Jointly Class-specific and Shared Discriminative Dictionary Learning for Classifying Surface Defects of Steel Sheet

Shiyang ZHOU,^{1,2)} Huaiguang LIU,¹⁾ Ketao CUI¹⁾ and Zhiqiang HAO^{1)*}

1) Key Laboratory of Metallurgical Equipment and Control Technology, Ministry of Education, School of Machinery and Automation, Wuhan University of Science and Technology, Wuhan, 430081 China.

2) Key Laboratory of Image Processing and Intelligent Control (Huazhong University of Science and Technology), Ministry of Education, Wuhan, 430074 China.

(Received on March 24, 2021; accepted on August 10, 2021; J-STAGE Advance published date: September 27, 2021)

For surface defect images that captured from a practical steel production line, feature extraction for classification is a challenging task due to inter-class similarity and intra-class difference of defect images. In this paper, we propose a class-specific and shared discriminative dictionary learning (CASDDL) method for extracting discriminative class-specific features to classify surface defects of steel sheet. Specifically, we learn a shared sub-dictionary as well as several class-specific sub-dictionaries to explicitly capture common information shared by all classes and class-specific information belonging to corresponding class. We adopt a mutual incoherence constrain for each sub-dictionary to encourage learned sub-dictionaries to be as independent as possible. A Fisher-like discriminative criterion is also introduced to coding vectors over all the class-specific sub-dictionaries, which can indirectly improve the discriminative ability of learned dictionary. In addition, we further impose a low-rank constrain on coding vector over shared sub-dictionary to guarantee learned dictionary has the abilities of same class of samples itself reconstruction and different class of samples reconstruction. Finally, classification can be efficiently performed by discriminative coding vector that obtained from a reconstructive and discriminative dictionary. Experimental results demonstrate that our proposed CASDDL method achieves a comparable or better performance than the state-of-the-art dictionary learning methods in classifying surface defects of steel sheet.

KEY WORDS: class-specific and shared dictionary; discriminative dictionary learning; image classification; surface defects of steel sheet.

1. Introduction

Though surface defects classification in steel sheet images has been studied extensively over the past decade, it still stands as a challenging problem owing to complex of defects object, whose representative factors include changes of shape, size, gray, texture and location. The approaches developed in previous works focused on two aspects of feature extraction and classifier design.^{1,2)} However, these methods are mostly exclusively designed for particular data sets and are highly dependent on specific hand-crafted features, limiting their performance on general classification of surface defect of steel sheet.

In the past few years, dictionary learning has been successfully applied to many machine vision problems and shown impressive performance, such as face recognition,³⁾ person re-identification,⁴⁾ image segmentation,⁵⁾ image

classification,⁶⁾ and surface defect classification of industrial products.⁷⁾ Although deep learning has achieved good performance in image classification,^{8,9)} it usually needs a large number of labeled training samples to obtain enough features. When the amount of labeled data is small, the deep model tends to over-fit the data easily. In addition, it consumes a significant amount of computing resources to keep running as its complex network structure. Besides, there is still little theoretical understanding of learning process in these networks. Many surface defective samples are difficult to obtain because of probability of defect occurrence is very low in industrial manufacturing, dictionary learning is more suitable and efficient for classifying surface defects of steel sheet, compared with deep learning.

Dictionary learning, as a fundamental problem in sparse representation, tries to learn a desired dictionary from training samples instead of using off-the-shelf bases. Then, the test sample can be sparsely represented by a small number of atoms of learned dictionary, and final classification can

* Corresponding author: E-mail: haozhiqiang@wust.edu.cn

be conducted by calculating the minimum reconstruction error. Sparse representation based classification (SRC)¹⁰⁾ method used original training data as dictionary directly, which show impressive performance in comparison with many conventional methods.¹¹⁾ Nevertheless, many methods demonstrate that the dictionary learned from training data can achieve better classification performance.^{12,13)} Aharon *et al.*¹⁴⁾ proposed K-SVD method to learn an overcomplete dictionary from original training data. Since the main goal of K-SVD is to reconstruct the training samples well, it's not applicable for classification tasks. Therefore, some discriminative dictionary learning approaches make full use of class information of training samples can improve joint representation and classification abilities. For example, Zhang *et al.*¹⁵⁾ proposed a discriminative K-SVD method to improve discriminative ability of learned dictionary by incorporating a classification error term into objective function. Jiang *et al.*¹⁶⁾ introduced a label consistent K-SVD (LCKSVD) method in which introducing a class label consistence term to encourage training samples from the same class to have similar representations and those from different classes to have dissimilar representations. What's more, the structure information of training samples is used to improve generalization and discriminability of dictionary. Ramirez *et al.*¹⁷⁾ developed a structured incoherence regularization term for dictionary learning (DLSI) to promote the independence between different sub-dictionaries. Gao *et al.*,¹⁸⁾ Li *et al.*¹⁹⁾ imposed graph Laplacian on coding vectors to preserve locality of training data. Yang *et al.*²⁰⁾ exploited Fisher discrimination dictionary learning (FDDL) to encourage coding vectors have both minimum within-class scatter and maximum inter-class scatter.

However, above methods still have a problem that learned class-specific sub-dictionaries have some similar or even same atoms if the training samples from different classes have some common components. These similar atoms from different class-specific sub-dictionaries might be helpful to better represent a test sample, but they may weaken the discriminative ability of learned dictionary. As shown in **Fig. 1**, defect object in the surface image can be regarded as local anomaly against relatively homogeneous background. The background texture of surface image, which is useful for reconstruction rather than discrimination, is nearly the same. For the aforementioned dictionary learning methods, in the worst case, most of atoms of learned dictionary are used to represent the background, causing only a small amount of atoms represent class-specific defect object. Therefore, the discrimination of coding vectors between different defect object will diminish, greatly degrading the classification performance. An intuitive way to address this issue is to

specially capture and separate those shared components from training samples.

Recent research efforts have yielded more promising results for image classification by using the idea of shared dictionary, which different classes not only have class-particular parts but also share commonality.^{21,22)} To fully explore the advantage of class-specific and shared sub-dictionaries, Zhou *et al.*²³⁾ constructed a joint dictionary learning algorithm to learned a hybrid of shared and class-specific dictionary with a discrimination promotion term according to the Fisher discrimination criterion. Gao *et al.*²⁴⁾ developed both some category-specific sub-dictionaries and a shared sub-dictionary by imposing cross-incoherence constraint between different sub-dictionaries and self-incoherence constraint in each sub-dictionary. Wang *et al.*²⁵⁾ established a category-specific and shared dictionary learning method (COPAR) by exploiting the information of particularity and commonality across all classes. Lin *et al.*²⁶⁾ constructed a class-shared, class-specific and disturbance dictionary by introducing a robust, discriminative and comprehensive dictionary learning (RDCDL) method. These methods overlook the low-rank ability of sub-dictionaries or coding vector over the shared sub-dictionary. To address this issue, some dictionary learning methods have been proposed by integrating low-rank constrains, which achieved impressive results. Jiang *et al.*²⁷⁾ proposed a sparse and dense hybrid representation method based on a supervised low-rank dictionary decomposition. Rong *et al.*²⁸⁾ used low-rank constrains on both class-specific sub-dictionary and class-shared sub-dictionary. Wen *et al.*²⁹⁾ introduced low-rank regularizations for feature matrices over a shared sub-dictionary. Furthermore, Vu *et al.*³⁰⁾ proposed a low-rank constraint on the shared dictionary (LRSDL) to encourage its subspace to be of low-dimensionality and its corresponding representations to be similar. Du *et al.*³¹⁾ presented a low-rank graph preserving discriminative dictionary learning (LRGPDDL) method by introducing a low-rank constraint on each sub-dictionary. These methods show that incorporating low-rank regularization term into dictionary learning can enhance robustness of the learned dictionary and achieved impressive results.

Inspired by the success of shared sub-dictionary, and taking advantage of low-rank constrain of coding vectors, we develop a class-specific and shared discriminative dictionary learning (CASDDL) method to improve discriminative power of learned dictionary, which is suitable for surface defect classification of steel sheet. Based on the analysis of different classes of surface defect image share similar background, the proposed CASDDL method constructs c class-specific sub-dictionaries associated with correspond-

Fig. 1. Examples of surface defects images of steel sheet: (a) Craze; (b) Inclusion; (c) Patch; (d) Pitted Surface; (e) Rolled-in Scale; (f) Non-defective.

ing classes and one shared sub-dictionary for all the classes, respectively. As illustrated in Fig. 2, each class-specific sub-dictionary is responsible for exploiting the exclusive class-specific information associated with corresponding classes, and a shared sub-dictionary is used to capture the explicit and hidden common information shared by all the classes. With these sub-dictionaries, the exclusive pattern and the shared pattern of surface defect image can be explicitly separated. CASDDL specially introduces incoherence promoting constraints on all the sub-dictionaries and low-rank constraints on coding vector over shared sub-dictionary to make learned dictionary more compact, discriminative and robust to variations. Also, a Fisher-like regularization term on coding vectors over class-specific sub-dictionaries ensures more coherence for within-class coding vectors and more disparity for between-class coding vectors, which enhance class discrimination further. Such process can considerably influence the quality of learned dictionary. In this way, the corresponding coding vectors will be more representative, discriminative as well as robust, which may directly improve subsequent classification performance. This paper is an extension of our previous works of class-specific and shared dictionary learning (CSDSL)³²⁾ with significant new proposals and more experiments and comparisons: firstly, we separate the coding vector over a shared sub-dictionary and class-specific sub-dictionaries, and exploit a class-specific and shared discriminative dictionary learning (CASDDL) method for classifying surface defects of steel sheet, which not only encourages intra-class samples to deliver the similar feature representation vector, but also minimizes the inter-class samples correlations; secondly, we constrain the coding vectors corresponding to the shared dictionary is low-rank, which is able to improve the adaptivity of defect classification and greatly reduces computation time; finally, more experiments and comprehensive comparisons are reported.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce some related works about classification of surface defects of steel sheet, basic formulations of dictionary learning based classification, respectively. Section 3 presents our proposed CASDDL method, including formulation and optimization. In Section 4, we validate proposed method in extensive experiments and compare it with the state-of-the-art methods. Some comments and future works are finally provided in Section 5.

2. Related Work

We briefly introduce two research topics that are related to our approach in this section, including surface defect classification, and dictionary learning.

2.1. Surface Defect Classification

For classifying surface defects of steel sheet based on machine vision, researchers often look for problem-specific visual cues, or features, in surface defect images in order to categorize a tissue image as one of the possible categories. These features might come from the distinguishable characteristics of defect, for example, size, shape or texture. Consequently, different customized feature extraction methods for a variety of problems have been developed based on these observed features. The representative feature extraction methods mainly include grayscale, shape, texture, morphological operator, gray-level co-occurrence matrix, local binary pattern, Fourier, Gabor and wavelet transform. For classification purposes, these features are combined with powerful classifiers, such as K-nearest neighbor (KNN), artificial neural networks, support vector machines (SVMs). Wang *et al.*³³⁾ used improved random forest algorithm with optimal multi-feature-set fusion for distributed defect recognition. Borwankar *et al.*³⁴⁾ used the discrete wavelet transform and rotated wavelet transform for feature extraction, while a KNN classifier for classification. Luo *et al.*³⁵⁾ exploited a generalized completed local binary patterns framework and simple nearest-neighbor classifier for steel surface defect classification. Ashour *et al.*³⁶⁾ developed a method combining the use of discrete shearlet transform and gray-level co-occurrence matrix to classify surface defects of hot-rolled steel strips.

2.2. Dictionary Learning

Mathematically, dictionary learning can be formulated as follows:

$$\min_{D, \mathbf{x}} \|\mathbf{y} - D\mathbf{x}\|_2^2 + \lambda \theta(D, \mathbf{x}) \dots \dots \dots (1)$$

where, $\|\cdot\|_2$ denotes l_2 norm, $\mathbf{y} \in \mathbb{R}^d$ denotes a given d -dimensional feature vector of training sample, $\mathbf{x} \in \mathbb{R}^K$ denotes coding vector of $\mathbf{y}$ over dictionary $D = (\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_j, \dots, \mathbf{d}_K) \in \mathbb{R}^{d \times K}$, $\mathbf{d}_j \in \mathbb{R}^d$ denotes the j -th atom of D , $\theta(D, \mathbf{x})$ denotes a regularization term to constrain D or $\mathbf{x}$, λ is a positive parameter that balances the tradeoff between recon-

Fig. 2. An illustration of the proposed CASDDL method, each image can be sparsely represented by a few atoms from the corresponding class-specific sub-dictionary $\{D_j\}_{j=1, 2, \dots, c}$ and a shared sub-dictionary D_{c+1} (brick-red). (Online version in color.)

structive error $\|y - Dx\|_2^2$ and $\theta(D, x)$.

For the classification task, recent researches on discriminative dictionary learning have demonstrated that a well-learned dictionary D will greatly boost classification performance. The discrimination could be developed from the dictionary, coding vectors, or both. Several regularization terms, such as sparsity, low-rank, neighborhood preservation of graph,³⁷⁾ entropy,³⁸⁾ small within-class scatter but large between-class scatter on coding vectors, incoherence constraint on sub-dictionaries, have been introduced into learning process to promote the discriminative power of learned dictionary.

Optimizing Eq. (1) jointly over D and x is a non-convex problem, but is convex with respect to each variable when the others are fixed. It's carried out by an iterative method composing two steps: (a) fixing D to update x ; (b) fixing x to update D . It can be solved efficiently by many kinds of algorithms.^{39,40)}

After coding vector x is obtained, test sample $\hat{y}$ is recognized as belonging to class k^* if it satisfies: $k^* = \operatorname{argmin}_k \|\hat{y} - D l_k(x)\|_2^2$, where, $l_k(x)$ denotes a vector only keeping the entries of x associated with the k -th class and changing others into zeros. As a result, $\hat{y}$ is assigned to the class k^* corresponding to the minimum reconstruction error $\|\hat{y} - D l_{k^*}(x)\|_2^2$.

3. Proposed Class-specific and Shared Discriminative Dictionary Learning (CASDDL) Framework

3.1. Formulation of CASDDL

Suppose $Y = [Y_1, Y_2, \dots, Y_i, \dots, Y_c] \in \mathbb{R}^{d \times N}$ denotes whole training samples of c classes, each column denotes one sample, where, $Y_i \in \mathbb{R}^{d \times n_i}$ denotes the i -th class training samples, d is dimension of one sample, n_i denotes number of sample from class i , $\sum_{i=1}^c n_i = N$, N is total number of training samples. Let $D = [D_1, D_2, \dots, D_j, \dots, D_c, D_{c+1}] = [D_{class}, D_{c+1}] \in \mathbb{R}^{d \times K}$ denotes learned dictionary of K atoms, $\{D_j\}_{j=1,2,\dots,c} \in \mathbb{R}^{d \times k_j}$ denotes the j -th class-specific sub-dictionary that trained from a corresponding training samples Y_i , $D_{c+1} \in \mathbb{R}^{d \times k_{c+1}}$ denotes a shared sub-dictionary that trained from whole training samples Y , where, $K = \sum_{j=1}^{c+1} k_j$, k_j denotes number of atoms from the j -th sub-dictionary. Let $X = [X_1, X_2, \dots, X_i, \dots, X_c] \in \mathbb{R}^{d \times N}$ denotes

coding matrix of Y over D , where, $X_i \in \mathbb{R}^{K \times n_i}$ denotes coding matrix of Y_i over D . Furthermore, X_i can be written as $X_i = [X_i^1; X_i^2; \dots; X_i^j; \dots; X_i^c; X_i^{c+1}] = [X_i^{class}; X_i^{c+1}]$, where, $X_i^j \in \mathbb{R}^{k_j \times n_i}$ denotes coding matrix of Y_i over sub-dictionary D_j , $X_i^{class} \in \mathbb{R}^{(K-k_{c+1}) \times n_i}$ denotes coding matrix of Y_i over all class-specific sub-dictionaries, $X_i^{c+1} \in \mathbb{R}^{k_{c+1} \times n_i}$ denotes coding matrix of Y_i over the shared sub-dictionary D_{c+1} .

To enhance the discriminative capability of dictionary and make the coding vector have more discriminability as much as possible for surface defect classification, it's ideally desired that for each class, its samples have non-zero coding coefficients intensively locating at the corresponding atoms, whereas the coding coefficients at other atoms be zero. The ideal structure of dictionary D and coding matrix X in the proposed CASDDL method is shown in **Fig. 3**. First of all, a sample is supposed to be represented only by the corresponding class-specific sub-dictionary, while can't be represented by other class-specific sub-dictionaries at the same time. It can enhance the discriminative capability of learned dictionary by forcing that all other discriminative sub-dictionaries have poor representative capability of non-corresponding samples. Besides, different sub-dictionaries should be low coherence, which can guide learned dictionary to be discriminative. What's more, in terms of intra-class compactness and inter-class separability, the coding vector of same samples class should be similar, while the coding vector of different samples class should be dissimilar. By directly constraining coding vectors, the separability and discriminability of coding vectors from different classes is increased and further enhanced. Lastly, according to the coding vectors corresponding to the shared dictionary should be similar, the corresponding coding matrix should be low-rank, which well addresses the redundant information in the shared sub-dictionary and promotes coding vectors compact. Not only the locality structured information of samples is preserved, but also it can avoid very large values in the coding vectors to reduce the error caused by overfitting.

Based on above discussions, the proposed CASDDL method can be modelled as a unitive framework which helps to relieve from the local optimal solution, including reconstructive error term, discriminative promotion constraint term, and low-rank regularization term. It can be formulated by the following optimization problem:

$$\min_{D, X} L_{reconstruction}(Y, D, X) + L_{incoherence}(D_i, D_j) + \dots + L_{exclusiveness}(X_i^{class}) + L_{lowrank}(X_i^{c+1}) \quad (2)$$

Fig. 3. Ideal structure of dictionary D and coding matrix X in the proposed CASDDL method. The block-diagonal constraints increase the discriminative capability of D and X . (Online version in color.)

where, $L_1 = L_{\text{reconstruction}}(Y, D, X)$ denotes the reconstruction error term; $L_2 = L_{\text{incoherence}}(D_i, D_j)$ denotes the sub-dictionary incoherence term; $L_3 = L_{\text{exclusiveness}}(X_i^{\text{class}})$ denotes the discriminative promotion term for coding vectors over all the class-specific sub-dictionaries; $L_4 = L_{\text{lowrank}}(X_i^{c+1})$ denotes the low-rank preserving term for coding vectors over the shared sub-dictionary.

3.1.1. Reconstruction Error Term L_1

To learn a representative and discriminative structured dictionary, each class-specific sub-dictionary should be supposed to well represent samples from the i -th class, but not other classes. As the common part, the most important property of the shared dictionary is to represent samples from all classes. According to $Y_i \approx DX_i = D_1X_i^1 + D_2X_i^2 + \dots + D_jX_i^j + \dots + D_cX_i^c + D_{c+1}X_i^{c+1} = D_{\text{class}}X_i^{\text{class}} + D_{c+1}X_i^{c+1}$, small value of $\|Y_i - D_{\text{class}}X_i^{\text{class}}\|_F^2$ ensures that the dictionary D can represent Y_i well, where, $\|\cdot\|_F$ denotes Frobenius-norm. Besides, small value of $\|Y_i - D_jX_i^j\|_F^2 = \|Y_i - (D_{\text{class}}S_j)(S_j^T X_i^{\text{class}})\|_F^2$ ($j = 1, 2, \dots, c$) makes sure that each class Y_i has good representation over corresponding class-specific sub-dictionary D_j , where, $S_j \in \mathbb{R}^{K \times k_j}$ is the selection operator that selects the j -th class-specific sub-dictionary D_j from D , each column of S_j has only one nonzero element 1, which the location is column index of corresponding class-specific sub-dictionary atom in D , $S_{-j} = [S_1, S_2, \dots, S_{j-1}, S_{j+1}, \dots, S_c] \in \mathbb{R}^{K \times (K - k_{c+1} - k_j)}$.

Meanwhile, the small value of $\|Y_i - D_{c+1}X_i^{c+1}\|_F^2$ ensures that the shared sub-dictionary D_{c+1} make contribution to represent Y_i . Hence, the reconstruction error term L_1 can be defined as follows:

$$L_1(Y, D, X) = \sum_{i=1}^c \left[\|Y_i - D_{\text{class}}X_i^{\text{class}}\|_F^2 + \|Y_i - (D_{\text{class}}S_i)(S_i^T X_i^{\text{class}})\|_F^2 + \|Y_i - D_{c+1}X_i^{c+1}\|_F^2 \right] \dots (3)$$

3.1.2. Sub-dictionary Incoherence Term L_2

To exploit desirable discriminative reconstruction capability of learned dictionary D , different sub-dictionaries should be as orthogonal as possible, which ensures that each class-specific sub-dictionary is exclusive to represent corresponding samples well. Therefore, the value of structural incoherence constraint $\|D_j^T D_{-j}\|_F^2$, $\|D_j^T D_j - I_{k_j}\|_F^2$ are supposed to be small, where, $D_{-j} \in \mathbb{R}^{K \times (K - k_j)}$ is the sub-matrix by removing D_j from D , I_{k_j} is an identity matrix. By adding these two terms, the redundancy among sub-dictionaries would be reduced effectively, which has a direct impact on the speed of computation. Hence, the sub-dictionary incoherence term L_2 can be defined as follows:

$$L_2 = \alpha \sum_{j=1}^{c+1} \left[\frac{n_j}{k_j(K - k_j)} \|D_j^T D_{-j}\|_F^2 + \frac{n_j}{k_j^2} \|D_j^T D_j - I_{k_j}\|_F^2 \right] \dots (4)$$

where, $n_{c+1} = N$, $\frac{n_j}{k_j(K - k_j)}$ and $\frac{n_j}{k_j^2}$ can alleviate the effect of imbalance between the number of samples and atoms.

3.1.3. Discriminative Promotion Term L_3

To further increase the discriminative capability of learned dictionary D , we enforce the coding vector over all the class-specific sub-dictionaries to be discriminative. Intuitively, the discrimination can be assessed by the similarity of pairs of coding vectors from the same training class and the dissimilarity of pairs of coding vectors from the different training classes. Based on Fisher's linear discriminant, which maximizes the ratio of between-class scatter matrix to within-class scatter matrix, we can minimize the within-class scatter matrix $S_W(X)$ and maximum the between-class scatter matrix $S_B(X)$. Denote $S_W(X) = \sum_{i=1}^c \sum_{l=1}^{n_i} (x_i^l - m_i)(x_i^l - m_i)^T$, $S_B(X) = \sum_{i=1}^c n_i(m_i - m)(m_i - m)^T$, where, x_i^l denotes the coding vector of the l -th training sample over the i -th class-specific sub-dictionary, $m_i = \frac{1}{n_i} \sum_{l=1}^{n_i} x_i^l$ and $m = \frac{1}{N} \sum_{i=1}^c \sum_{l=1}^{n_i} x_i^l$ are mean vector of X_i and X , respectively. Thus, $\text{tr}(S_W) = \sum_{i=1}^c \|X_i^{\text{class}} - M_i\|_F^2$, $\text{tr}(S_B) = \sum_{i=1}^c \|M_i - M\|_F^2$, where, $M_i \in \mathbb{R}^{K \times n_i}$, each column equals to m_i , $M \in \mathbb{R}^{K \times n_i}$, each column equals m . Hence, the discriminative coding vector term L_3 can be defined as follows:

$$L_3 = \beta \sum_{i=1}^c \left(\|X_i^{\text{class}} - M_i\|_F^2 + \|M_i - M\|_F^2 + \|X_i^{\text{class}}\|_1 \right) \dots (5)$$

3.1.4. Low-rank Preserving Term L_4

As nuclear norm $\|\cdot\|_*$ is the convex relaxation of $\text{rank}(\cdot)$, the low-rank preserving term L_4 can be defined as follows:

$$L_4 = \gamma \sum_{i=1}^c \|X_i^{c+1}\|_* \dots (6)$$

Taking all mentioned above into consideration, we have the following CASDDL model:

$$\begin{aligned} \min_{D, X} \sum_{i=1}^c \left[\|Y_i - D_{\text{class}}X_i^{\text{class}}\|_F^2 + \|Y_i - (D_{\text{class}}S_i)(S_i^T X_i^{\text{class}})\|_F^2 + \|Y_i - D_{c+1}X_i^{c+1}\|_F^2 \right] &+ \alpha \sum_{j=1}^{c+1} \left[\frac{n_j}{k_j(K - k_j)} \|D_j^T D_{-j}\|_F^2 + \frac{n_j}{k_j^2} \|D_j^T D_j - I_{k_j}\|_F^2 \right] \\ &+ \beta \sum_{i=1}^c \left(\|X_i^{\text{class}} - M_i\|_F^2 + \|M_i - M\|_F^2 + \|X_i^{\text{class}}\|_1 \right) + \gamma \sum_{i=1}^c \|X_i^{c+1}\|_* \dots (7) \end{aligned}$$

3.2. Optimization of CASDDL

The objective function Eq. (7) can be divided into two sub-problems to jointly learn dictionary D , and coding matrix X : (1) updating X with fixed D ; (2) updating D with fixed X . These two sub-problems are optimized alternatively by updating one variable and fixing the other ones, through an iterative process. In order to learn a discriminative dictionary better, K -means algorithm is chosen to initialize the dictionary at first: each class-specific sub-dictionary is initialized as cluster centers of corresponding training samples, a shared sub-dictionary is initialized as cluster center of whole training samples. As the dissimilarity between different cluster centers is high, the initial atoms in class-specific sub-dictionaries obtain approximately discrimination ability. We summarize our proposed CASDDL

method in **Table 1** and discuss optimization details in the following sub-sections.

3.2.1. Update Coding Matrix X

When D is fixed, Eq. (7) becomes a coding problem of computing $X = [X_1, X_2, \dots, X_i, \dots, X_c]$. Each class of X_i can be successively computed class by class. When computing X_i , all X_j ($j \neq i$), are fixed. Thus, the optimization problem of Eq. (7) can be simplified as Eq. (8), and some details of optimization algorithm are presented in Appendix 1.

$$\begin{aligned} \min_X \sum_{i=1}^c & \left(\|Y_i - D_{class} X_i^{class}\|_F^2 + \|Y_i - (D_{class} S_i)(S_i^T X_i^{class})\|_F^2 \right) + \\ & \beta \sum_{i=1}^c \left(\|X_i^{class} - M_i\|_F^2 + \|M_i - M\|_F^2 + \|X_i^{class}\|_1 \right) + \\ & \sum_{i=1}^c \|Y_i - D_{c+1} X_i^{c+1}\|_F^2 + \gamma \sum_{i=1}^c \|X_i^{c+1}\|_* \end{aligned} \quad (8)$$

3.2.2. Update Dictionary D

When X is fixed, $\{D_j\}_{j=1, 2, \dots, c, c+1}$ can be updated one by one. When updating one sub-dictionary, the other sub-dictionaries are fixed. Therefore, the optimization problem of Eq. (7) can be simplified as Eq. (9), and some details of optimization algorithm are presented in Appendix 2.

$$\begin{aligned} \min_D \sum_{i=1}^c & \left(\|Y_i - D_{class} X_i^{class}\|_F^2 + \|Y_i - (D_{class} S_i)(S_i^T X_i^{class})\|_F^2 + \right. \\ & \left. \|Y_i - D_{c+1} X_i^{c+1}\|_F^2 \right) + \alpha \sum_{j=1}^{c+1} \left[\frac{n_j}{k_j(K - k_j)} \|D_j^T D_{-j}\|_F^2 + \right. \\ & \left. \frac{n_j}{k_j^2} \|D_j^T D_j - I_{k_j}\|_F^2 \right] \end{aligned} \quad (9)$$

3.3. Classification Scheme

The proposed CASDDL method especially emphasizes class discrimination of both dictionary atoms and coding vector, which not only contributes for learning class-oriented discriminative dictionary, but also results in discriminative coding vector, which has a finer adaptation for classification.

Different from traditional classification method that treat the coding vector just as input to sophisticated classifiers, we can directly makes full use of the discriminative capability of coding vector for a simple and efficient classification scheme, without adding any parameters to be learned.

For a test sample $\hat{y}$, we first simply use the obtained dictionary D to compute its coding vector $\hat{x} = [x_1; x_2; \dots; x_i; \dots; x_c]$, where, x_i is the coding sub-vector associated with class-specific sub-dictionary $\{D_i\}_{i=1, 2, \dots, c}$. Considering the discrimination of $\hat{x}$, if $\hat{y}$ is from class i , x_i will be large than other part. Therefore, the class of $\hat{y}$ is determined by $\arg \max_i \|x_i\|_2^2$.

4. Experimental Results

In this section, various issues involved in the proposed CASDDL method are discussed, including convergence analysis, parameters analysis, number of atoms in sub-dictionary, robustness to noise, visualization of coding vectors, and comparison of classification results. Two kinds of typical surface defects (Inclusion, Rolled-in Scale) and one kind of non-defective are selected in the following experiments. There are 300 grayscale images per class, whose image size is 200×200 pixels.⁴¹⁾ All the defect images are normalized and resized to 40×40 pixels. We randomly divide the whole dataset into a training dataset and a test dataset in the 1:1 ratio. The classification accuracy is defined as N_R/N , where, N_R is the number of test samples that are correctly classified, N is the total number of test samples.

4.1. Convergence Analysis

Although Eq. (7) is non-convex, the optimization algorithm actually adopts an alternatively updating fashion, and the convergence of each sub-problem can be guaranteed. On the one hand, for updating X with D fixed, the optimal solution is gained by TwIST and ALM. On the other hand, in the process of updating D with X fixed, each atom is optimally renewed for the sub-problem. As a consequence, the objective function is non-increasing during the whole process of alternatively updating X and D . In addition, we provide the empirical evidence to illustrate the good convergence behavior of our proposed CASDDL method in **Fig. 4**, which

Table 1. Algorithm of class-specific and shared discriminative dictionary learning.

Algorithm 1: Class-specific and Shared Discriminative Dictionary Learning

Input: Training samples $Y = \{Y_i\}_{i=1, 2, \dots, c}$; number of atoms k_j in class-specific dictionary $\{D_i\}_{i=1, 2, \dots, c}$; number of atoms k_{c+1} in shared sub-dictionary D_{c+1} ; parameters α, β , and γ .

Initialize: The class-specific sub-dictionary $\{D_j\}_{j=1, 2, \dots, c}$ is initialized by K -means in Y_i , the shared sub-dictionary D_{c+1} is initialized by K -means in Y .

While not converged **do**

step 1: Update X_i^{class} by Eq. (A1-12);

step 2: Update X_i^{c+1} by Table 2;

step 3: Update $\{D_i\}_{i=1, 2, \dots, c}$ by Eq. (A2-5);

step 4: Update D_{c+1} by Eq. (A2-7);

End While

Output: The optimal dictionary $D = \{D_j\}_{j=1, 2, \dots, c, c+1}$.

Table 2. Algorithm of solving Eq. (A1-15) via ALM.

Algorithm 2: Solving Eq. (A1-15) via ALM

Input: Training samples Y_i , shared sub-dictionary D_{c+1} , parameter $\gamma > 0$

Initialize: $J^{(0)} = (X_i^{c+1})^{(0)} = 0$, $H^{(0)} = 0$, $\mu^{(0)} = 0.1$, $\mu_{\max} = 10^5$, $\rho = 1.1$, $k = 0$, $k_{\max} = 10$

While not converged **do**

step 1: Update $J^{(k+1)}$ by Eq. (A1-18);

step 2: Update $(X_i^{c+1})^{(k+1)}$ by Eq. (A1-20);

step 3: Update $H^{(k+1)}$ by Eq. (A1-21);

step 4: Update $\mu^{(k+1)}$ by Eq. (A1-22);

step 5: Check the convergence condition $k < k_{\max}$;

step 6: Update k by $k = k + 1$;

End While

Output: The optimal solution X_i^{c+1} .

Fig. 4. Convergence curve of the proposed CASDDL method.

shows the curves of error values versus iteration numbers respectively. It can be seen that with the increase of iteration numbers, the curves of reconstructed error gradually decrease and eventually become flat and stable. Clearly, the proposed CASDDL method enjoys a good convergence performance and low complexity.

4.2. Parameters Analysis

The tuning three regularization parameters α , β , γ in Eq. (7) are chosen by 5-fold cross validation. α controls mutual incoherence between each sub-dictionary, β controls discrimination of coding vectors over all the class-specific sub-dictionaries, γ controls the low-rank ability of coding vector over the shared sub-dictionary. Following the work in,¹⁸⁾ we set $\alpha = 0.1$. Because there are many combinations of remaining two parameters β and γ , we search for their best values in a small set $\{0, 0.6, 0.8, 1.0\}$, $\{0, 0.5, 0.7, 0.9\}$, respectively.

Let k_c , k_s denotes number of atoms of class-specific sub-dictionary and shared sub-dictionary, respectively. We vary k_c from 10 to 45 with five interval, k_s from 2 to 30 with four interval, to study the parameters effect for classification performance. For each parameter combination, we compute the classification accuracy of all the sub-dictionary combinations in terms of mean value, and illustrate the classification accuracy in Table 3. The bottom row denotes the mean value of classification accuracy with one β corresponding to different γ , the right column denotes the mean value of classification accuracy with one γ corresponding to different β . As shown in Table 3, classification accuracy rises as the increase of β at first, but a further increase of β over a proper value will decrease the performance. The classification performance will degrade with a small value of β , which shows that the discriminative coding vector term is useful in learning class-oriented dictionary. Comparably, we can obtain a similar conclusion for γ , a larger value of γ will capture the inter-class similarity. In this scenario, the shared sub-dictionary is more readily to capture the commonality components. However, so large value of γ will decrease the representation ability of shared sub-dictionary; hence the classification performance will be degraded. For γ ,

Table 3. Classification accuracy with different parameters α , β , and γ of the proposed CASDDL method, fix $\alpha = 0.1$ to tune β and γ .

$\beta \backslash \gamma$	0	0.6	0.8	1	
0	0.8585	0.8918	0.8971	0.8936	0.8852
0.5	0.8882	0.8955	0.9028	0.8997	0.8965
0.7	0.8815	0.8932	0.8977	0.8973	0.8933
0.9	0.8825	0.8955	0.8990	0.8964	0.8924
	0.8976	0.8940	0.8991	0.8976	

we empirically observe that a value lying in the range $[0.5, 0.9]$ can always achieve an acceptable result. Furthermore, the classification accuracy with $\gamma = 0$ are lower than that with $\gamma = 0.5$, which illustrates the importance of the low-rank preserving term.

From Table 3, the highest performance is achieved when $\alpha = 0.1$, $\beta = 0.8$, and $\gamma = 0.5$. In the rest of experiments, we will use these parameter combinations. Besides, we observe that the classification performance is robust to different parameter values with the accuracies being greater than 89% in most cases. Therefore, it clearly demonstrates a positive performance for our proposed CASDDL method will indeed help to promote the classification performance, and exhibits robustness to parameter choices.

4.3. Number of Atoms in Sub-dictionary

According to k_c , k_s denotes number of atoms of class-specific sub-dictionary and shared sub-dictionary, we illustrate corresponding classification accuracy in Fig. 5 and Table 4. By checking the classification accuracy, we can get the following conclusions: firstly, we can observe that increasing k_c will lead to a higher classification performance. The possible reason is that more discriminative information can be captured by a larger class-specific sub-dictionary; secondly, when k_s is fixed, the classification accuracy is dropped as the increase of k_c . The possible reasons are that smaller shared sub-dictionary is enough to capture the shared features of defect images, and that larger shared sub-dictionary tends to absorb class-specific features into the shared sub-dictionary, causing many entries in feature vectors that carried discriminative information will be lost.

As shown in Fig. 5 and Table 4, we can see that our proposed CASDDL method always achieves higher classification accuracy despite different number of atoms. It indicates that CASDDL has a better ability to reconstruct defect images, even if learned dictionary has a small size. In fact, larger size of dictionary may have stronger representative ability of details and achieve better classification performance at the expense of increasing computational load. Therefore, we should make a tradeoff between classification performance and computational efficiency. When k_c is 30, the classification accuracy gain is merely promoted very little ($\sim 1\%$) as the increase of k_c . When number of atoms of shared sub-dictionary $k_s = 2$ and the size of class-specific sub-dictionary $k_c = 30$, CASDDL can still have higher classification accuracy 94.25%. Therefore, the proposed CASDDL method is able to efficiently reduce computational data redundancy and running time. In the rest of experiments, k_c , k_s are set to 30, 2, respectively.

4.4. Robustness to Noise

We evaluate the robustness of the proposed CASDDL method by corrupting original surface defect images with additive Gaussian noise in different signal to noise ratio

Fig. 5. Classification accuracy versus different number of atoms in class-specific sub-dictionary and shared sub-dictionary, k_c from 10 to 45 with five interval, k_s from 2 to 30 with four interval. (Online version in color.)

(SNR), including 24 dB, 20 dB, and 16 dB. According to **Table 5**, we can find that: (1) the classification result is decreased slower when the noise level is increased, *i.e.*, the corruptions clearly have the negative effects on the classification results; (2) CASDDL can achieve 81.37% classification accuracy even at 20 dB noise, which is considered as less sensitive to the noise.

4.5. Visualization of Coding Vectors

The proposed CASDDL method aims to get highly-discriminative coding vector, through the learned dictionary, to achieve the purpose of surface defect classification. As shown in **Fig. 6**, it illustrates the coding vectors of training and testing samples are approximately block-diagonal, which further shows the class-label discriminative information in coding vectors and can improve classification performance.

4.6. Comparison of Classification Results

We compare the proposed CASDDL method with other well-known dictionary learning methods, including SRC,¹⁰⁾

Table 5. Classification accuracy with different additive Gaussian noise.

SNR (dB)	Classification Accuracy (%)
16	75.09 ± 1.80
20	81.37 ± 1.39
24	88.39 ± 0.75

Table 4. Classification accuracy with different number of atoms k_c and k_s .

$k_s \backslash k_c$	10	15	20	25	30	35	40	45
2	0.8702	0.9158	0.9249	0.9364	0.9425	0.9381	0.9458	0.9473
6	0.8447	0.8987	0.9127	0.9251	0.9289	0.9342	0.9400	0.9427
10	0.8276	0.8984	0.9164	0.9204	0.9280	0.9320	0.9347	0.9422
14	0.8309	0.8764	0.9087	0.9149	0.9273	0.9287	0.9262	0.9300
18	0.8251	0.8624	0.8927	0.9093	0.9180	0.9171	0.9249	0.9300
22	0.8089	0.8602	0.8847	0.9071	0.9109	0.9256	0.9213	0.9198
26	0.8073	0.8658	0.8862	0.9067	0.9080	0.9207	0.9091	0.9216
30	0.8111	0.8536	0.8778	0.8971	0.9018	0.9187	0.9207	0.9187

Fig. 6. Visualization of coding vectors in training and testing process of the proposed CASDDL method. (Online version in color.)

Fig. 7. Visualization of multiple classification results. (Online version in color.)

Table 6. Classification accuracy comparison between the proposed method and other approaches.

Method	Classification Accuracy (%)
SRC	68.71±2.1
LCKSVD	76.98±1.2
DLSI	90.13±1.0
FDDL	68.98±1.7
COPAR	82.27±3.5
RDCDL	76.42±1.2
LRSDL	74.13±2.4
LRGPDLL	84.76±2.2
CASDDL	94.91±0.4

LCKSVD,¹⁶⁾ DLSI,¹⁷⁾ FDDL,²⁰⁾ COPAR,²⁵⁾ RDCDL,²⁶⁾ LRSDL,³⁰⁾ LRGPDLL.³¹⁾ To obtain a stable results, we repeat all the experiments 10 times, and report classification results in terms of mean±standard deviation.

As shown in Fig. 7 and Table 6, CASDDL achieves 94.91% classification accuracy, compared to 84.76% for LRGPDLL, 82.27% for COPAR, 90.13% for DLSI and 76.98% for LCKSVD. Compared to SRC, which is the baseline method in the experiment, CASDDL improves the classification accuracy with a margin of more than 26%. Among above approaches, DLSI performs the second best, which is inferior to ours by 4.12%. Besides, CASDDL outdoes LRGPDLL by a significant improvement of above 10%. Therefore, the proposed CASDDL method has better classification performance than these methods.

5. Conclusion

In this paper, we propose a CASDDL method to learn a discriminative dictionary that consists of several class-specific sub-dictionaries associated with corresponding classes and a shared sub-dictionary shared by all the classes, in which class-specific sub-dictionaries are responsible for exploiting class-specific information, and the shared sub-dictionary is used for capturing and separating the common information.

By imposing low-rank constraints on coding vectors over the shared sub-dictionary, it can effectively reduce redundancy in training samples. Moreover, the discriminative ability of learned dictionary is significantly improved by imposing mutual incoherence constraints on each sub-dictionary and Fisher-like discriminative constraints on coding vectors over all the class-specific sub-dictionaries, respectively. An efficient iterative algorithm is also presented to solve the optimization problem of CASDDL. Experimental results verify the effectiveness and robustness of CASDDL for classifying surface defects of steel sheet.

Acknowledgements

This research was funded by National Natural Science Foundation of China, under grant number 51805386, open fund of Key Laboratory of Image Processing and Intelligent Control (Huazhong University of Science and Technology), Ministry of Education, under grant number IPIC2019-03, open fund of Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering at Wuhan University of Science and Technology, under grant number 2018A04.

REFERENCES

- Q. W. Luo, X. X. Fang, L. Liu, C. H. Yang and Y. C. Sun: *IEEE Trans. Instrum. Meas.*, **69** (2020), 626. <https://doi.org/10.1109/TIM.2019.2963555>
- T. Czimmernann, G. Ciuti, M. Milazzo, M. Chiurazzi, S. Roccella, C. M. Oddo and P. Dario: *Sensors*, **20** (2020), 1459. <https://doi.org/10.3390/s20051459>
- J. W. Lu, V. E. Liong and J. Zhou: *IEEE Trans. Pattern Anal. Mach. Intell.*, **40** (2018), 1979. <https://doi.org/10.1109/TPAMI.2017.2737538>
- S. Li, M. Shao and Y. Fu: *IEEE Trans. Pattern Anal. Mach. Intell.*, **40** (2018), 2963. <https://doi.org/10.1109/TPAMI.2017.2764893>
- J. C. Yao, H. M. Yu and R. Hu: *Vis. Comput.*, **33** (2017), 179. <https://doi.org/10.1007/s00371-015-1171-2>
- J. Ling, Z. Z. Chen and F. Wu: *IEEE Trans. Circuits Syst. Video Technol.*, **30** (2020), 2155. <https://doi.org/10.1109/TCSVT.2019.2918852>
- J. Z. Wang, Q. Y. Li, J. R. Gan, H. M. Yu and X. Yang: *IEEE Trans. Ind. Inform.*, **16** (2020), 141. <https://doi.org/10.1109/TII.2019.2917522>
- X. H. Dong, C. J. Taylor and T. F. Coates: *IEEE Trans. Signal Process.*, **68** (2020), 6055. <https://doi.org/10.1109/TSP.2020.3031188>
- J. B. Zhang, H. Su, W. Zou, X. Y. Gong, Z. T. Zhang and F. Shen: *Pattern Recognit.*, **109** (2021), 107571. <https://doi.org/10.1016/j.patcog.2020.107571>
- J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma: *IEEE Trans. Pattern Anal. Mach. Intell.*, **31** (2009), 210. <https://doi.org/10.1109/TPAMI.2008.79>
- J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang and S. C. Yan: *Proc. IEEE*, **98** (2010), 1031. <https://doi.org/10.1109/JPROC.2010.2044470>
- R. Rubinstein, A. M. Bruckstein and M. Elad: *Proc. IEEE*, **98** (2010), 1045. <https://doi.org/10.1109/JPROC.2010.2040551>
- H. Cheng, Z. C. Liu, L. Yang and X. W. Chen: *Signal Process.*, **93** (2013), 1408. <https://doi.org/10.1016/j.sigpro.2012.09.011>
- M. Aharon, M. Elad and A. K. Bruckstein: *IEEE Trans. Signal Process.*, **54** (2006), 4311. <https://doi.org/10.1109/TSP.2006.881199>
- Q. Zhang and B. X. Li: 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, IEEE, Piscataway, NJ, (2010), 2691. <https://doi.org/10.1109/CVPR.2010.5539989>
- Z. L. Jiang, Z. Lin and L. S. Davis: *IEEE Trans. Pattern Anal. Mach. Intell.*, **35** (2013), 2651. <https://doi.org/10.1109/TPAMI.2013.88>
- I. Ramirez, P. Sprechmann and G. Sapiro: 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, IEEE, Piscataway, NJ, (2010), 3501. <https://doi.org/10.1109/CVPR.2010.5539964>
- S. H. Gao, I. W. H. Tsang and L. T. Chia: *IEEE Trans. Pattern Anal. Mach. Intell.*, **35** (2013), 92. <https://doi.org/10.1109/TPAMI.2012.63>
- Z. M. Li, Z. H. Lai, Y. Xu, J. Yang and D. Zhang: *IEEE Trans. Neural Netw. Learn. Syst.*, **28** (2017), 278. <https://doi.org/10.1109/TNNLS.2015.2508025>
- M. Yang, L. Zhang, X. C. Feng and D. Zhang: *Int. J. Comput. Vis.*, **109** (2014), 209. <https://doi.org/10.1007/s11263-014-0722-8>
- C. J. Zhang, C. Liang, L. Li, J. Liu, Q. M. Huang and Q. Tian:

- IEEE Trans. Neural Netw. Learn. Syst., **28** (2017), 1550. <https://doi.org/10.1109/TNNLS.2016.2545112>
- 22) W. H. Deng, J. N. Hu and J. Guo: *IEEE Trans. Pattern Anal. Mach. Intell.*, **40** (2018), 2513. <https://doi.org/10.1109/TPAMI.2017.2757923>
- 23) N. Zhou and J. P. Fan: *IEEE Trans. Pattern Anal. Mach. Intell.*, **36** (2014), 715. <https://doi.org/10.1109/TPAMI.2013.189>
- 24) S. H. Gao, I. W. H. Tsang and Y. Ma: *IEEE Trans. Image Process.*, **23** (2014), 623. <https://doi.org/10.1109/TIP.2013.2290593>
- 25) D. H. Wang and S. Kong: *Pattern Recognit.*, **47** (2014), 885. <https://doi.org/10.1016/j.patcog.2013.08.004>
- 26) G. J. Lin, M. Yang, J. Yang, L. L. Shen and W. C. Xie: *Pattern Recognit.*, **81** (2018), 341. <https://doi.org/10.1016/j.patcog.2018.03.021>
- 27) X. D. Jiang and J. Lai: *IEEE Trans. Pattern Anal. Mach. Intell.*, **37** (2015), 1067. <https://doi.org/10.1109/TPAMI.2014.2359453>
- 28) Y. Rong, S. W. Xiong and Y. S. Gao: *Pattern Recognit.*, **72** (2017), 419. <https://doi.org/10.1016/j.patcog.2017.06.038>
- 29) Z. D. Wen, B. Hou and L. C. Jiao: *IEEE Trans. Cybern.*, **47** (2017), 3758. <https://doi.org/10.1109/TCYB.2016.2581861>
- 30) T. H. Vu and V. Monga: *IEEE Trans. Image Process.*, **26** (2017), 5160. <https://doi.org/10.1109/TIP.2017.2729885>
- 31) H. S. Du, L. G. Ma, G. D. Li and S. Wang: *Knowl.-Based Syst.*, **187** (2020), 104823. <https://doi.org/10.1016/j.knosys.2019.06.031>
- 32) S. Y. Zhou, Y. P. Chen, D. L. Zhang, J. M. Xie and Y. F. Zhou: *ISIJ Int.*, **57** (2017), 123. <https://doi.org/10.2355/isijinternational>. ISIJINT-2016-478
- 33) Y. L. Wang, H. B. Xia, X. F. Yuan, L. Li and B. Sun: *Multimed. Tools Appl.*, **77** (2018), 16741. <https://doi.org/10.1007/s11042-017-5238-0>
- 34) R. Borwankar and R. Ludwig: *IEEE Trans. Instrum. Meas.*, **67** (2018), 690. <https://doi.org/10.1109/TIM.2017.2783098>
- 35) Q. W. Luo, Y. C. Sun, P. C. Li, O. Simpson, L. Tian and Y. G. He: *IEEE Trans. Instrum. Meas.*, **68** (2019), 667. <https://doi.org/10.1109/TIM.2018.2852918>
- 36) M. W. Ashour, F. Khalid, A. A. Halin, L. N. Abdullah and S. H. Darwish: *Arab. J. Sci. Eng.*, **44** (2019), 2925. <https://doi.org/10.1007/s13369-018-3329-5>
- 37) H. Foroughi, N. Ray and H. Zhang: *IEEE Trans. Image Process.*, **27** (2018), 806. <https://doi.org/10.1109/TIP.2017.2766446>
- 38) A. Abdi, M. Rahmati and M. M. Ebadzadeh: *Pattern Recognit.*, **110** (2021), 107634. <https://doi.org/10.1016/j.patcog.2020.107634>
- 39) A. M. Bruckstein, D. L. Donoho and M. Elad: *SIAM Rev.*, **51** (2009), 34. <https://doi.org/10.1137/060657704>
- 40) S. Ravishanker, J. C. Ye and J. A. Fessler: *Proc. IEEE*, **108** (2019), 86. <https://doi.org/10.1109/JPROC.2019.2936204>
- 41) Y. He, K. C. Song, Q. G. Meng and Y. H. Yan: *IEEE Trans. Instrum. Meas.*, **69** (2020), 1493. <https://doi.org/10.1109/TIM.2019.2915404>
- 42) J. M. Bioucas-Dias and M. A. T. Figueiredo: *IEEE Trans. Image Process.*, **16** (2007), 2992. <https://doi.org/10.1109/TIP.2007.909319>
- 43) I. Ramirez, F. Lecumberry and G. Sapiro: *Sparse Modeling with Universal Priors and Learned Incoherent Dictionaries*, University of Minnesota, Minneapolis, (2009), IMA Preprint Series # 2279.

Appendices

Appendix 1. Update Coding Matrix X

A. Update X_i^{class}

With fixed D and X_i^{c+1} , Eq. (8) can be rewritten as follows:

$$\min_{X_i^{class}} \|Y_i - D_{class} X_i^{class}\|_F^2 + \|Y_i - (D_{class} S_i)(S_i^T X_i^{class})\|_F^2 + \beta \left(\|X_i^{class} - M_i\|_F^2 + \sum_{i=1}^c \|M_i - M\|_F^2 \right) + \beta \|X_i^{class}\|_1 \quad \text{..... (A1-1)}$$

It can be rewritten as follows:

$$\min_{X_i^{class}} F(X_i^{class}) + 2\tau \|X_i^{class}\|_1 \quad \text{..... (A1-2)}$$

where, $F(X_i^{class}) = \|Y_i - D_{class} X_i^{class}\|_F^2 + \|Y_i - (D_{class} S_i)(S_i^T X_i^{class})\|_F^2 + \beta \left(\|X_i^{class} - M_i\|_F^2 + \sum_{i=1}^c \|M_i - M\|_F^2 \right)$, $\tau = \frac{\beta}{2}$.

Because $F(X_i^{class})$ is convex and differentiable to X_i^{class} , a two-step iterative shrinkage/thresholding (TwIST) algorithm⁴²⁾ can be employed to solve Eq. (A1-2).

$$\text{Denote } E_i^j = \begin{pmatrix} 1 \cdots 1 \\ \vdots \\ 1 \cdots 1 \end{pmatrix}_{n_i \times n_j}, I_{n_i \times n_i} = \begin{pmatrix} 1 \cdots 0 \\ \vdots \\ 0 \cdots 1 \end{pmatrix}_{n_i \times n_i}, O_i = I_{n_i \times n_i} - \frac{E_i^i}{n_i},$$

then, we have

$$\begin{aligned} \|X_i^{class} - M_i\|_F^2 &= \left\| X_i^{class} \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}_{n_i \times n_i} - X_i^{class} \begin{pmatrix} \frac{1}{n_i} & \cdots & \frac{1}{n_i} \\ \vdots & \ddots & \vdots \\ \frac{1}{n_i} & \cdots & \frac{1}{n_i} \end{pmatrix}_{n_i \times n_i} \right\|_F^2 \\ &= \left\| X_i^{class} I_{n_i \times n_i} - X_i^{class} \frac{E_i^i}{n_i} \right\|_F^2 \\ &= \|X_i^{class} O_i\|_F^2 \quad \text{..... (A1-3)} \end{aligned}$$

$$\text{Denote } Q_i^j = \frac{E_i^j}{N}, P_i = \frac{E_i^i}{n_i} - \frac{E_i^i}{N} = \frac{E_i^i}{n_i} - Q_i^i, R = \sum_{j=1}^c X_j^{class} Q_j^j,$$

then, we have

$$\begin{aligned} \sum_{i=1}^c \|M_i - M\|_F^2 &= \left\| X_i^{class} \frac{E_i^i}{n_i} - \left(X_i^{class} \frac{E_i^i}{N} + \sum_{j=1}^c X_j^{class} \frac{E_j^i}{N} \right) \right\|_F^2 \\ &= \left\| X_i^{class} \left(\frac{E_i^i}{n_i} - Q_i^i \right) - \sum_{j=1}^c X_j^{class} Q_j^i \right\|_F^2 \\ &= \left\| X_i^{class} P_i - \sum_{j=1}^c X_j^{class} Q_j^i \right\|_F^2 \quad \text{..... (A1-4)} \end{aligned}$$

$$\text{For the } i\text{-th class, we have } \left\| X_i^{class} P_i - \sum_{j=1}^c X_j^{class} Q_j^i \right\|_F^2, \text{ for}$$

$$\text{the non } i\text{-th class, we have } \sum_{j=1}^c \left\| X_j^{class} P_j - \sum_{l=1}^c X_l^{class} Q_l^j \right\|_F^2.$$

Choose the j -th part of the i -th class, we have

$$\begin{aligned} \left\| X_j^{class} P_j - \sum_{l=1}^c X_l^{class} Q_l^j \right\|_F^2 &= \left\| X_j^{class} \frac{E_j^j}{n_j} - \left(X_j^{class} Q_j^j + \sum_{l=1}^c X_l^{class} Q_l^j \right) \right\|_F^2 \quad \text{.... (A1-5)} \end{aligned}$$

Separate X_i^{class} , we have

$$\begin{aligned} \left\| X_j^{class} \frac{E_j^j}{n_j} - \left(X_i^{class} Q_i^j + \sum_{l=1, l \neq i}^c X_l^{class} Q_l^j \right) \right\|_F^2 &= \left\| X_j^{class} \frac{E_j^j}{n_j} - \sum_{l=1, l \neq i}^c X_l^{class} Q_l^j - X_i^{class} Q_i^j \right\|_F^2 \quad \text{... (A1-6)} \\ &= \|T_j - X_i^{class} Q_i^j\|_F^2 \end{aligned}$$

Expand the anyone non i -th class, $\sum_{j \neq i}^c \|T_j - X_i^{class} Q_j^i\|_F^2$, then, we have

$$\begin{aligned} & \|X_i^{class} - M_i\|_F^2 + \sum_{i=1}^c \|M_i - M\|_F^2 \\ &= \|X_i^{class} O_i\|_F^2 - \|X_i^{class} P_i - R\|_F^2 - \sum_{j \neq i}^c \|T_j - X_i^{class} Q_j^i\|_F^2 \\ & \dots\dots\dots (A1-7) \end{aligned}$$

The first derivative of $F(X_i^{class})$ with respect to X_i^{class} is calculated, we have

$$\frac{\partial \|X_i^{class} O_i\|_F^2}{\partial X_i^{class}} = 2X_i^{class} O_i O_i^T \dots\dots\dots (A1-8)$$

$$\frac{\partial \|X_i^{class} P_i - R\|_F^2}{\partial X_i^{class}} = 2(X_i^{class} P_i P_i^T - R P_i^T) \dots\dots (A1-9)$$

$$\frac{\partial \sum_{j \neq i}^c \|T_j - X_i^{class} Q_j^i\|_F^2}{\partial X_i^{class}} = 2 \sum_{j \neq i}^c [X_i^{class} Q_j^i (Q_j^i)^T - T_j (Q_j^i)^T] \dots\dots\dots (A1-10)$$

Then we have

$$\begin{aligned} \nabla_{X_i^{class}} F(X_i^{class}) &= 2D_{class}^T (D_{class} X_i^{class} - Y_i) + \\ & 2S_i S_i^T D_{class}^T (D_{class} S_i S_i^T X_i^{class} - Y_i) + 2\beta \left\{ X_i^{class} O_i O_i^T + \right. \\ & \left. X_i^{class} P_i P_i^T - R P_i^T + \sum_{j \neq i}^c [X_i^{class} Q_j^i (Q_j^i)^T - T_j (Q_j^i)^T] \right\} \\ & \dots\dots\dots (A1-11) \end{aligned}$$

$$\text{where, } E_i^j = \begin{pmatrix} 1 \dots 1 \\ \vdots \vdots \vdots \\ 1 \dots 1 \end{pmatrix}_{n_i \times n_j}, O_i = I_{n_i \times n_i} - \frac{E_i^i}{n_i}, P_i = \frac{E_i^i}{n_i} - \frac{E_i^i}{N} = \frac{E_i^i}{n_i} - Q_i^i,$$

$$R = \sum_{j \neq i}^c X_j Q_j^i, T_j = X_j \frac{E_j^i}{n_j} - \sum_{l \neq i}^c X_l Q_l^j.$$

According to TwIST algorithm, we have

$$\begin{aligned} (X_i^{class})^{(t+1)} &= (1-\xi)(X_i^{class})^{(t-1)} + (\xi-\nu)(X_i^{class})^{(t)} \\ &+ \nu \Psi_{\frac{\tau}{\sigma}} \left[(X_i^{class})^{(t)} - \frac{1}{2\sigma} \nabla_{X_i^{class}} F(X_i^{class}) \right] \\ & \dots\dots\dots (A1-12) \end{aligned}$$

where, $\Psi_{\frac{\tau}{\sigma}}(M)$ denotes soft-thresholding shrinkage opera-

$$\text{tor, } \Psi_{\frac{\tau}{\sigma}}(M) = \begin{cases} m_{ij} - \frac{\tau}{\sigma} & m_{ij} \geq \frac{\tau}{\sigma} \\ 0 & m_{ij} < \frac{\tau}{\sigma} \end{cases}, m_{ij} \text{ stands for the } (i,j)\text{-th}$$

element of matrix M ; $(X_i^{class})^{(t-1)}$ is the previous value of

X_i^{class} , $(X_i^{class})^{(t)}$ is the current value of X_i^{class} , $(X_i^{class})^{(t+1)}$ is the next value of X_i^{class} ; $\xi > 0$, $\nu > 0$, $\sigma > 0$.

B. Update X_i^{c+1}

With fixed D and X_i^{class} , Eq. (8) is further reduced to:

$$\min_{X_i^{c+1}} \|Y_i - D_{c+1} X_i^{c+1}\|_F^2 + \gamma \|X_i^{c+1}\|_* \dots\dots\dots (A1-13)$$

According to inexact augmented Lagrangian multipliers (ALM) algorithm, introducing the auxiliary variable $J = X_i^{c+1}$, Eq. (A1-13) is turned into the form of an augmented Lagrangian function, which is defined as follows:

$$\begin{aligned} \mathcal{O}(X_i^{c+1}, J, H, \mu) &= \|Y_i - D_{c+1} J\|_F^2 + \gamma \|X_i^{c+1}\|_* \\ &+ \langle H, J - X_i^{c+1} \rangle + \frac{\mu}{2} \|J - X_i^{c+1}\|_F^2 \dots\dots (A1-14) \end{aligned}$$

where, $\langle \cdot, \cdot \rangle$ means the inner product operator for two matrices; $\|\cdot\|_F$ denotes the Frobenius norm, which is defined as the sum of squares of each element of matrix; H is a Lagrange multiplier; $\mu > 0$ is a penalty parameter.

Furthermore, we have

$$\begin{aligned} \mathcal{O}(X_i^{c+1}, J, H, \mu) &= \frac{1}{2} \left\| J + \frac{H}{\mu} - X_i^{c+1} \right\|_F^2 \\ &+ \frac{1}{\mu} \|Y_i - D_{c+1} J\|_F^2 + \frac{\gamma}{\mu} \|X_i^{c+1}\|_* \dots\dots\dots (A1-15) \end{aligned}$$

The detailed procedure of solving Eq. (A1-15) is presented in **Table 2**.

① Update J when fix the others

$$\frac{1}{2} \left\| J + \frac{H}{\mu} - X_i^{c+1} \right\|_F^2 + \frac{1}{\mu} \|Y_i - D_{c+1} J\|_F^2 \dots\dots (A1-16)$$

Differentiating it with respect to J , and let it to be zero:

$$J + \frac{H}{\mu} - X_i^{c+1} + \frac{2}{\mu} D_{c+1}^T (Y_i - D_{c+1} J) = 0 \dots\dots (A1-17)$$

we have

$$J^{(k+1)} = \left(I - \frac{2}{\mu^{(k)}} D_{c+1}^T D_{c+1} \right)^{-1} \left[(X_i^{c+1})^{(k)} - \frac{2}{\mu^{(k)}} D_{c+1}^T Y_i - \frac{H^{(k)}}{\mu^{(k)}} \right] \dots\dots\dots (A1-18)$$

② Update X_i^{c+1} when fix the others

$$\frac{1}{2} \left\| J + \frac{H}{\mu} - X_i^{c+1} \right\|_F^2 + \frac{\gamma}{\mu} \|X_i^{c+1}\|_* \dots\dots\dots (A1-19)$$

we have

$$(X_i^{c+1})^{(k+1)} = U \Psi_{\frac{\gamma}{\mu}} \{\Sigma\} V^T \dots\dots\dots (A1-20)$$

where, $(U, \Sigma, V) = \text{svd} \left(J^{(k+1)} + \frac{H^{(k)}}{\mu^{(k)}} \right)$, $\text{svd}(\cdot)$ denotes singular

value decomposition (SVD) operation, $\Sigma = \text{diag}(\{\sigma_i\}_{1 \leq i \leq r})$ is the diagonal matrix with $\sigma_1, \sigma_2, \dots, \sigma_r$ on the diagonal and zeros elsewhere, σ_i is the i -th singular value of $J^{(k+1)} + \frac{H^{(k)}}{\mu^{(k)}}$, $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{N \times r}$ are left, right singu-

lar matrices, respectively.

③ Update H

$$H^{(k+1)} = H^{(k)} + \mu^{(k)} \left(J^{(k+1)} - (X_i^{c+1})^{(k+1)} \right) \dots (A1-21)$$

④ Update μ

$$\mu^{(k+1)} = \min(\rho\mu^{(k)}, \mu_{\max}) \dots (A1-22)$$

where, $\rho = 1.1$, $\mu_{\max} = 10^5$.

Appendix 2. Update Dictionary D

A. Update $\{D_i\}_{i=1,2,\dots,c}$

With fixed X and other sub-dictionaries, Eq. (9) can be rewritten as follows:

$$\begin{aligned} \min_{D_i} \sum_{i=1}^c & \left(\|Y_i - D_{class} X_i^{class}\|_F^2 + \|Y_i - (D_{class} S_i)(S_i^T X_i^{class})\|_F^2 \right) + \\ & \alpha \frac{n_i}{k_i(K-k_i)} \|D_i^T D_{-i}\|_F^2 + \alpha \frac{n_i}{k_i^2} \|D_i^T D_i - I_{k_i}\|_F^2 \\ & \dots (A2-1) \end{aligned}$$

We optimize $\{D_i\}_{i=1,2,\dots,c}$ class-by-class and meanwhile, make all other D_j ($j \neq i$) fixed. As a result, we have

$$\begin{aligned} \min_{D_i} & \|Y - D_{class}(X_1^{class}, \dots, X_c^{class})\|_F^2 + \\ & \|Y - D_{class}(S_1 S_1^T X_1^{class}, \dots, S_c S_c^T X_c^{class})\|_F^2 + \\ & \alpha \frac{n_i}{k_i(K-k_i)} \|D_i^T D_{-i}\|_F^2 + \alpha \frac{n_i}{k_i^2} \|D_i^T D_i - I_{k_i}\|_F^2 \\ & \dots (A2-2) \end{aligned}$$

Denote $A = [Y, Y] \in \mathbb{R}^{d \times 2N}$, $B = [X_1^{class}, \dots, X_c^{class}, S_1 S_1^T X_1^{class}, \dots, S_c S_c^T X_c^{class}] \in \mathbb{R}^{(K-k_{c+1}) \times 2N}$, we have

$$\begin{aligned} \min_{D_i} & \|A - D_{class} B\|_F^2 + \alpha \frac{n_i}{k_i(K-k_i)} \|D_i^T D_{-i}\|_F^2 + \\ & \alpha \frac{n_i}{k_i^2} \|D_i^T D_i - I_{k_i}\|_F^2 \\ & \dots (A2-3) \end{aligned}$$

Therefore

$$\begin{aligned} \min_{D_i} & \left\| A - \sum_{j=1, j \neq i}^c D_j B^j - D_i B^i \right\|_F^2 + \alpha \frac{n_i}{k_i(K-k_i)} \\ & \|D_i^T D_{-i}\|_F^2 + \alpha \frac{n_i}{k_i^2} \|D_i^T D_i - I_{k_i}\|_F^2 \\ & \dots (A2-4) \end{aligned}$$

Denote $\tilde{A} = A - \sum_{j=1, j \neq i}^c D_j B^j \in \mathbb{R}^{d \times 2N}$, we have

$$\begin{aligned} \min_{D_i} & \|\tilde{A} - D_i B^i\|_F^2 + \alpha \frac{n_i}{k_i(K-k_i)} \|D_i^T D_{-i}\|_F^2 + \\ & \alpha \frac{n_i}{k_i^2} \|D_i^T D_i - I_{k_i}\|_F^2 \\ & \dots (A2-5) \end{aligned}$$

where, $B^i \in \mathbb{R}^{k_i \times 2N}$.

Equation (A2-5) can be solved by method of optimal coherence-constrained directions (MOCOD) algorithm.⁴³⁾

B. Update D_{c+1}

With fixed X and all the class-specific sub-dictionaries, by ignoring irrelevant terms, Eq. (9) can be rewritten as follows:

$$\begin{aligned} \min_{D_{c+1}} & \sum_{i=1}^c \|Y_i - D_{c+1} X_i^{c+1}\|_F^2 + \alpha \frac{n_{c+1}}{k_{c+1}(K-k_{c+1})} \|D_{c+1}^T D_{-(c+1)}\|_F^2 + \\ & \alpha \frac{n_{c+1}}{k_{c+1}^2} \|D_{c+1}^T D_{c+1} - I_{k_{c+1}}\|_F^2 \\ & \dots (A2-6) \end{aligned}$$

Denote $X^{c+1} = [X_1^{c+1}, X_2^{c+1}, \dots, X_c^{c+1}] \in \mathbb{R}^{k_{c+1} \times N}$ is the coding matrix of Y over shared sub-dictionary D_{c+1} , then we have

$$\begin{aligned} \min_{D_{c+1}} & \|Y - D_{c+1} X^{c+1}\|_F^2 + \alpha \frac{n_{c+1}}{k_{c+1}(K-k_{c+1})} \|D_{c+1}^T D_{class}\|_F^2 + \\ & \alpha \frac{n_{c+1}}{k_{c+1}^2} \|D_{c+1}^T D_{c+1} - I_{k_{c+1}}\|_F^2 \\ & \dots (A2-7) \end{aligned}$$

Similar to Eqs. (A2-5), (A2-7) can be solved by MOCOD algorithm.