Improved Dictionary Learning Scheme with Cross-Label for Scene Classification

Tian Zhou, Liya Huang, Sujuan Yang, Yu Zhao and Guan Gui

College of Telecommunication and Information Engineering
*Nanjing University of Posts and Telecommunications, Nanjing 210003, China
E-mails: huangly@njupt.edu.cn, guiguan@njupt.edu.cn

***Abstract*—In recent decades, dictionary learning has been attracted strong attentions due to its great performances when applied in many applications such as signal reconstruction and scene classification. Conventional dictionary learning schemes have been developed to the scene classification but it hard to make a suitable tradeoff between accuracy and efficiency. This paper proposed an improved method of dictionary learning with cross-label and group regularization, in order to achieve a tradeoff between accuracy of classification and efficiency of algorithm execution. Demonstrated by the experiment results on the Scene15 dataset. Our proposed approach method is obviously improved owing to the algorithm can not only obtain a desired classification performance, but also decrease much of the computational time during the process of dictionary learning.**

Keywords: Cross-label suppression, dictionary learning, scene classification, compressive sensing.

I. INTRODUCTION

Dictionary learning has been revealed to achieve a good performance in signal processing over the past a few decades, and have been applied to many aspects, such as, signal reconstruction [1]–[3], cluster [4] and classification [5]. Moreover, dictionary learning is such a method to reconstruct a physical signal from its specific sparse representation through a learnt matrix which made up with the linear combination of representative vectors, where the significant learnt matrix is named dictionary and each column of the dictionary is called 'atom'.

The method of SVD is adopted by the classic K-SVD to learn the atoms in dictionary one by one [6]. A structured dictionary model has been utilized in [7] to decrease the computational time in learning process. Supervised learning methods is demonstrated that to have more advantages for pattern classification, based on this, the sparse representation based classification (SRC) method proves it [5].

In most cases, ℓ_0-norm [8] or ℓ_1-norm [9] is commonly utilized to as a prior item avoid overfitting. However, the redundancy of computational time is obviously a major drawback in these works. Thus, ℓ_2-norm can be adopted in dictionary learning to obtain an accurate solution with closed form, also smooth the iterative manner. Furthermore, cross-label suppression as well as group regularization (CLS-GR) is adopted to achieve a higher accuracy in scene classification.

In this paper, we propose an improved method in order to obtain a tradeoff between accuracy and efficiency with cross-label suppression and group regularization. It is obviously observed that, if the dimension of the sparse codes is too few, the sparse representations do not have the ability to understand all the valid information of the physical signals. In contrast, if the dimension of the sparse representation is too high, then the computational complexity of the training samples will suffer a disastrous growth. What's more, a quantitative relationship is built between sparse dimension and cost time without traversing the situation of all dimensions. Owing to the purpose of saving time, a model inspired by K-SVD is utilized to estimate the residuals of signal recovery without operating the algorithms.

II. BASIC MODEL OF DICTIONARY LEARNING WITH CROSS – LABEL AND GROUP REGULARIZATION

In this section, we briefly introduce the basic model of dictionary learning model with cross-label as well as the group regularization, and the estimation model inspired by the classic dictionary learning method K-SVD.

A. Generally Dictionary Model with CLS-GR [10]

Assume that C classes of samples are given in total, the CLS-GR of dictionary learning model [4] can be formulated as (1).

$$\begin{aligned}\min_{\mathbf{D},\mathbf{X}} \ & \sum_{c=1}^{C}\left(\| \mathbf{Y}^c - \mathbf{D}\mathbf{X}^c \|_F^2 + \beta\|\mathbf{X}^c\|_F^2\right) \\ & + \lambda\sum_{c=1}^{C}\left(\|\mathbf{P}^c\mathbf{X}^c\|_F^2\right) + \gamma tr\left(\mathbf{X}^c\tilde{\mathbf{L}}^c(\mathbf{X}^c)\right)^T \\ \text{s. t.} \ & \| d_k \|_2 = 1, \forall k\end{aligned} \tag{1}$$

where $\mathbf{Y} \in R^{m\times p}$ denotes the image samples of the facial features and the superscript c represents that the sample belongs to the c-th class, $\mathbf{D} \in R^{m\times n}$ is the learnt dictionary obtained by learning of the training samples, $\mathbf{X} \in R^{n\times p}$ represents the specific sparse codes, and $\mathbf{P}^c \in R^{n\times n}$ is the matrix for the cross-label and is defined as (2).

$$\mathbf{P}^c(m,n) = \begin{cases} 1 & m = n,\ m \notin (L^c \bigcup L^0) \\ 0 & \text{elsewhere} \end{cases} \tag{2}$$

where $c = 1,2,\ldots,C$, $\mathbf{P}^c(m,n)$ denotes the (m,n)-th entry

of $\mathbf{P}^c$. Thus, the Laplacian matrix of an N-vertex is defined as (3).

$$\tilde{\mathbf{L}} = \mathbf{M}^{-\frac{1}{2}}\mathbf{L}\mathbf{M}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{M}^{-\frac{1}{2}}\mathbf{W}\mathbf{M}^{-\frac{1}{2}} \tag{3}$$

where $\mathbf{W}$ and $\mathbf{M}$ are the adjacency matrix of the graph and the degree matrix, respectively. Moreover, with K maps in total, the total variation ($TotalVar(f)$) can be obtained as the following form.

$$\begin{aligned} TotalVar(f) &= \sum_{k=1}^{K} \mathbf{f}_k^T \tilde{\mathbf{L}} \mathbf{f}_k \\ &= \mathrm{tr}\left(\begin{bmatrix} \mathbf{f}_1^T \\ \mathbf{f}_2^T \\ \cdots \\ \mathbf{f}_K^T \end{bmatrix} \tilde{\mathbf{L}} [\mathbf{f}_1, \cdots, \mathbf{f}_k] \right) \\ &= \mathrm{tr}\left(\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T \right) \end{aligned} \tag{4}$$

In aspect of mathematics, the smaller $TotalVar(f)$ is kept, the smoother the map is, thereby reducing the differences between the images that belong to the same class.

B. Corresponding dimensions of the residuals inspired by K-SVD

Inspired by the classic K-SVD [1] and considering the CLS-GR method of dictionary learning, the energy of sparse representations is concentrated on some particular areas of the structured matrix. Thus, the residuals can be described as (5).

$$\| \mathbf{Y}^c - \mathbf{D}\mathbf{X}^c \|_F^2 \approx \| \mathbf{Y}^c - \bar{\mathbf{D}}^c \bar{\mathbf{X}}_{L^c}^c \|_F^2 \tag{5}$$

Similarly, take $\mathbf{Y}^c$ separated by SVD, the residuals can be estimated as (6).

$$\begin{aligned} \| \mathbf{Y}^c - \mathbf{D}\mathbf{X}^c \|_F^2 &= \| \mathbf{Y}^c - \textstyle\sum_{j \notin L^c} \mathbf{d}_j \bar{\mathbf{x}}_j^c - \mathbf{D}^c \mathbf{X}_{L^c}^c \|_F^2 \\ &= \| \mathbf{E}^c - \bar{\mathbf{D}}^c \bar{\mathbf{X}}_{L^c}^c \|_F^2 \\ &\approx \| \mathbf{Y}^c - \textstyle\sum_{i=1}^{s} \mathbf{\Lambda}(i,i) \mathbf{U}_{c=i} \mathbf{V}_{c=i}^T \|_F^2 \end{aligned} \tag{6}$$

where $\mathbf{D}^c$ is a matrix with s columns; next, take $\mathbf{Y}^c$ to conduct the method of SVD, $\mathbf{Y}^c = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$. Inspired by the K-SVD algorithm, we denote the first s columns of $\mathbf{U}$ corresponding to the number of columns of $\mathbf{D}^c$ with the notation $\bar{\mathbf{D}}^c$, where $\mathbf{U}_{c=i}$ represents the i-th column in $\mathbf{U}$. Moreover, the diagonal matrix composed of the first s singular values multiplying the transpose of the first s columns in $\mathbf{V}^T$ corresponds to $\mathbf{X}_{L^c}^c$, where $\bar{\mathbf{X}}_{L^c}^c$ denotes the rows in $\bar{\mathbf{X}}^c$ that belong to the c-th Label. The residuals can be given as (7).

$$\| \mathbf{Y}^c - \mathbf{D}\mathbf{X}^c \|_F^2 \approx \left(\sum \mathbf{\Lambda}^2(i,i) \right) - \textstyle\sum_{i=1}^{s} \mathbf{\Lambda}^2(i,i) \tag{7}$$

Particularly, s denotes the dimension of sparse representation. According to the proposed estimation model inspired by K-SVD with low complexity, the residuals can be predicted without practical experiments.

III. Improved Model of Dictionary Learning

A. Cubic fitting method for time estimation

Considering the time complexity of each step, the time complexity of k-means [5] is $O(p \cdot n \cdot m)$, where p represents the number of samples, n represents the number of clusters as well as the number of atoms in dictionary, m represents the number of features of each sample.

In the part of initializing the sparse representations, given the initialized dictionary $\mathbf{D}_0$ by k-means, $\mathbf{X}_0$ can be obtained as (8).

$$\mathbf{X}_0 = (\mathbf{D}_0^T \mathbf{D}_0 + \beta \mathbf{E})^{-1} \mathbf{D}_0^T \mathbf{Y} \tag{8}$$

where the time complexity of initial method is $O(n^3)$.

Instead of renewing the whole dictionary at the same time, the model updates the dictionary atom by atom to fully utilize those that have already been updated. The dictionary is updated by the formula as (9) and (10).

$$\tilde{\mathbf{Z}} = \mathbf{Y} - \sum_{k \notin L^c} \mathbf{d}_k \bar{\mathbf{x}}_k - \sum_{k \in L^c, k \neq i} \mathbf{d}_k \bar{\mathbf{x}}_k \tag{9}$$

$$\hat{\mathbf{d}}_i^c = \frac{\tilde{\mathbf{Z}} \cdot \bar{\mathbf{x}}^T}{\| \tilde{\mathbf{Z}} \cdot \bar{\mathbf{x}}^T \|_2} \tag{10}$$

where $\hat{\boldsymbol{d}}_i^c$ represents the estimated value of i-th column which belongs to the c-th class, $\bar{\mathbf{x}}_k$ denotes estimated value of the k-th row in the whole code matrix $\mathbf{X}$. Its time complexity is $O(p \cdot n \cdot m)$, where n represents the number of atoms of the dictionary, m represents the number of features of each sample, p represents the number of samples. A parameter estimation method of LMS (least mean square) is adopted to update the sparse representation $\mathbf{X}^c$ of the c-th class, formulated as (11).

$$\hat{\mathbf{X}}^c = [\mathbf{D}^T\mathbf{D} + \lambda(\mathbf{P}^c)^T(\mathbf{P}^c) + (\beta+\gamma)\mathbf{E}]^{-1}(\mathbf{D}^T\mathbf{Y}^c - \gamma\mathbf{X}^c\bar{\mathbf{L}}^c) \tag{11}$$

where the time complexity is $O(n^3)$ and n represents the number of atoms of the dictionary matrix. According to this, the time cost can be obtained by (12).

$$\textstyle\sum_{c=1}^{C} t_c \approx \omega_1 \cdot s^3 + \omega_2 \cdot s^2 + \omega_3 \cdot s^1 + \omega_4 \tag{12}$$

where $\omega_1, \omega_2, \omega_3$, and ω_4 represent coefficients of the cubic function, and $\sum_{c=1}^{C} t_c$ represents the total time cost of the algorithm. For the least-square solution, (12) can be rewritten as an optimization model as (13) and (14).

$$\arg\min_{\omega_i} \sum\nolimits_s \left(\textstyle\sum_{c=1}^{C} t_c - \hat{t}(s) \right)^2 \tag{13}$$

$$\hat{t} = \omega_1 \cdot s^3 + \omega_2 \cdot s^2 + \omega_3 \cdot s^1 + \omega_4 \tag{14}$$

where $\hat{t}(s)$ represents the estimated value of the total cost time of operating the method. Through determining a reasonable number of samples for cubic fitting, a more precise estimated value of the time cost of the algorithms can be obtained without traversing all the situations.

B. Optimized object corresponding to the sparse dimension

In order to achieve a balance between residual and accuracy, the improved approach method for extracting the information of dimension can be defined as (15).

$$\begin{cases} \min\limits_{s} \sum_{c=1}^{C} \left(\| \mathbf{Y}^c - \mathbf{D}\mathbf{X}^c \|_F^2 + \delta t^c(s) \right) \\ \min\limits_{\mathbf{D},\mathbf{X}} \sum_{c=1}^{C} \left(\| \mathbf{Y}^c - \mathbf{D}\mathbf{X}^c \|_F^2 + \beta \|\mathbf{X}^c\|_F^2 \right) \\ \quad + \lambda \sum_{c=1}^{C} \left(\|\mathbf{P}^c \mathbf{X}^c\|_F^2 \right) + \gamma tr\left(\mathbf{X}^c \tilde{\mathbf{L}}^c (\mathbf{X}^c) \right)^T \\ \text{s. t.} \quad \| d_k \|_2 = 1, \forall k \end{cases} \tag{15}$$

Mathematically, the LMS is utilized to obtain a solution of the above optimization model with closed form as (16).

$$\begin{aligned} s &= \arg\min_{s} \sum_{c=1}^{C} \| \mathbf{Y}^c - \bar{\mathbf{D}}^c \bar{\mathbf{X}}_{L^c}^c \|_F^2 + \delta' \hat{t}^c(s) \\ &= \arg\min_{s} \frac{1}{C} \sum_{c=1}^{C} \left(1 - \frac{\sum_{i=1}^{s} (\Lambda^c(i,i))^2}{\sum (\Lambda^c(i,i))^2} \right) \\ &\quad + \delta'(\omega_1 \cdot s^3 + \omega_2 \cdot s^2 + \omega_3 \cdot s^1 + \omega_4) \end{aligned} \tag{16}$$

where δ' is the normalization parameter to balance the magnitude of residuals and cost time, the value is defined as (17).

$$\delta' = \frac{\| \mathbf{Y}^c \|_F^2 \cdot \delta}{\hat{t}^c(trNum)} = \frac{p \cdot \delta}{C \cdot \hat{t}^c(trNum)} \tag{17}$$

Thus, the normalized residuals and the cost time can be obtained, and the hyper-parameter $\boldsymbol{\delta}$ is utilized to measure the influence of the cost time in the model.

C. Global Coding Classifier

A global coding classifier with reasonable sparse dimension (RSD-GCC) is adopted to predict the unknown label of samples. Given an scene image sample $\mathbf{y}$ and the dictionary $\mathbf{D}$ learnt by the training samples, the sparse representation with a general model can be defined as

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \| \mathbf{y} - \mathbf{D}\mathbf{x} \|_2^2 + \beta \| \mathbf{x} \|_2^2 = (\mathbf{D}^T\mathbf{D} + \beta\mathbf{I})^{-1}\mathbf{D}^T\mathbf{y} \tag{18}$$

Due to the structured dictionary [7] is adopted in the proposed learning method, if the image in form of column vector $\mathbf{y}$ belongs to the c-th class, the large coefficients shall be distributed in the particular rows of sparse matrix belonging to the c-th label, and the predicted label can be obtained from (19).

$$\text{label}(\mathbf{y}) = \arg\min_{c} \frac{\| \mathbf{y} - \sum_{k \notin L^c} \mathbf{d}_k \mathbf{x}_{L^c} \|_2^2}{\sum_{k \notin L^c} | \mathbf{x}_{L^c} |} \tag{19}$$

where $\bar{\mathbf{X}}_{L^c}$ represents the rows belonging to the L^c of the sparse code corresponding to the image signal.

IV. Experimental Results and Discussions

In this section, some experiments on Scene15 databases [12] are set for scene classification. Because both the efficiency and residuals are taken into consideration, we will show both the two aspects in our experiment; the performances of our approach are demonstrated in the following results. In the experiments, to fairly evaluate the computational efficiency, the proposed approach will be operated on the platform of MATLAB2017b applied in one PC with a 64-bit Windows 10 operating system and equipped with Intel i7-6700H 3.4 GHz CPU, and 8 GB memory.

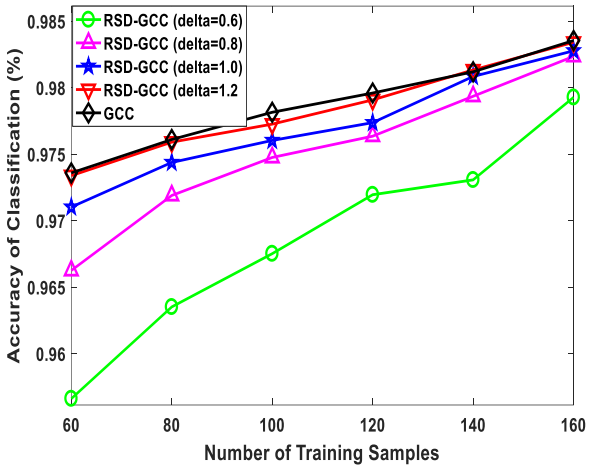

Fig. 1 . Accuracies of scene classification on the Scene15 dataset.

In this section, the Scene15 dataset is considered by us, each category has 200 to 400 images with a total number of 4485, and the average image size is about 250 × 300 pixels. For a fair comparison, the 3000-dimensional SIFT based features applied by LC-KSVD [13] is adopted.

In order to evaluate different algorithms fairly, the parameters β, γ, λ are set to 2×10^{-3}, 1 and 2×10^{-1} respectively, the value of δ is increasing from 0.6 to 1.2 with a step of 0.2. In order to acquire a stable classification rate, each situation is operated for 30 times, performances of the two classifications scheme are shown from two parts including the average accuracies and the average time costs of training samples as the followings.

According to Fig. 1 and Fig. 2, compared to GCC, our proposed approach method corresponding to a more reasonable sparse dimension can not only obtain a desired accuracy but also can decrease much redundancy during the process of computation on the Scene15 dataset. For example, in Fig. 1, our proposed RSD-GCC sacrifices about 0.1 to 2 percent of the classification accuracy, but save about 30 to 60 percent of the cost time. In another way, the accuracy function is convergent. Demonstrated by the experimental results, the dictionary learning model of CLS-GR dismisses the information of sparse dimension, and results in much computational redundancy. Our

proposed approach method solve the problem, and obtain a great tradeoff of between accuracy and efficiency.

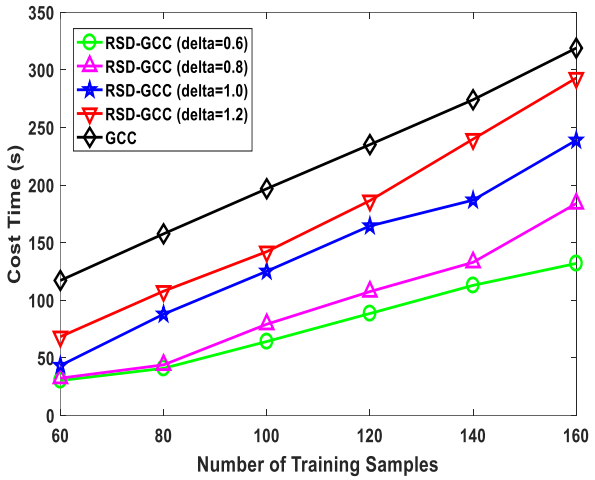

Fig. 2. Time costs of scene classification on the Scene15 dataset.

V. CONCLUSIONS

In this paper, a cubic fitting model was utilized for quantization of the time cost in dictionary learning methods, and an estimation method based on SVD was adopted to estimate the residuals in signal reconstruction, to consider the tradeoff between accuracy and computational complexity. The experimental results show that our proposed approach method can effectively improve the dictionary learning method of CLS-GR from both aspects of accuracy and efficiency.

REFERENCES

[1] 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, 2006.

[2] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Res*., vol. 37, no. 23, pp. 3311–3325, 1997.

[3] Y. Li et al., "Sparse Adaptive Iteratively-Weighted Thresholding Algorithm (SAITA) for Lp-Regularization Using the Multiple Sub-Dictionary Representation," *Sensors*, vol. 17, no. 12, pp. 2920–2936, 2017.

[4] F. Wang, N. Lee, J. Sun, J. Hu, and S. Ebadollahi, "Automatic Group Sparse Coding," *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence* (AAAI), San Francisco, California, USA, August 7-11, 2011.

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

[6] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Expert Syst. Appl*., vol. 40, no. 1, pp. 200–210, 2013.

[7] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Trans. Signal Process*., vol. 58, no. 3, pp. 1553–1564, 2010.

[8] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*., pp. 2691–2698, 2010.

[9] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3501–3508.

[10] X. Wang and Y. Gu, "Cross-Label Suppression: A Discriminative and Fast Dictionary Learning with Group Regularization," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3859–3873, 2017.

[11] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[12] S.Lazebnik, C.Schmid, and J.Ponce, "Beyond bags of features:spatial pyramid matching for recognizing natural scene categories," in Proc. *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2169-2178, 2007.

[13] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, 2013.