$ measurements and the results obtained are reported in Table 1. In each case, the PSNR values were obtained by averaging the results over 10 iterations with different random measurement matrices. The proposed approaches outperform the traditional dictionary learning in all cases. The improvement in reconstruction performance obtained with increasing number of boosted dictionaries is demonstrated in Figure 2. We also observed that the performance of the proposed approaches are not affected even if different corruption matrices are employed during training and testing phases.

5. CONCLUSIONS

We proposed an algorithm for learning an ensemble of weak dictionaries that incorporates the knowledge of degradation of data and provided a method for restoring the data from degraded observations. Results with compressed recovery show that the proposed method outperforms dictionaries obtained with the traditional alternating optimization method. Possible future research involves analyzing the convergence and generalization behavior of the algorithm apart from testing the proposed method for several types of degradations.

6. REFERENCES

[1] D.J. Field, "What is the goal of sensory coding?," *Neural computation*, vol. 6, no. 4, pp. 559–601, 1994.

[2] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*, Springer, 2010.

[3] J.J. Thiagarajan, K.N. Ramamurthy, and A. Spanias, "Multilevel dictionary learning for sparse representation of images," in *IEEE DSPE Workshop*, 2011, pp. 271–276.

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

[5] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *IEEE CVPR*, 2010, pp. 3501–3508.

[6] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," *Advances in neural information processing systems*, 2008.

[7] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. ICML*, 2009, pp. 689–696.

[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[9] H. Lee, A. Battle, R. Raina, and A.Y. Ng, "Efficient sparse coding algorithms," *Advances in neural information processing systems*, vol. 19, pp. 801, 2007.

[10] M. Aharon, M. Elad, and A. Bruckstein, "The K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[11] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, pp. 1612, 1999.

[12] G. Yu, G. Sapiro, and S. Mallat, "Image modeling and enhancement via structured sparse model selection," in *IEEE ICIP*, 2010, pp. 1641–1644.

[13] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *IEEE ICCV*, 2009, pp. 2272–2279.

[14] W. Zhang, A. Surve, X. Fern, and T. Dietterich, "Learning non-redundant codebooks for classifying complex objects," in *Proc. ICML*, 2009, pp. 1241–1248.

[15] J. Wang, Y. Li, Y. Zhang, H. Xie, and C. Wang, "Boosted learning of visual word weighting factors for bag-of-features based medical image retrieval," in *International Conference on Image and Graphics*, 2011, pp. 1035–1040.

[16] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.

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

[18] "Berkeley segmentation dataset," Available at http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/.