A Fast Dictionary Learning Method for Coupled Feature Space Learning

Farshad G. Veshki and Sergiy A. Vorobyov, Fellow, IEEE

Abstract

In this letter, we propose a novel computationally efficient coupled dictionary learning method that enforces pairwise correlation between the atoms of dictionaries learned to represent the underlying feature spaces of two different representations of the same signals, e.g., representations in different modalities or representations of the same signals measured with different qualities. The jointly learned correlated feature spaces represented by coupled dictionaries are used in sparse representation based classification, recognition and reconstruction tasks. The presented experimental results show that the proposed coupled dictionary learning method has a significantly lower computational cost. Moreover, the visual presentation of jointly learned dictionaries shows that the pairwise correlations between the corresponding atoms are ensured.

Index Terms

Coupled dictionary learning, feature space learning, sparse representation.

I. INTRODUCTION

Sparsity and overcompleteness has been successfully used for diverse applications in signal processing over the last decade [1]–[4]. The fact exploited is that signals can be compactly modelled using an overcomplete dictionary as a linear combination of only few *atoms*.

Formally, the basic synthesis model suggests that the signal x can be described as a linear combination of few atoms over an overcomplete dictionary D, and the problem of seeking such sparse representation can be formulated as $\min_{\alpha} \|\alpha\|_0$ s.t. $x \approx D\alpha$, where α is the sparse vector of coefficients for atoms in the dictionary D and $\|\cdot\|_0$ denotes the operator that counts the number of non-zero entries in a vector.

F. G. Veshki and S. A. Vorobyov are with Aalto University, Dept. Signal Processing and Acoustics, FI-00076, AALTO, Finland. E-mails: farshad.ghorbaniveshki@aalto.fi, svor@ieee.org **Corresponding author** is S. A. Vorobyov.

Many applications have benefited remarkably from using the above approach with learned overcomplete dictionary [5]–[8]. Representative examples of dictionary learning algorithms include the K-SVD method [9], the method of optimal directions (MOD) [10], the online dictionary learning (OLD) method [11], and their variants [12]–[14]. "Good" dictionaries are expected to be highly adaptive to the observed signals and to lead to accurate sparse representations.

While the *single dictionary* model has been extensively studied, there exists also a *coupled dictionary* viewpoint to sparsity and overcompleteness, where a coupled dictionary is needed to represent the double feature space (e.g., focused and blurred image patches in image processing). The combination of learned coupled dictionary and sparse approximation is shown to be superior for representing double feature spaces [15]–[22].

The coupled dictionary learning aims to find a pair of dictionaries $[D_1, D_2]$ best representing two subsets of *n* training signals $X_1 = [[x_1]_1, \dots, [x_1]_n]$ and $X_2 = [[x_2]_1, \dots, [x_2]_n]$ in such a way that the atoms of D_1 and D_2 are pairwise correlated, and if a linear combination of atoms of D_1 models a signal in X_1 , the same linear combination of atoms of D_2 also models the corresponding signal in X_2 . This can be insured by enforcing an identical sparse representation matrix Γ for both X_1 and X_2 while learning D_1 and D_2 . Then the coupled dictionary learning problem can be formulated as the following optimization problem [15]

$$\min_{\boldsymbol{D}_{1},\boldsymbol{D}_{2},\boldsymbol{\Gamma}} \|\boldsymbol{X}_{1} - \boldsymbol{D}_{1}\boldsymbol{\Gamma}\|_{2}^{2} + \|\boldsymbol{X}_{2} - \boldsymbol{D}_{2}\boldsymbol{\Gamma}\|_{2}^{2}$$
s.t. $\|\boldsymbol{\gamma}_{i}^{c}\|_{0} \leq T_{0}, \|[\boldsymbol{d}_{1}]_{t}\|_{2} = 1, \|[\boldsymbol{d}_{2}]_{t}\|_{2} = 1, \forall t, i$

$$(1)$$

where $[d_2]_t$ are the *t*-th dictionary atoms (columns) of D_1 and D_2 , respectively, T_0 is the constraint value on sparsity, and $\|\cdot\|_2$ is the Euclidian norm of a vector. The notation γ_i^c is used for *i*-th column of Γ , to be distinct from the notation that later is used for the rows of the same matrix.

The methods in [15]–[18] address (1) to model the function between observation and latent feature spaces (e.g., noisy and clear data), so that they can recover the unknown higher quality signals from their available low quality versions. Inverse problems such as image superresolution [15], [16], and speech signal bandwidth extension [17] are then examples of applications. For such methods, the corresponding dictionaries are expected to yield accurate sparse approximations. There are also methods that employ coupled dictionary learning techniques to solve problems such as cross-modal matching [19], cross-domain image recognition [20], and multifocus image fusion [21], as examples of classification and recognition applications. In latter

applications, the learned dictionaries are not required to provide accurate sparse recovery, but the objective is to learn the underlying feature spaces of X_1 and X_2 , i.e., the coupled dictionary.

A majority of existing coupled dictionary learning algorithms address (1) by learning two correlated feature spaces through burdensome complex procedures, while the computationally demanding nature of dictionary learning algorithms becomes more restrictive when we need to learn two dictionaries simultaneously. In this letter, we propose a fast coupled dictionary learning scheme that dramatically reduces the computational costs that brings it below the one by the K-SVD method even for a single dictionary.

II. A NEW PROPOSED METHOD

The optimization variables in problem (1) can be split into two subsets, where one subset consists of the common sparse representation matrix Γ , and the other includes the dictionaries D_1 and D_2 . Then (1) can be addressed in alternating manner by iterating between two phases, where in the first phase Γ is optimized under the constraint $\|\gamma_i^c\|_0 \leq T_0 - a$ *joint sparse coding* problem, and in the second phase D_1 and D_2 are optimized under the constraints $\|[d_1]_t\|_2 = 1$ and $\|[d_2]_t\|_2 = 1$, respectively– *dictionary update* problems. The general procedure of the proposed coupled dictionary learning is summarized in the block-diagram presented in Fig. 1. In the dictionary update phase, after updating each atom, all nonzero coefficients of its corresponding row of Γ have to be updated. The dashed arrow in the block diagram indicates that in order to preserve the same sparse representation for both D_1 and D_2 , the updates of Γ need to be performed jointly also during the dictionary update phase. Other operations, e.g., substituting unused atoms with better ones, are performed based on the common sparse representation matrix, thus the enforced atom-wise correlations in the joint sparse coding phase are preserved. The dictionaries can be initialized by any fixed basis overcomplete dictionary, e.g., discrete cosine transform (DCT) dictionary. *A. Joint Sparse Coding*

The joint sparse coding is the problem of finding optimal in least squares (LS) sense sparse representations of the joint dataset $\boldsymbol{X} \triangleq \left[\boldsymbol{X}_{1}^{T}, \boldsymbol{X}_{2}^{T}\right]^{T}$ over the joint dictionary $\boldsymbol{D} \triangleq \left[\boldsymbol{D}_{1}^{T}, \boldsymbol{D}_{2}^{T}\right]^{T}$, that is,

$$\min_{\boldsymbol{\Gamma}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{\Gamma}\|_{2}^{2} \quad \text{s.t.} \quad \|\boldsymbol{\gamma}_{i}^{c}\|_{0} \leqslant T_{0}, \ \forall i.$$
(2)

Problem (2) is known to be NP-hard, but by replacing $\|\cdot\|_0$ with l_1 -norm, it can be turned to a convex problem that is solvable by many existing methods. There are also sparse approximation

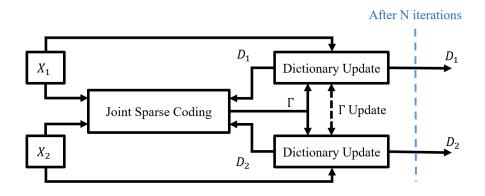


Fig. 1: Block-diagram of the proposed coupled dictionary learning method.

methods known as matching pursuits (MP) [23] which despite of not using explicit l_1 -norm term, are proved to yield approximations for l_1 -norm minimization problems [24].

Here, we use the orthogonal matching pursuit (OMP) method [25] to address the joint sparse coding problem. OMP is an iterative method that sequentially adds coefficients to the sparse representation vector in two steps.

The first step is to find the best matching atom for the signal (or residual). The standard formulation for the matching problem is as follows

$$\boldsymbol{d}_{best} = \operatorname*{argmax}_{\boldsymbol{d}_t} \mid \boldsymbol{d}_t^T \boldsymbol{r} \mid; \quad \boldsymbol{d}_t \in \boldsymbol{D}$$
(3)

where d_{best} denotes the best matching atom from the joint dictionary D for the joint residual $r \triangleq x_i - D\gamma_i^c$ and $x_i \in X$. "Matching" is measured by the absolute value of correlation, i.e., $|d_t^T r|$.

The second step is to calculate the coefficients for the atoms that are selected so far. This can be formulated as the following LS problem

$$\min_{\boldsymbol{\gamma}_{i}^{\mathrm{c}(m)}} \left\| \boldsymbol{r}^{(m)} - \boldsymbol{D}^{(m)} \boldsymbol{\gamma}_{i}^{\mathrm{c}(m)} \right\|_{2}^{2}$$
(4)

where $r^{(m)}$ is the residual, $\gamma_i^{c(m)}$ is the sparse representation vector, and $D^{(m)}$ is the subset of chosen atoms, all at *m*-th iteration. Problem (4) is equivalent to (1) optimized over Γ only, that is,

$$\min_{\boldsymbol{\gamma}_{i}^{c(m)}} \left\| \boldsymbol{r}_{1}^{(m)} - \boldsymbol{D}_{1}^{(m)} \boldsymbol{\gamma}_{i}^{c(m)} \right\|_{2}^{2} + \left\| \boldsymbol{r}_{2}^{(m)} - \boldsymbol{D}_{2}^{(m)} \boldsymbol{\gamma}_{i}^{c(m)} \right\|_{2}^{2}, \forall i$$
(5)

where r_1 and r_2 are the residuals from $[x_1]_i \in X_1$ and $[x_2]_i \in X_2$, respectively. Thus, OMP approximates a common sparse representation matrix for X_1 and X_2 over D_1 and D_2 , respectively.

At the end of each iteration, the residuals need to be updated. The algorithm iterates until the remainder error which is calculated as the norm squared of the residuals $e = \|\boldsymbol{r}^{(m)}\|_2^2$ satisfies the error threshold ϵ or the number of coefficients reaches its limit T_0 , i.e., the constraint $\|\boldsymbol{\gamma}_i^c\|_0 \leq T_0$ is satisfied as equality.

B. Dictionary Update

For the common sparse representation Γ , problem (1) needs to be solved then over the coupled dictionary D. Since the objective function of (1) is separable with respect to the dictionaries D_1 and D_2 , and different sets of constraints are applied to the atoms of D_1 and D_2 , problem (1) can be split into two subproblems of finding updates for the dictionaries D_1 and D_2 separately, although the similarity of the sparse representations has to be maintained. Thus, we explain the proposed dictionary update for a single dictionary D_i , i = 1, 2.

The corresponding optimization problem is given as

$$\boldsymbol{D}_{i} = \operatorname{argmin}_{\boldsymbol{D}_{i}} \left\| \boldsymbol{X}_{i} - \sum_{t} [\boldsymbol{d}_{i}]_{t} \boldsymbol{\gamma}_{t}^{\mathrm{r}} \right\|_{\mathrm{F}}^{2}$$
(6)

subject to the constraints in (1) applicable to corresponding atoms. Here γ_t^r is the *t*-th row of Γ . Note that in (6), we rewrite the product $D\Gamma$ as the sum of vector outer products $[d_i]_t \gamma_t^r$. After such modification, it appears that each atom can be updated disjoint from the others. Thus, to update the atom $[d_i]_t$, we fix the remaining atoms, and rewrite optimization problem (6) as

$$[\boldsymbol{d}_{i}]_{t} = \arg\min_{[\boldsymbol{d}_{i}]_{t}} \left\| \left(\boldsymbol{X}_{i} - \sum_{s \neq t} [\boldsymbol{d}_{i}]_{s} \boldsymbol{\gamma}_{s}^{\mathrm{r}} \right) - [\boldsymbol{d}_{i}]_{t} \boldsymbol{\gamma}_{t}^{\mathrm{r}} \right\|_{\mathrm{F}}^{2}.$$
(7)

Columns of $X_i - \sum_{s \neq t} [d_i]_s \gamma_s^r$ that correspond to zero entries of γ_t^r can be ignored. Thus, we define the vector ω_t representing the subset of indices where $\gamma_t^r \neq 0$, that is, $\omega_t = \{i | [\gamma_t^r]_i \neq 0\}$. Then the error matrix $[E_i]_t$ is formed as

$$[\boldsymbol{E}_i]_t \triangleq \left[\boldsymbol{X}_i - \sum_{s \neq t} [\boldsymbol{d}_i]_s \boldsymbol{\gamma}_t^{\mathrm{r}} \right]_{\boldsymbol{\omega}_t}.$$
(8)

Then optimization problem (7) can be further rewritten as the following simple rank-1 LS approximation problem

$$[\boldsymbol{d}_i]_t = \arg\min_{[\boldsymbol{d}_i]_t} \| [\boldsymbol{E}_i]_t - [\boldsymbol{d}_i]_t [\boldsymbol{\gamma}_t^{\mathrm{r}}]_{\boldsymbol{\omega}_t} \|_{\mathrm{F}}^2$$
(9)

where $[\boldsymbol{\gamma}_t^{\mathrm{r}}]_{\omega_t}$ contains nonzero entries of $\boldsymbol{\gamma}_t^{\mathrm{r}}$. There is no sparsity constraint in LS problem (9), thus, it can be easily solved as $\boldsymbol{d}_t = \boldsymbol{E}_t[\boldsymbol{\gamma}_t^{\mathrm{r}}]_{\omega_t}^T/\|[\boldsymbol{\gamma}_t^{\mathrm{r}}]_{\omega_t}\|_2^2$. The normalization term $\|[\boldsymbol{\gamma}_t^{\mathrm{r}}]_{\omega_t}\|_2^2$ can be dropped, since we need to normalize the l_2 -norm of each atom to one anyway. Then the atom update rule is

$$[\boldsymbol{d}_i]_t = [\boldsymbol{E}_i]_t [\boldsymbol{\gamma}_t^{\mathrm{r}}]_{\boldsymbol{\omega}_t}^T.$$
(10)

If ω_t is empty, $[d_i]_t$ is updated as the column-wise average of error matrix $[E_i]_t = X_i - D_i \Gamma$. To avoid the scale ambiguity in sparse approximation, the updated atoms are then normalized.

After updating $[d_i]_t$, we need to update $[\gamma_t^r]_{\omega_t}$ accordingly. Since $[d_i]_t$ is a unit vector, the solution of (9), this time over $[\gamma_t^r]_{\omega_t}$, can be efficiently found as $[\gamma_t^r]_{\omega_t} = [d_i]_t^T [\boldsymbol{E}_i]_t$. However, this solution is different for each feature space, i.e., i = 1 and i = 2. Thus, the optimal common nonzero coefficients can be found for the joint atom $d_t = [[d_1]_t^T, [d_2]_t^T]^T$ and joint error matrix $\boldsymbol{E}_t = [[\boldsymbol{E}_1]_t^T, [\boldsymbol{E}_2]_t^T]^T$, as

$$[\boldsymbol{\gamma}_t^{\mathrm{r}}]_{\boldsymbol{\omega}_t} = \frac{1}{2} \boldsymbol{d}_t^T \boldsymbol{E}_t. \tag{11}$$

The complexity orders of (10) and (11) are both O(mn), which is much smaller than that of singular value decomposition (SVD) in [9] with complexity order of $O(\max(m, n)^2 \times \min(m, n))$.

C. Maximum Number of Nonzero Coefficients

In each iteration, the majority of the existing two-phased alternating dictionary learning methods (including [9]–[14]) first find Γ over D, then update the atoms to reduce the error $||X - D\Gamma||_2^2$ in order to have a sparser Γ in the next iteration. That means that Γ is not sparse enough at the beginning. This backward approach imposes unnecessary extra computational costs, since a larger number of nonzero entries in sparse representation matrix leads to higher computational costs in both sparse coding and dictionary update phases.

Another drawback of this backward approach is that it reduces the effectiveness of the dictionary update phase. Each atom d_t is updated according to the error matrix E_t , which represents a potential amount of error that the atom update can compensate for in the total approximation error. When the dictionary is not learned to yield sparse enough approximations, the backward approach adds more coefficients to the sparse representations to minimize the approximation error, which leads to smaller entries for E_t , thus reducing the learning potential for updating d_t .

These issues can be easily addressed. Instead of setting the maximum number of nonzero coefficients as a constant number, we can gradually increase it. As a result, the first iterations become computationally cheap and the dictionary update phase becomes more effective. For example, we can form a vector of a size equal to the number of update cycles of dictionary learning algorithm, and set its values as equally spaced numbers between a minimum (e.g., 1) and the maximum number of nonzero coefficients. This simple change significantly reduces the computational cost without sacrificing the performance, even slightly.

D. Summary of the Algorithm

The overall algorithm for coupled dictionary learning can be then summarized as in Algorithm 1, where lines 3 to 11 represent the sparse coding phase, and lines 12 to 18 represent the dictionary update phase.

III. EXPERIMENTAL RESULTS

In this section, we first demonstrate that the proposed coupled dictionary learning method is able to provide the desired pairwise correlation between the atoms of two jointly learned dictionaries. As an example experiment, we generate two subsets of 20,000 focused and blurred 8×8 grayscale image patches taken from Lytro image dataset [26], and use them as X_1 (focused data) and X_2 (blurred data), where the patches (signals) in X_2 are blurred versions of their corresponding focused patches in X_1 . The columns of X_1 and X_2 are vectorized image patches. We apply our method to the double feature space and learn the correlated dictionaries D_1 and D_2 (see Figs. 2.(a) and (b)), then we visually compare it to the case where D_1 and D_2 are learned separately from the same feature spaces (see Figs. 2.(c) and (d)).

From the visual representations of atoms if Fig. 2, the pairwise correlations can be observed only between the atoms of dictionaries learned by the proposed coupled dictionary learning method. Those correlations are obtained by enforcing identical sparse representations through the proposed method and ensure that D_1 and D_2 represent corresponding features from the focused and blurred feature spaces.

Next, we compare our proposed dictionary learning method to the K-SVD and ODL methods, in terms of runtime, obtained number of nonzero coefficients, and average learning error $\sqrt{\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{\gamma}_i^c)^2}/n$, for learning a dictionary from a single feature space.¹ The experiment

¹Note that the proposed method is applicable without any change to a single dictionary learning as well.

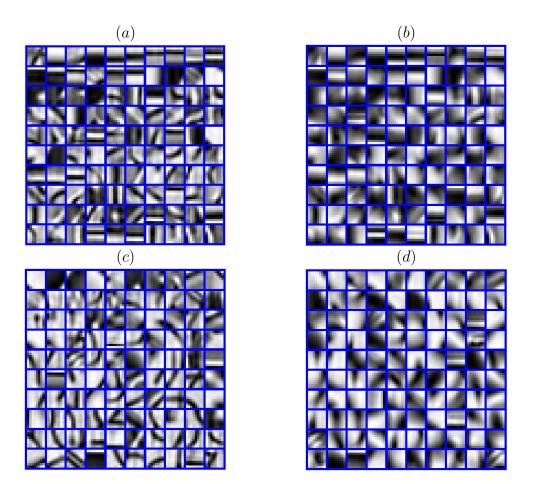


Fig. 2: Visual comparison between coupled learned dictionaries: (a) D_1 and (b) D_2 , and separately learned dictionaries: (c) D_1 and (d) D_2 .

is performed on a PC running an Intel(R) Xeon(R) 3.40GHz CPU. The learning dataset includes 10,000 mean centred grayscale image patches with the size of 8×8 , taken from Lytro image dataset. The tolerance error is set as $\epsilon = 4$, and the maximum number of nonzero coefficients is set to 32 (half of the size of vectorized patches). We run the K-SVD method for 16, the proposed algorithm for 32, and the ODL method for 256 dictionary learning cycles. The numbers of learning cycles are chosen with regards to the computational costs of the iterations of the algorithms, in a way that the ultimate runtimes are almost the same, so we can compare the results.

From Fig. 3(a), it can be observed that the dictionary learned by the proposed method yields significantly sparser representations in a much shorter time, comparing to those learned by the other methods. To explain this result, we visualize the changes in average learning error in

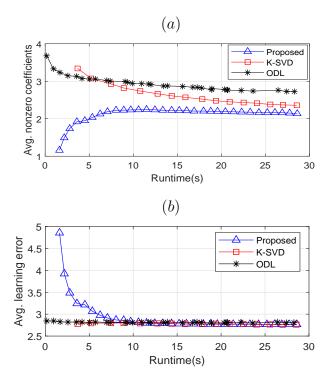


Fig. 3: The results for (a) the average number of nonzero coefficients, and (b) the average learning error versus the runtime. For the K-SVD and proposed methods, the markers indicate each iteration. For the ODL method, the markers show each 8 iterations.

Fig. 3(b). As explained in Subsection II-C, in the proposed method, we increase the maximum number of nonzero coefficients gradually. As a result, in the first iterations, the average error is high, however those iterations are faster. In this experiment, when the K-SVD method finishes its fifth iteration, the proposed algorithm has iterated 12 times. After about 11 seconds (13 iterations), the proposed method reaches the same average error and obtains a sparsity level which the K-SVD method achieves in about 28 seconds (16 iterations).

IV. CONCLUSION

A novel fast coupled dictionary learning algorithm that enforces common sparse approximations for double feature spaces and learns correlated pairs of atoms representing corresponding features from different feature spaces has been developed. The proposed dictionary learning method reduces dramatically the computational cost, which is important for computationally costly tasks such as coupled dictionary learning. The proposed method can be straightforwardly extended to find joint dictionaries for more than two feature spaces.

Algorithm 1 Coupled Dictionary Learning.

Input: Two training datasets of N signals X_1 and X_2 , and $D_0 = DCT$ dictionary.

1: Initialization: Set $D_1 := D_0$, $D_2 := D_0$.

Number of update cycles := N.

maxNum = A sequence of N equally spaced numbers between 1 and the maximum number of nonzero coefficients.

2: for
$$k = 1 \cdots N$$
 do

3: for
$$i = 1 \cdots n$$
 do

4: Set
$$\boldsymbol{r} = \left[[\boldsymbol{x}_1]_i^T, \ [\boldsymbol{x}_2]_i^T \right]^T;$$

 $m \leftarrow 1;$

5: while $e > \epsilon$ and $m \leq maxNum(k)$

6: Find d_{best} by solving (3);

7: Find
$$\gamma_t^{c(m)}$$
 by solving (4);

8: Update
$$\boldsymbol{r}^{(m)} = \boldsymbol{x}_i - \boldsymbol{D} \boldsymbol{\gamma}_t^{\mathrm{c}};$$

9: Update
$$e = \|\boldsymbol{r}^{(m)}\|_2^2;$$

 $m \leftarrow m + 1;$

```
10: end while
```

```
11: end for
```

```
12: for t = 1 \cdots number of atoms do
```

- 13: Find $\boldsymbol{\omega}_t = \{i | [\boldsymbol{\gamma}_t^{\mathrm{r}}]_i \neq 0\};$
- 14: Find $[\mathbf{E}_1]_t$ and $[\mathbf{E}_2]_t$ for $[\mathbf{d}_1]_t$ and $[\mathbf{d}_2]_t$ using (8);
- 15: Update $[d_1]_t$ and $[d_2]_t$ using (10);

```
16: Normalize the atoms:
```

$$[d_1]_t = [d_1]_t / ||[d_1]_t||_2$$
 and $[d_2]_t = [d_2]_t / ||[d_2]_t||_2$;

- 17: Update $\gamma_t^{\rm r}$ using (11);
- 18: **end for**

```
19: end for
```

Output: The pairwise correlated dictionaries D_1 and D_2 .

REFERENCES

- J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Mia, "Robust Face Recognition via Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
- [2] J. Yang, J. Wright, T. S. Huang, and Y. Mia, "Image Super-Resolution Via Sparse Representation," *IEEE Transactions on Image Processing.*, vol. 19, no. 11, pp. 2861–2873, May. 2010.
- [3] L. Zhang, W. D. Zhou, P. C. Cheng, J. Liu, Z. Yan, and T. Wang, "Kernel Sparse Representation-Based Classifier," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1684–1695, April. 2012.
- [4] Z. Tian, and G. B. Giannakis, "Compressed Sensing for Wideband Cognitive Radios," *IEEE International Conference on Acoustics, Speech and Signal Processing.*, honolulu, HI, USA, 2007.
- [5] I. ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*, San Francisco, CA, USA, 2010.
- [6] W. Dong, X. Li, L. Zhang, and G. Shi, "Sparsity-based image denoising via dictionary learning and structural clustering," *IEEE–CVPR.*, Colorado Springs, CO, USA, 2011.
- [7] Q. Xu, H. Yu, X. Mou, L. Zhang, J. Hsieh,and G. Wang, "Low-Dose X-ray CT Reconstruction via Dictionary Learning," *IEEE Transactions on Medical Imaging*., vol. 31, no. 9, pp. 1682–1697, Nov. 2012.
- [8] R. Rubinstein, T. Peleg, and M. Elad, "Analysis K-SVD: A Dictionary-Learning Algorithm for the Analysis Sparse Model," *IEEE Trans. Signal Process.*, vol. 61, no. 3, pp. 661–677, Feb. 2013.
- [9] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
- [10] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, Phoenix, AZ, USA, 1999, pp. 2443–2446.
- [11] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. ACM Int. Conf. Mach. Learn.*, Montreal, QC, Canada, 2009, pp. 689–696.
- [12] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit," *CS Technion*, vol. 40, no. 8, pp. 1–15, Apr. 2008.
- [13] Q. Zhang, and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*, San Francisco, CA, USA, Aug. 2010.
- [14] Z. Jiang, Z. Lin, and L. S. davis, "Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*., vol. 35, no. 11, pp. 2651–2664, May. 2013.
- [15] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467–3478, Aug. 2012.
- [16] J. Ahmed, R. A. Memon, M. Waqas, M. I. Mangrio, and S. Ali, "Selective sparse coding based coupled dictionary learning algorithm for single image super-resolution," in *Int. Conf. on Computing, Mathematics and Engineering Technologies* (*iCoMET*)., Sukkur, Pakistan, 2018.
- [17] J. Sadasivan, S. Mukherjee, and C. S. Seelamantula, "Joint dictionary training for bandwidth extension of speech signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, Shanghai, China, 2016, pp. 5925–5929.
- [18] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Rhode Island, USA, 2012, pp. 2216– 2223.

- [19] D. Mandal, and S. Biswas, "Generalized coupled dictionary learning approach with applications to cross-modal matching," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3826–3837, Jun. 2016.
- [20] F. Huang, and Y. F. Wang, "Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition" *IEEE Int. Conf. on Comp. Vision.*, Sydney, NSW, Australia, 2013.
- [21] R. Gao, S. A. Vorobyov, and H. Zhao, "Multi-focus image fusion via coupled dictionary training," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process., Shanghai, China, 2016, pp. 1666–1670.
- [22] T. Peleg, and M. Elad, "A statistical prediction model based on sparse representations for single image super-resolution," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2569–2582, Jun. 2014.
- [23] S. G. Mallat, Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing.*, vol. 41, no. 12, pp. 3397–3415, 1993
- [24] F. Locatello, R. Khanna, M. Tschannen, and M. Jaggi"A unified optimization view on generalized matching pursuit and Frank-Wolfe," in *Proc. 20th Intern. Conf. Artificial Intelligence and Statistics.*, vol. 54, pp. 860–868, Jul. 2017.
- [25] J. A. Tropp, and A.C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," in *Proc. IEEE Trans. Inf. Proc.*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
- [26] http://mansournejati.ece.iut.ac.ir/content/lytro-multi-focus-dataset.