A HIERARCHICAL-STRUCTURED DICTIONARY LEARNING FOR IMAGE CLASSIFICATION

Jaesik Yoon, Jinho Choi, Chang D. Yoo

Korea Advanced Institute of Science and Technology
Department of Electrical Engineering

ABSTRACT

This paper proposes a hierarchical-structured discriminative dictionary learning algorithm for image classification. Hierarchical structure of the overall dictionary is learned such that the upper-level dictionaries are specific in representing patterns common across a wide set of class images while lower-level dictionaries are specific in representing patterns localized to a narrow set of class images. Therefore the root dictionary can represent patterns common to all classes, while the leaf dictionaries can represent patterns specific only to a single distinct class. The learned dictionary is efficient in its use of the bases, and leads to a more discriminative representation than that led by previous dictionaries which is devoid of any structure and contains redundant bases. This hierarchical-structured dictionary is learned by solving a constraint optimization problem that minimized reconstruction error of a given image while using dictionaries in the hierarchical structure pertaining only to the class of the image. Sparse representation is pursued in addition, and it acts a regularizer to improve generalization. The representation is as distinct as the paths to each of the class in the hierarchical structure are divergent. To evaluate the effectness of the hierarchical-structured dictionary, classification is performed on three benchmark datasets: Extended Yale B database, Caltech 101 and Caltech 256 dataset, and based on a common features, the proposed algorithm performs better than other state-of-the-art dictionary learning algorithms.

Index Terms— feature learning, discriminative dictionary learning, sparse coding, classification

1. INTRODUCTION

Originally proposed for representing data with sparse coefficients, dictionary learning in the context in the sparse coding algorithm has been attracting considerable attention in recent years for its application in various computer vision tasks: image denoising [1], image super-resolution [2] and image classification [3, 4, 5] applications. In image classification, the sparse coefficients are used as discriminative features, and various dictionary learning algorithms have been developed to design discriminative sparse coefficients. In this paper, a hierarchical-structured discriminative dictionary learning algorithm for image classification is considered.

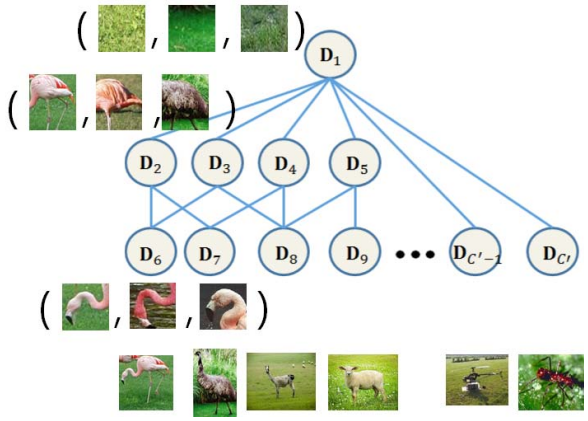

Fig. 1. Proposed algorithm using hierarchical structure : Image classification algorithms using common sub-dictionary have been proposed. This sub-dictionary is used to represent similar condition among entire classes, as background clutter or illumination or noise. However, different levels of similarities are not considered. We propose a hierarchical-structured discriminative dictionary learning algorithm, which uses those similarities to classification by using hierarchical structure.

Previous efforts to design discriminative sparse coefficients can be categorized in two ways: (1) efforts to directly formulate discriminative sparse coefficients and (2) efforts to construct incoherent dictionaries among different classes [6]. In the first way, Zhang and Li, Jiang *et al*. and Yang *et al*. proposed discriminative dictionary learning algorithms [7, 3, 4] and in the second way, Ramirez *et al*. proposed a constraint optimization that included an additional constraint in the dictionary topology to introduce incoherency among sub-dictionaries of each class while having a common sub-dictionary for the overall class [8]. Kong and Wang proposed a dictionary learning algorithm that combined the two different ways [9].

In this paper, a hierarchical-structured discriminative dictionary learning algorithm is considered in providing different levels of commonality. The upper-level dictionaries in the hierarchical structure provide basis for representing patterns that are more common across a wide set of class images while lower-level dictionaries provide basis specific in representing patterns localized to a narrow set of class images. The proposed algorithm uses a unbalanced hierarchical-structured dictionary such that the leafs are at different depth with respect the root as shown in Fig. 1.

In Section2, a dictionary learning is introduced. Section 3 describes the hierarchical-structured discriminative dictionary learning algorithm. The hierarchical-structured dictionary is learned by solving a constraint optimization problem that minimized reconstruction error of a given image while using dictionaries in the hierarchical structured pertaining only to the class of the image. Sparse representation is pursued in addition, and it acts a regularizer to improve generalization. Section 4 evaluates the effectness of the hierarchical-structured dictionary, classification is performed on three benchmark datasets: Extended Yale B database, Caltech 101 and Caltech 256 dataset. Section 5 concludes and provides a summary.

2. BACKGROUND

2.1. Dictionary learning and sparse coding

Dictionary learning with a sparse coding algorithm learns a dictionary to represent data with a few bases and sparse coefficients. Let us consider a set of *M*-dimensional *N* data $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N] \in \mathbb{R}^{M \times N}$. The input data $\mathbf{Y}$ can be reconstructed as a linear combination of two matrices, $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_P] \in \mathbb{R}^{M \times P}$ and $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{P \times N}$, where $\mathbf{D}$ denotes a *M*-dimensional learned dictionary containing *P* bases, and $\mathbf{X}$ denotes sparse coefficients for *N* data. It can be achieved by optimizing the following equation:

$$\min_{\mathbf{D},\mathbf{X}} \sum_{i=1}^{N} \frac{1}{2} ||\mathbf{y}_i - \mathbf{D}\mathbf{x}_i||_2^2 + \lambda_s \Omega(\mathbf{x}_i) \tag{1}$$

where, Ω is regularization term, usually used as l_0-norm or l_1-norm to make sparsity, whose effect is balanced by the regularization parameter $\lambda_s > 0$.

If this objective function is a convex optimization problem, there can be various methods to solve the problem. However, this problem is not a convex optimization problem, thus, this problem is transposed to convex optimization problem by solving each of the variables $\mathbf{D}$ and $\mathbf{X}$ separately. In optimization procedure for $\mathbf{D}$, this function is differentiable; therefore, the general gradient descent method or a method using a pseudo inverse matrix (this is called MOD [10]) can be used. However, for $\mathbf{X}$, while l_0-norm is used as a regularization term, this function is not a convex function, and this is an NP-hard problem; therefore, a variety of greedy algorithms such as Matching Pursuit [11] or Orthogonal Matching Pursuit[12] are used to solve this problem. With l_1-norm, this function is a convex, but non-derivative function; thus, many researchers have developed various algorithms to solve this problem quickly and accurately [13, 14].

3. DISCRIMINATIVE DICTIONARY LEARNING

3.1. Constraint optimization

Dictionary learning classification algorithms provide discriminative sparse coefficients by solving a constraint optimization problem with an objective function consisting of a representation error term, a discriminative term and a sparsity term [7, 3, 15, 16, 4, 8, 9]. However, no term used in those algorithms can have a positive effect on any other term. For example, if the algorithm uses a term to make a dictionary incoherent and a term to make sparse coefficients discriminative, separately, then a more incoherent dictionary cannot make more discriminative sparse coefficients and more discriminative sparse coefficients cannot make a more incoherent dictionary; thus, even if two terms can help make highly discriminative sparse coefficients, using two discriminative terms separately will not lead to one common goal, which would be to make highly discriminative sparse coefficients with a properly incoherent dictionary. Thus, the proposed algorithm uses one term, which makes a properly incoherent dictionary that is sufficient to make highly discriminative sparse coefficients; this is given as follows.

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{X}} \sum_{c=1}^{C} [\frac{1}{2} ||\mathbf{Y}_c - \mathbf{D}\mathbf{X}_c||_F^2 + \\ &\frac{\lambda_D}{2} ||(\mathbf{D}\mathbf{W}\mathbf{Q}_c)^T \mathbf{D}\mathbf{W}\mathbf{Q}_{\setminus c}\mathbf{Q}_{\setminus c}\mathbf{X}_c||_F^2] + \lambda_s \sum_{i=1}^{N} ||\mathbf{x}_i||_1 \end{aligned} \tag{2}$$

where $\mathbf{Y}_c \in \mathbb{R}^{M \times N_c}$ and $\mathbf{X}_c \in \mathbb{R}^{P \times N_c}$ are the data and sparse coefficients matrix of class c. $\sum_{c=1}^{C} N_c = N$. $\mathbf{D} \in \mathbb{R}^{M \times P}$ is Dictionary matrix and $\mathbf{Q}_c$ is diagonal matrix, and, if i th basis is included in class c, then $\mathbf{Q}_c(i,i) = 1$, else $\mathbf{Q}_c(i,i) = 0$ and if this objective function is used with the above hierarchical structure, $\mathbf{Q}_c$ includes bases of specific common levels corresponding to class c and the entire common level. $\mathbf{W}$ is also diagonal matrix, and, $\mathbf{W}(i,i)$ has different value in accordance with level of sub-dictionary including i th basis. $\mathbf{Q}_{\setminus c}$ is $\mathbf{I}(P,P) - \mathbf{Q}_c$ for P by P identity matrix $\mathbf{I}$. $\mathbf{Q}_c, \mathbf{Q}_{\setminus c}, \mathbf{W}$ is provided by hierarchical structure, which is constructed with KL-divergence of data.

As for the optimization process, if sparse coefficients are not made sufficiently discriminative in the sparse coding step, then $\mathbf{Q}_{\setminus c}\mathbf{X}_c$ is not equal to zero, and the second term has an effect of making the dictionary incoherent in the dictionary learning step; and, if sparse coefficients are made sufficiently discriminative in the sparse coding step, then $\mathbf{Q}_{\setminus c}\mathbf{X}_c$

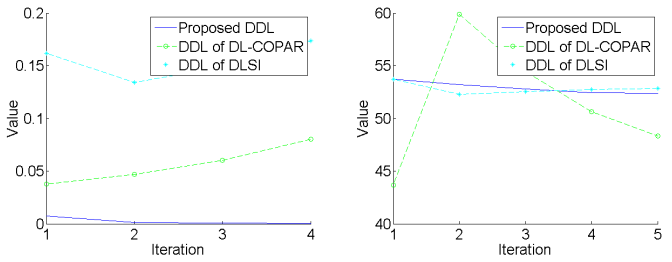

Fig. 2. The comparison of various DDL process of same structure on the Caltech 101 sub dataset. (a) shows how data are represented by using sub dictionries of wrong classes. Y-axis is $||\mathbf{D}\mathbf{Q}_{\setminus c}\mathbf{X}_c||_F$ and x-axis is iteration number. (b) shows how sub-dictionaries are incoherenced. Y-axis is correlation of each sub-dictionary $\sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}||(\mathbf{D}\mathbf{Q}_i)^T\mathbf{D}\mathbf{Q}_j||_F$ and x-axis is also iteration number.

is equal to zero and, in the dictionary learning step, the second term does not have the effect on the learning of the dictionary. Eventually, the goal of this objective function is to make a properly incoherent dictionary sufficient for making highly discriminative sparse coefficients This process and comparison of other algorithm are described in Fig 2.

3.2. Optimization

The objective function in Eq. 2 can be optimized for each variable, dictionary $\mathbf{D}$ and sparse coefficients $\mathbf{X}$, separately to make this function as convex optimization problem. First, for sparse coefficients $\mathbf{X}$, the objective function is not differentiable (l_1-norm is summation of absolute element value). Thus, the general gradient descent method or a method using pseudo inverse matrix are not used to solve it for $\mathbf{X}$. A variety of algorithms have also been developed to solve this problem by using various approaches [17, 14, 18]. The proposed algorithm uses H. Lee's algorithm [14], which has shown good performance in several application tasks [2, 19]. This algorithm can find sparse coefficients for the squared error term. However, the criterion of the proposed algorithm has a discriminative term. Thus , we make the squared loss term and discriminative term, as shown in the following equation, into one squared loss term.

In the second step, we use a matrix differential to learn the dictionary with a gradient descent method for the entire dictionary matrix $\mathbf{D}$.

4. EXPERIMENTS

Experiments were performed on one public face database, the Extended YaleB database[22], and multi-class object category datasets: the Caltech 101[23] and Caltech 256[24]. Random faces, which are made by projecting a face image onto a random vector, are used for the feature descriptor in the Extended YaleB data base. The size of a random face feature in the Extended YaleB is 504. For Caltech 101 and Caltech 256, we use a dense SIFT(DSIFT) descriptor. The DSIFT descriptor is extracted from 25×25 pixel patch which is densely sampled on a dense grid with 8 pixels. Then, we extract the sparse coding spatial pyramid matching(ScSPM) feature[25], which is concatenation of vector pooled from words of the extracted DSIFT descriptor. Dimension of words is 1024 and max pooling techique is used with pooling grid of $1 \times 1, 2 \times 2, 4 \times 4$. Thus, dimension of ScSPM feature is 1024*21 = 21504, and for processing speed of classification, it is reduced to 3000 by using PCA.

Before optimization process, K-SVD algorithm [20] is used for each sub-dictionary initialization and specific common and entire common sub-dictionary is initialized with data of included classes. Stepsize parameter, st is 1 and the number of iteration for sparse coding and dictionary learning are 10 and the number of iteration for gradient descent method is 5. $\mathbf{W}(\mathbf{i},\mathbf{i})$ has 1, 2 and 4 values, when i th basis is included to common, specific common and class sub-dictionary. $\lambda_{\mathbf{D}}$ is 2 and sparsity parameters of each algorithm, λ_s, are 0.1 to SRC, 0.05 to DLSI, DL-COPAR and proposed DDL and 0.01 to HSDDL, because, structure having more common levels needs more bases to representation. The sub-dictionary size of common and specic level is 3. SRC classication procedure is used for experiments.

4.1. Extended YaleB Database

The Extended YaleB database consists of 2,414 frontal-face images of 38 persons [22]. There are about 64 images for each person, and the face images, whose sizes are 192×168 after cropping and normalization, are under various illumination conditions and have various expressions. For the experiment, we randomly selected half of the images (about 32 images per person) for training; the rest of the images were used for testing. The dictionary size of HSDDL is about 500, which is smaller than dictionary size of other algorithms, 1216 or 575. The proposed algorithm is compared with various discriminative dictionary algorithms and is found to outperform those other algorithms in Table 1. The highest accuracy of the discriminative dictionary algorithm DL-COPAR (using one entire common level) was 98.3%; the HSDDL outperforms this value by about 0.3% and as shown in table 2, proposed algorithm is 5 times faster than algorithms showing similar accuracy for dictionary learning.

4.2. Caltech101 Dataset

The Caltech101 dataset consists of 102 classes (including background class). The number of images in the Caltech101 dataset is 9144 images and the number of images of each class is from 31 to 800. Most images are about 300 x 300

	Acc.(%, training data : 32)
SRC	97.2
K-SVD	93.1
D-KSVD	94.1
LC-KSVD	96.7
DLSI	96.5
FDDL	97.9
DL-COPAR	98.3
proposed DDL	98.4
proposed HSDDL	**98.6**

Table 1. Recognition results using random-face features on the Extended YaleB database.

	LC-KSVD	DLSI	DL-COPAR	HSDDL
DL time(sec.)	10.62	56.65	58.33	11.48

Table 2. Computation time for dictionary learning using random-face features on the Extended YaleB database.

pixels; images have high shape variability. We compared the results for the proposed algorithm with those from a variety of discriminative dictionary learning algorithms. Detailed comparison results are shown in Table 3 and HSDDL shows higher accuracy than previous algorithms. In this experiments, dictionary size of HSDDL is about 2500 to training data : 30 and about 1200 to training data : 15. They are smaller than dictionary used in previous algorithm, about 3000 to training data : 30 and about 1500 to training data : 15. As can be seen, our proposed algorithm leads to lower accuracy than that of DL-COPAR*, by about 3.5%; the reason for this is that DL-COPAR* is used to represent local features. On the other hand, our proposed algorithm is used to represent sparse coding spatial pyramid matching features. Except for DL-COPAR* whose feature extraction algorithm we could not duplicate, the proposed algorithm outperformed various other state-of-the-art dictionary learning based image classification algorithms.

4.3. Caltech256 Dataset

The Caltech256 dataset includes a total of 30,607 images classified into 257 object categories(including background class). The number of images of each class is a minimum of 80 and goes up to 827. Variability in object size, location, etc. is higher than that of the Caltech101 dataset. Dictionary size of HSDDL is about 5000, which is smaller than dictionary size of other algorithms, about 7700. The experiment results are enumerated in Table 4 and it is seen to achieve higher accuracy than that of previous algorithms, about 2%.

Acc.(%)	training data : 15	training data : 30
SRC	70.9	78.02
D-KSVD	71.10	78.50
LC-KSVD	71	78.30
DLSI	71.78	78.58
DL-COPAR	72.10	78.58
Proposed DDL	72.80	78.62
Proposed HSDDL	**73.04**	**79.24**

Table 3. Recognition results on the Caltech101 dataset.

	Acc.(%, training data : 30)
D-KSVD	34.9
LC-KSVD	34.9
DLSI	37.2
DL-COPAR	38.2
proposed DDL	38.8
Proposed HSDDL	**40.1**

Table 4. Recognition results using DSIFT features on the Caltech256 dataset.

5. CONCLUSION

This paper proposes a hierarchical-structured discriminative dictionary learning algorithm for image classification. Hierarchical structure of the overall dictionary is learned such that the upper-level dictionaries are specific in representing patterns common across a wide set of class images while lower-level dictionaries are specific in representing patterns localized to a narrow set of class images. The representation is as distinct as the paths to each of the class in the hierarchical structure are divergent and the learned dictionary is efficient in its use of the bases, and leads to a more discriminative representation than that led by previous dictionaries which is devoid of any structure and contains redundant bases. To evaluate the effectness of HSDDL, classification is performed on three benchmark datasets: Extended Yale B database, Caltech 101 and Caltech 256 dataset, and based on a common features, the proposed algorithm performs better than other state-of-the-art dictionary learning algorithms.

6. ACKNOWLEDGE

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No.NRF-2010-0028680) and by the Industrial Strategic Technology Development Program(10044009) funded by the Ministry of Knowledge Economy(MKE, Korea)

7. REFERENCES

[1] Michael Elad and Michal Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *Image Processing, IEEE Transactions on*, vol. 15, no. 12, pp. 3736–3745, 2006.

[2] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma, "Image super-resolution via sparse representation," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2861–2873, 2010.

[3] Zhuolin Jiang, Zhe Lin, and Larry S Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *CVPR*, 2011.

[4] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang, "Fisher discrimination dictionary learning for sparse representation," in *ICCV*, 2011.

[5] Ning Zhou, Yi Shen, Jinye Peng, and Jianping Fan, "Learning inter-related visual dictionary for object recognition," in *CVPR*, 2012.

[6] Shu Kong and Donghui Wang, "A brief summary of dictionary learning based approach for classification (revised)," *arXiv preprint arXiv:1205.6544*, 2012.

[7] Qiang Zhang and Baoxin Li, "Discriminative k-svd for dictionary learning in face recognition," in *CVPR*, 2010.

[8] Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *CVPR*, 2010.

[9] Shu Kong and Donghui Wang, "A dictionary learning approach for classification: separating the particularity and the commonality," in *ECCV*. 2012.

[10] Kjersti Engan, Sven Ole Aase, and John Håkon Husøy, "Multi-frame compression: Theory and design," *Signal Processing*, vol. 80, no. 10, pp. 2121–2140, 2000.

[11] Stephane G Mallat and Zhifeng Zhang, "Matching pursuits with time-frequency dictionaries," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, 1993.

[12] Yagyensh Chandra Pati, Ramin Rezaiifar, and PS Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*. IEEE, 1993, pp. 40–44.

[13] Bruno A Olshausen, David J Field, et al., "Sparse coding with an overcomplete basis set: A strategy employed by vi?," *Vision research*, vol. 37, no. 23, pp. 3311–3326, 1997.

[14] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng, "Efficient sparse coding algorithms," *Advances in neural information processing systems*, vol. 19, pp. 801, 2007.

[15] Huimin Guo, Zhuolin Jiang, and Larry S Davis, "Discriminative dictionary learning with pairwise constraints," in *ACCV 2012*, pp. 328–342. Springer, 2013.

[16] Ke Huang and Selin Aviyente, "Sparse representation for signal classification," in *Advances in neural information processing systems*, 2006, pp. 609–616.

[17] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[18] Amir Beck and Marc Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[19] Y-L Boureau, Francis Bach, Yann LeCun, and Jean Ponce, "Learning mid-level features for recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2559–2566.

[20] Michal Aharon, Michael Elad, and Alfred Bruckstein, "K-svd: Design of dictionaries for sparse representation," *Proceedings of SPARS*, vol. 5, pp. 9–12, 2005.

[21] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 689–696.

[22] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 6, pp. 643–660, 2001.

[23] Li Fei-Fei, Rob Fergus, and Pietro Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *CVIU*, 2007.

[24] Gregory Griffin, Alex Holub, and Pietro Perona, "Caltech-256 object category dataset," *California Institute of Technology*, 2007.

[25] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR*, 2009.