Incrementally Built Dictionary Learning for Sparse Representation

Ludovic Trottier, Brahim Chaib-draa, and Philippe Giguère

Department of Computer Science and Software Engineering,
Université Laval, Québec (QC), G1V 0A6, Canada
ludovic.trottier.1@ulaval.ca
{chaib,philippe.giguere}@ift.ulaval.ca
http://www.damas.ift.ulaval.ca

Abstract. Extracting sparse representations with Dictionary Learning (DL) methods has led to interesting image and speech recognition results. DL has recently been extended to supervised learning (SDL) by using the dictionary for feature extraction and classification. One challenge with SDL is imposing diversity for extracting more discriminative features. To this end, we propose Incrementally Built Dictionary Learning (IBDL), a supervised multi-dictionary learning approach. Unlike existing methods, IBDL maximizes diversity by optimizing the between-class residual error distance. It can be easily parallelized since it learns the class-specific parameters independently. Moreover, we propose an incremental learning rule that improves the convergence guarantees of stochastic gradient descent under sparsity constraints. We evaluated our approach on benchmark digit and face recognition tasks, and obtained comparable performances to existing sparse representation and DL approaches.

Keywords: supervised dictionary learning, sparse representation, digit recognition, face recognition

1 Introduction

Feature extraction is a crucial step for improving the performance of machine learning algorithms. Various engineering methods were proposed in past decades for extracting the most useful information from raw observations [1]. However, recent developments in *representation learning* (RL) showed that feature engineering has limitations [2]. In particular, engineered features do not generalize to a large amount of problems due to their task-specific design, and the approaches are theoretically cumbersome to analyze and improve.

On the contrary, RL approaches do not suffer from these limitations because they aim to learn the relevant features instead of relying on expert knowledge for creating the pipeline of preprocessing transformations. Their goal is to construct in an unsupervised fashion a high-level representation from unlabeled data and use it for feature extraction. An important class of RL approaches, known as *dictionary learning* (DL), uses sparse modeling to construct efficient data representations by linearly combining a small number of typical patterns (atoms)

learned from data. Significant practical and theoretical contributions for learning a collection of such patterns (called a dictionary) have lead to state-of-the-art results in many signal processing and vision-related tasks [3] [4].

Recently, DL has been extended to *supervised dictionary learning* (SDL) by taking into account the label information instead of only relying on unlabeled data [5]. Among the different ways to extend a DL approach to supervised learning, one is to learn class-specific dictionaries. For instance, Ramirez et al. [6] proposed *dictionary learning with structured incoherence* (DLSI), a multi-dictionary approach minimizing the correlation between the class-specific dictionaries. Also, Yang et al. [7] incorporated a Fisher discrimination penalty in their *Fisher discrimination dictionary learning* (FDDL) approach to make the dictionaries more discriminative. Another way is to learn a joint dictionary with class-specific atoms. For example, Zhang et al. [8] proposed *discriminative KSVD* and Jiang et al. [9] proposed *label consistent KSVD*, both extending the well-known KSVD algorithm [10] to supervised learning. Finally, Wright et al. [11] proposed using the whole dataset as the dictionary in their *sparse representation-based classification* (SRC) approach and applied it on face recognition tasks.

The critical part when using multiple dictionaries in SDL is encouraging dictionary diversity for learning discriminative patterns [12]. The most common way to achieve this goal is by incorporating a discriminative term to the learning framework. For instance, DLSI uses the correlation between the class-specific dictionaries while FDDL uses a Fisher discrimination criterion. However, these new terms greatly complexify the learning phase. In this paper, we propose a framework for learning class-specific dictionaries under a diversity constraint without adding new discriminative terms. Moreover, we propose a learning algorithm that simultaneously treats the dictionaries in parallel. We finally propose an incremental learning rule that improves the convergence guarantees of stochastic gradient descent under sparsity constraints.

This paper is organized as follows. We introduce the reader to dictionary learning in Sec. 2 and present the proposed approach in Sec. 3. We show the experimentations in Sec. 4 and conclude in Sec. 5.

2 Background on Dictionary Learning

Let us define $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_N] \in \mathbb{R}^{M \times N}$ as the data matrix containing N M-dimensional observations and $\mathbf{D} = [\mathbf{d}_1 \dots \mathbf{d}_K] \in \mathbb{R}^{M \times K}$ as the dictionary with K atoms. Dictionary learning aims to represent each observation $\mathbf{x}_n$ as a sparse linear composition $\mathbf{w}_n$ of the dictionary atoms $\mathbf{d}_k$ with minimal residual error:

$$\min_{\mathbf{D},\mathbf{W}} \quad \|\mathbf{X} - \mathbf{D}\mathbf{W}\|_F^2 + \lambda \|\mathbf{W}\|_1 \quad s.t. \quad \|\mathbf{d}_k\|_2 = 1 \quad \forall k \,, \tag{1}$$

where $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_N] \in \mathbb{R}^{K \times N}$ is the weight matrix, $\|\cdot\|_F^2$ and $\|\cdot\|_1$ are respectively the squared Frobenius and entry-wise ℓ_1 norms. We define the residual error as $\|\mathbf{X} - \mathbf{D}\mathbf{W}\|_F^2$ and the sparsity constraint as $\lambda \|\mathbf{W}\|_1$. The hyper-parameter $\lambda > 0$ governs the sparsity of the weight vectors $\mathbf{w}_n$ and is usually chosen by

cross-validation. Constraining the dictionary atoms with $\|\mathbf{d}_k\|_2 = 1$ is needed for removing the trivial solution where $\mathbf{W} \approx 0$ and $\mathbf{D} \approx \infty$.

Minimizing Eq. 1 for both the dictionary $\mathbf{D}$ and the weight matrix $\mathbf{W}$ is however computationally too expensive, due to the coupling between $\mathbf{D}$ and $\mathbf{W}$ and the large amount of data. Approximating the optimum with an *iterative-alternative optimization* scheme is the usual choice for finding a suitable solution. The procedure works as follows. First, initialize randomly the dictionary $\mathbf{D}$. Second, minimize Eq. 1 w.r.t. the weight matrix $\mathbf{W}$, considering $\mathbf{D}$ fixed. We call this step *sparse coding*. Then, minimize Eq. 1 w.r.t. the dictionary $\mathbf{D}$, considering $\mathbf{W}$ fixed. We refer to this step as *dictionary learning*. Finally, alternate the two last steps until convergence.

Sparse coding is generally reduced to the LASSO [13] problem and has been solved by many approaches such as *orthogonal matching pursuit* (OMP) [14], *lest angle regression* (LARS) [15] and *marginal regression* (MR) [16]. On the other hand, dictionary learning is usually viewed as a constrained least squares problem. *Method of optimal direction* (MOD) [17], *online dictionary learning* [18] and KSVD [10] are examples of well-known methods for solving it.

3 Incrementally Built Dictionary Learning

In this section, we describe the proposed approach for imposing dictionary diversity. We first develop the optimization framework that will be used for learning the parameters and elaborate on its convergence guaranty. The section ends with a discussion on sparse-coding based features.

3.1 Approach Description

Let $\mathcal{D}$ be the unknown data distribution that generated the dataset $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$, where observation $\mathbf{x}_n \in \mathbb{R}^M$ has class label $y_n \in \{1 \dots C\}$. Let us define the following residual error-based model:

$$f_c(\mathbf{x}) = \min_{\mathbf{w}} \|\mathbf{x} - \mathbf{D}_c\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1 \,, \tag{2}$$

where $\mathbf{D}_c$ are class-specific dictionaries. Our goal is then to learn the dictionaries minimizing the expected classification error:

$$h = \arg\min_h \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\ell(h(\mathbf{x}), y)\right] \qquad \text{s.t.} \qquad h(\mathbf{x}) = \arg\min_c f_c(\mathbf{x}) \,, \tag{3}$$

where ℓ is the 0-1 loss. The classification rule $h(\mathbf{x})$ assigns class label c to observation $\mathbf{x}$ when the combined residual error and sparsity penalty using dictionary $\mathbf{D}_c$ is the smallest. Since Eq. 3 is a multi-dictionary approach, the main concern is imposing dictionary diversity for extracting discriminative features. We propose to maximize the distance between $f_c(\mathbf{x}_i)$ and $f_c(\mathbf{x}_j)$ where observations $\mathbf{x}_i$ are labeled c ($y_i = c$) and observations $\mathbf{x}_j$ are not labeled c ($y_j \neq c$). This learning principle is contrary to the one generally used in a SDL approach where

class-dictionary $\mathbf{D}_c$ must achieved the smallest residual error on observations from class c. Here, we rather encourage that the residual $f_c(\mathbf{x}_i)$ is far enough from the residual $f_c(\mathbf{x}_j)$. In a sense, we want the dictionaries to learn features maximizing the residual error margin to improve the separation of the classes. We believe that better classification performances could be achieved by tuning the class-c dictionary to increase the residual for observations not from class c rather than to reduce the residual for observations from class c. Therefore, we define the following optimization problem, $\forall c \in \{1 \dots C\}$:

$$\mathbf{D}_c = \arg\min_{\mathbf{D}_c} \sum_{\substack{\mathbf{x}_i \\ s.t.\ y_i = c}} \sum_{\substack{\mathbf{x}_j \\ s.t.\ y_j \neq c}} \mathcal{L}(\gamma f_c(\mathbf{x}_i) - f_c(\mathbf{x}_j)) \quad s.t. \quad \|\mathbf{d}_{ck}\|_2 = 1 \quad \forall k\,, \quad (4)$$

where $\mathcal{L}(x) = \max\{x, 0\}$ is a Hinge-based loss. The hyper-parameter $\gamma \geq 1$ governs the importance of minimizing the residual for $\mathbf{x}_i$ and will be inferred by cross validation. We emphasis that our framework is easily parallelized by solving Eq. 4 simultaneously for each class c. Minimizing Eq. 4 is done with projected stochastic gradient descent and the gradient of the loss function $\mathcal{L}(\gamma f_c(\mathbf{x}_i) - f_c(\mathbf{x}_j))$ is computed as follows:

$$\nabla_{\mathbf{D}_c} \mathcal{L} = \begin{cases} \mathbf{D}_c \left(\gamma \tilde{\mathbf{w}}_i \tilde{\mathbf{w}}_i^\top - \tilde{\mathbf{w}}_j \tilde{\mathbf{w}}_j^\top\right) + \mathbf{x}_j \tilde{\mathbf{w}}_j^\top - \gamma \mathbf{x}_i \tilde{\mathbf{w}}_i^\top & \text{if } \gamma f_c(\mathbf{x}_i) > f_c(\mathbf{x}_j) \\ 0 & \text{otherwise} \end{cases}, \quad (5)$$

where $\tilde{\mathbf{w}}_i = \arg\min_{\mathbf{w}} \|\mathbf{x}_i - \mathbf{D}_c \mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1$ is the sparse coding solution. The final update rule is then:

$$\mathbf{D}_c \leftarrow \Pi_{\mathbb{D}} \left(\mathbf{D}_c - \alpha \nabla_{\mathbf{D}_c} \mathcal{L}(\gamma f_c(\mathbf{x}_i) - f_c(\mathbf{x}_j))\right) \quad (6)$$

where $\alpha > 0$ is the learning rate and $\Pi_{\mathbb{D}}$ is a projection onto the subspace $\mathbb{D}$ of dictionaries having normalized columns.

3.2 Incremental Learning Rule

Our approach does not have a convergence guarantee. Due to sparsity, many dictionary atoms are never updated and only a small subset truly represent the latent structure. Consequently, we observe a *rich get richer* phenomenon where the same atoms get activated over and over again resulting in suboptimal dictionaries. To prevent this, we propose the following incremental learning rule. We initialize the dictionary $\mathbf{D}_c$ with $K_0 < K$ atoms and perform gradient descent. At each iteration, we keep track of the usage statistics by incrementing $\mathbf{u}_k$ by 1 when atom $\mathbf{d}_k$ is active ($\tilde{\mathbf{w}}_{ik} \neq 0$), $\forall k \in \{1 \dots K\}$. After T_0 iterations, we reconsider the dictionary in two ways. Let $u^* = \arg\min_k \mathbf{u}_k$ denotes the least used dictionary atom [1] and $\mathbf{p} = \mathbf{u}/\|\mathbf{u}\|_2$ denotes the probability distribution computed from the usage statistics. We first resample the least used dictionary

[1] In the case where there are several atoms, randomly select one of them.

Algorithm 1 Incrementally Built Dictionary Learning

Initialize $\mathbf{D}_c, \forall c \in \{1 \dots C\}$, each with K_0 random and normalized atoms.
for $c = 1 \dots C$ **do**
  Initialize the usage statistics: $\mathbf{u} \leftarrow \mathbf{0}$
  for $t = 1 \dots T$ **do**
    Sample $\mathbf{x}_i \sim \mathcal{D}$, $\mathbf{x}_j \sim \mathcal{D}$ such that $y_i = c$ and $y_j \neq c$
    Compute $\tilde{\mathbf{w}}_i$ and $\tilde{\mathbf{w}}_j$ using, for example, LARS.
    $\mathbf{D}_c \leftarrow \mathbf{D}_c - \alpha \nabla_{D_c} \mathcal{L}(\gamma f_c(\mathbf{x}_i) - f_c(\mathbf{x}_j))$ (Eq. 5)
    $\mathbf{d}_{ck} \leftarrow \mathbf{d}_{ck} / \|\mathbf{d}_{ck}\|_2, \forall k \in \{1 \dots K\}$
    $\mathbf{u} \leftarrow \mathbf{u} + \mathbf{1}_{\tilde{\mathbf{w}}_i \neq \mathbf{0}}$ (element-wise indicator function)
    if $0 \equiv t \mod T_0$ **and** $K_0 < K$ **then**
      $u^* = \arg\min_k \mathbf{u}_k$
      $i, j \sim \mathcal{GB}(\mathbf{p})$, where $\mathbf{p} = \mathbf{u}/\|\mathbf{u}\|_2$
      $\mathbf{d}_{cu^*} \sim \mathcal{N}(\mathbf{d}_{ci}, \sigma^2 \mathbf{1})$, $\mathbf{d}_+ \sim \mathcal{N}(\mathbf{d}_{cj}, \sigma^2 \mathbf{1})$
      $\mathbf{D}_c \leftarrow \mathbf{D}_c \cup \{\mathbf{d}_+\}$, $K_0 \leftarrow K_0 + 1$, $\mathbf{u} \leftarrow \mathbf{0}$
    end if
  end for
end for

atom $\mathbf{d}_{u^*}$ and second add a new atom to the dictionary according to the following scheme:

$$i, j \sim \mathcal{GB}(\mathbf{p}), \qquad \mathbf{d}_{u^*} \sim \mathcal{N}(\mathbf{d}_i, \sigma^2 \mathbf{1}) \qquad \mathbf{d}_+ \sim \mathcal{N}(\mathbf{d}_j, \sigma^2 \mathbf{1}), \tag{7}$$

where $\mathcal{GB}$ is the generalized Bernoulli distribution and $\mathcal{N}$ is the Gaussian distribution with variance parameter σ^2. Based on our experimentations, we found that $\sigma^2 = 1/M$ and $T_0 = T/2K$ achieved good performances, where T is the total number of iterations. The resulting algorithm is presented in Alg. 1 and we refer to our approach as *incrementally build dictionary learning* (IBDL).

There are two aspects motivating this incremental learning. First, unused dictionary atoms are either useless or specialized. At the beginning of the gradient descent, unused atoms are necessarily useless for reasons given earlier. However, at the end of the descent, unused atoms might represent a relevant structure even though they are rarely used for sparse coding. This explains our choice for T_0 because no atom resampling is permitted after $T/2$ iterations to allow learning specialized atom. Second, overused atoms need specialization. To understand this principle, let us study the following limit case. Suppose that a dictionary atom $\mathbf{d}_p$ has discovered such a complex structure that it always gets activated for encoding. However, most observations only uses subregions of $\mathbf{d}_p$ (corresponding to simpler structures) due to its overly complex nature. As a consequence, the gradient descent updates only target subregions of $\mathbf{d}_p$ (those found in the observation used for the gradient descent) and the content of $\mathbf{d}_p$ never settles to a stationary point. Sampling $\mathbf{d}_{u^*}$ and $\mathbf{d}_+$ according to the usage statistic allows them to share the complex structure found by $\mathbf{d}_p$.

	IBDL-E IBDL-C	SDL-G SDL-D	REC L REC BL	ℓ_2-KNN	SVM-Gauss	FDDL	DLSI	SRSC
MNIST	2.88 2.21	3.56 1.05	4.33 3.41	5.00	1.4	-	1.26	-
USPS	4.63 3.99	6.67 3.54	6.83 4.38	5.2	4.2	3.69	3.98	6.05

Table 1: Error rate results (%) of our approaches (IBDL-E and IBDL-C) and the state-of-the-art on the MNIST and USPS digit recognition tasks.

	MNIST					USPS				
	Encoder	K	λ	γ	α	Encoder	K	λ	γ	α
IBDL-E	LARS	75	0.66	11.36	0.00009	LARS	50	0.13	2.86	0.00006
IBDL-C	MR	25	0.06	10.60	0.0003	MR	50	11.18	1.84	0.0001
	5NN classifier					SVM, $gam = 0.00005$, $C = 11851.15$				

Table 2: Optimal hyper-parameter values of our approaches on the MNIST and USPS digit recognition tasks.

3.3 Sparse Coding-based Feature Extraction

Another alternative to using the residual error for classification is to train a classifier on sparse coding-based features. We therefore define $\mathbf{e} = [f_1(\mathbf{x}) \dots f_C(\mathbf{x})]$ the error vector and $\mathbf{r} = [\mathbf{w}_1 \dots \mathbf{w}_C]$ the representation vector of observation $\mathbf{x}$ computed from the class-specific dictionaries. We construct a feature vector by concatenating the normalized vectors $\mathbf{e}$ and $\mathbf{r}$ to form the vector $\boldsymbol{\phi} = [\frac{\mathbf{e}}{\|\mathbf{e}\|_2}, \frac{\mathbf{r}}{\|\mathbf{r}\|_2}]$ and use it as input for the classifier.

4 Experimentations

We tested IBDL on digit recognition using MNIST and USPS datasets and on face recognition using the Extended Yale B dataset. We evaluated the two proposed types of classification: (1) minimal residual error (IBDL-E), as defined by Eq. 3, and (2) classification with our sparse coding-based features (IBDL-C), as defined in Sec. 3.3. For all tasks, we cross-validated (3-fold) the hyper-parameters $\Theta = \{\lambda, \gamma, \alpha\}$ using Bayesian optimization [19] and tested the OMP, MR and LARS sparse coding algorithms with $K \in \{25, 50, 75\}$ and $T \in \{10000, 25000, 50000\}$. We report the test score of the approach achieving the best validation score. For IBDL-C, we evaluated the KNN and RBF-SVM classifiers. We cross-validated their hyper-parameters using Bayesian optimization [19] and report the test score of the best approach.

4.1 Digits Recognition

The USPS dataset contains 7,291 training and 2,007 testing 16×16 images and the MNIST contains 60,000 training and 10,000 testing 28×28 images. We report in Tab. 1 the performances of our approach in comparison to others in the

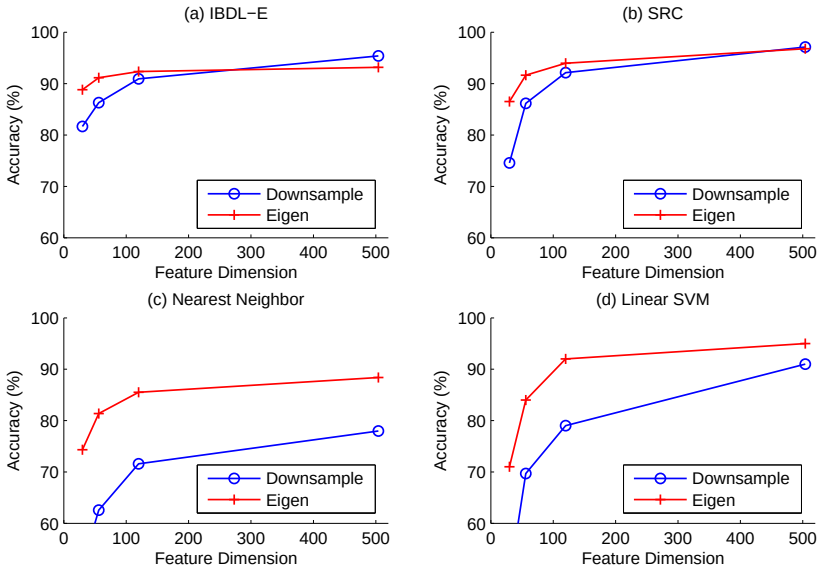

Fig. 1: Face recognition results on the Extended Yale B dataset.

literature: REC-L, REC-BL, SDL-G and SDL-D [5], KNN, SVM and DLSI [6], FDLL [7] and SRSC [20]. The optimal hyper-parameters are also reported in Tab. 2.

Discussion

Our approach IBDL-C has competitive performances on the USPS dataset and comparable performances on MNIST with the state-of-the-art. The main difference between these two tasks is the number of observations and the image size. Since our approach samples one observation from class c and another one from class $\neg c$, it appears that more iterations are needed when dealing with larger datasets. Due to the curse of dimensionality, the image size may also affect the accuracy. Even though IBDL-E does not work well on either tasks, our results show that the KNN classifier has better accuracy with our sparse coding-based features. We believe that using both the structure of the weight vectors $\mathbf{w}_c$ and the error values $\mathbf{e}$ for constructing the feature vector $\boldsymbol{\phi}$ makes it more discriminative. Further investigations for explaining why IBDL-C achieve better accuracy than IBDL-E is needed.

4.2 Face Recognition

The Extended Yale B dataset contains 2,414 frontal-face images of 38 individuals (approximatively 64 images per subject). We used the experimental setup of [11] and compared against SRC, NN and SVM from [11] for downsampled images and Eigenfaces. The accuracy results are reported in Fig. 1. We tested for $K = 25$

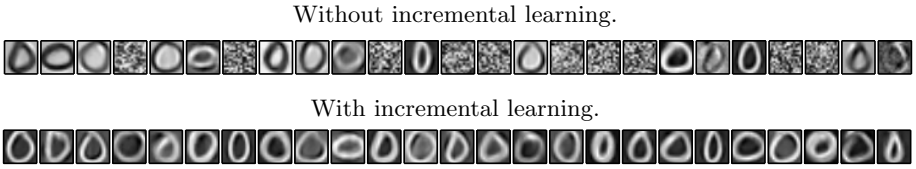

Fig. 2: The effects of incremental learning on the USPS dataset for class digit 0. Each image corresponds to a dictionary atom $\mathbf{d}_k$. The dictionary learned without incremental learning (top row) contains uninformative features in comparison to the dictionary learned with incremental learning (bottom row).

and $T = 10000$. IBDL-E achieved a maximum accuracy of 95.39% with the OMP encoding and $\lambda = 10$, $\gamma = 14.66$ and $\alpha = 0.000009$.

Discussion

The IBDL-E approach has state-of-the-art performances on face recognition. Interestingly, our approach achieved its best accuracy with downsampled images. This was unexpected since eigenfaces are usually known to outperform them (it is indeed the case for SRC, NN and SVM). We believe that the eigenface transformation removes important information during the orthogonalization affecting the latent structure learning. Using downsampled images is advantageous because no training data is required for extracting the features. Also, IBDL-E has better accuracy than SRC with low-dimensional features but does not outperform it with high-dimensional features. We believe that this is due to the curse of dimensionality. However, since our approach learns the dictionaries independently, the 38 class-specific models were trained in parallel. This parallel linear speed up is important for recognition task with many classes, such as face recognition which requires on class per individual.

4.3 The Effects of Incremental Learning

In this section, we demonstrate the beneficial effects of the proposed incremental learning rule. We trained two IBDL models on the USPS dataset by optimizing Eq. 4, one with incremental learning and the other without, using a LARS encoding with hyper-parameters $\lambda = 3$, $\alpha = 0.0001$, $\gamma = 15$, $T = 5000$ and $K = 25$. Fig. 2 shows the 25 learned dictionary atoms of both dictionaries for the class digit 0. As explained in Sec. 3, the dictionary trained without incremental learning is suboptimal containing atoms unrelated to the structure of a 0. This can clearly be seen in the top row of Fig. 2 where many entries (e.g. 4th) are just noise. Those atoms were never updated during the gradient descent (their $\mathbf{u}_k$ were 0). On the contrary, the dictionary learned with incremental learning (bottom row of Fig. 2) used all atoms to represent the latent structure. The same phenomenon appeared for all class-specific dictionaries. Therefore, this shows empirically that incremental learning improves the convergence guarantee of the gradient descent under a sparsity constraint for finding a more representative dictionary.

5 Conclusion

In this paper, we proposed *incrementally built dictionary learning* (IBDL), a supervised multi-dictionary approach for classification. The IBDL aims to learn class-specific dictionaries with high diversity by optimizing the between-class residual error distance. We proposed a parallel optimization framework based on stochastic gradient descent that allows learning the dictionaries simultaneously. The preliminary experimental results on digit and face recognition show that IBDL achieves good accuracy on face recognition and improved the KNN classifier performances with the proposed sparse coding-based features. As future work, we will extend the notion of diversity by using probability simplexes as dictionary atoms and by considering a Kullback-Leibler based distance between the dictionaries.

References

1. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.: Feature Extraction, Foundations and Applications. Springer, New York (2006)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. In 2013 IEEE Transactions on Pattern Analysis and Machine Intelligence 35(8), pp. 1798–1828 (2013)
3. Kreutz-Delgado, K., Murray, J. F., Rao, B. D., Engan, K., Lee, T. W., Sejnowski, T. J.: Dictionary Learning Algorithms for Sparse Representation. Neural computation 15(2), pp. 349–396 (2003)
4. Coates, A., Ng, A. Y.: The Importance of Encoding Versus Training with Sparse Coding and Vector Quantization. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 921–928 (2011)
5. Mairal, J., Ponce, J., Sapiro, G., Zisserman, A., Bach, F. R.: Supervised dictionary learning. In Advances in Neural Information Processing Systems, pp. 1033–1040 (2009)
6. Ramirez, I., Sprechmann, P., Sapiro, G.: Classification and Clustering via Dictionary Learning with Structured Incoherence and Shared Features. In 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3501–3508 (2010)
7. Yang, M, Zhang, D, Feng, X: Fisher Discrimination Dictionary Learning for Sparse Representation. In 2011 IEEE International Conference on Computer Vision (ICCV), pp. 543–550 (2011)
8. Zhang, Q., Li, B.: Discriminative K-SVD for Dictionary Learning in Face Recognition. In 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2691–2698 (2010)
9. Jiang, Z., Lin, Z., Davis, L. S.: Learning a Discriminative Dictionary for Sparse Coding via Label Consistent K-SVD. In 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1697–1704 (2011)
10. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. In IEEE Transactions on Signal Processing 54, no. 11, pp. 4311–4322 (2006)
11. Wright, J, Yang, A. Y., Ganesh, A., Sastry, S. S., Ma, Y.: Robust Face Recognition via Sparse Representation. In 2009 IEEE Transactions on Pattern Analysis and Machine Intelligence, 31, no. 2, pp. 210–227 (2009)

12. Donoho, D. L., Elad, M.: Optimally Sparse Representation in General (nonorthogonal) Dictionaries via ℓ_1 Minimization. Proceedings of the National Academy of Sciences, 100(5), pp. 2197-2202 (2003)
13. Tibshirani, Robert: Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), pp. 267–288 (1996)
14. Pati, Y. C., Rezaiifar, R., Krishnaprasad, P.: Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition. In 1993 Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, pp. 40–44 (1993)
15. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least Angle Regression. The Annals of statistics, 32(2), pp. 407–499 (2004)
16. Donoho, D. L., Johnstone, I. M.: Adapting to Unknown Smoothness via Wavelet Shrinkage. Journal of the american statistical association, 90(432), pp. 1200–1224 (1995)
17. Engan, K., Aase, S. O., Husoy, J. H.: Method of Optimal Directions for Frame Design. In 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 2443–2446 (1999)
18. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online Learning for Matrix Factorization and Sparse Coding. The Journal of Machine Learning Research, 11, pp. 19–60 (2010)
19. Snoek, J., Larochelle, H., Adams, R. P.: Practical Bayesian Optimization of Machine Learning Algorithms. In Advances in Neural Information Processing Systems, pp. 2951–2959 (2012)
20. Huang, K., Aviyente, S.: Sparse Representation for Signal Classification. In Advances in neural information processing systems, pp. 609–616 (2006)