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

Shiyang ZHOU, Youping CHEN, Dailin ZHANG,* Jingming XIE and Yunfei ZHOU

School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, 430074 China.

(Received on August 1, 2016; accepted on September 20, 2016; J-STAGE Advance published date: November 26, 2016)

An approach to a class-specific and shared dictionary learning (CSDSL) for sparse representation is proposed to classify surface defects of steel sheet. The proposed CSDSL algorithm is modelled as a unified objective function, covering reconstructive error, sparse and discriminative promotion constraints. With the high-quality dictionary, the compact, reconstructive and discriminative feature representation of an image can be extracted. Then the classification can be efficiently performed by discriminative information obtained from the reconstructive error or the sparse vector. Based on a dataset of surface images captured from a practical steel production line, the CSDSL algorithm is carried out to verify its effectiveness. Experimental results indicate that the CSDSL algorithm is more effective in classifying surface defects of steel sheet than other algorithms.

KEY WORDS: surface defects of steel sheet; image classification; sparse representation; discriminative dictionary learning.

1. Introduction

Machine vision-based classification of surface defects of steel sheet has attracted much attention due to the potential value and challenges for some practical industrial applications.^{1–8)} As a classical pattern recognition problem, feature extraction (learning) and designing of classifier construct the core. Most of existing works have focused on these two aspects to enhance the classification performance over the past few years.^{9–20)} A set of features are studied, including grayscale features,^{9,10)} geometric features,^{9,10)} shape features,^{9–11)} texture features,^{10,12)} Gabor filter output features,¹³⁾ Fourier spectral,¹⁴⁾ wavelet transform,^{14–17)} local binary pattern (LBP),^{18,19)} shearlet transform,²⁰⁾ and the classifier of artificial neural network (ANN)^{12,13,15)} and support vector machine (SVM)^{9,10,14,17–20)} is discussed. However, these approaches, which treat feature extraction and classifier training as two separating steps, are difficult to control the interaction of two steps. In fact, sparse representation and dictionary learning based classification methods, with no explicit stage of feature extraction, may be more effective in classification of surface defects of steel sheet. The sparse representation and dictionary learning, which can be formulated as follows, has also been successfully applied in many machine vision applications,²¹⁾ such as face recognition.²²⁾

$$\min_{\mathbf{x}} \|\mathbf{y} - D\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1 \quad \dots \quad (1)$$

where $\mathbf{y} \in \mathbb{R}^d$ is a given feature vector, $D \in \mathbb{R}^{d \times K}$ is a dictionary, $\|\mathbf{x}\|_1$ is l_1 -norm of a coding vector $\mathbf{x}$, l_2 -norm of an atoms

$\mathbf{d}_i$ is less than one, λ is a positive parameter that balances the tradeoff between reconstructive error and sparsity. The optimization of $\mathbf{x}$ and D is carried out by an iterative method composing two steps: (a) fixing D to update $\mathbf{x}$; (b) fixing $\mathbf{x}$ to update D . Equation (1) can be solved efficiently by many kinds of algorithms,^{36–40)} such as orthogonal matching pursuit (OMP).⁴¹⁾

A choice of dictionary D plays a key role to obtain a sparse, high-fidelity and discriminative feature representation of an image. Many state-of-the-art results have shown that a good dictionary is learned from the training dataset itself.^{23,24)} For example, given a set of feature vectors $\{\mathbf{y}_i\}_{i=1,2,\dots,N}$ in $\mathbb{R}^d$, the goal of dictionary learning is to learn a dictionary $D = \{\mathbf{d}_i\}_{i=1,2,\dots,K}$ in $\mathbb{R}^{d \times K}$ that each $\mathbf{y}_i$ can be compactly represented as a sparse linear combination of a few atoms from D , meanwhile keeping the reconstruction error $\|\mathbf{y}_i - D\mathbf{x}_i\|_2^2$ as small as possible. Many existing dictionary learning approaches can be divided into two categories: unsupervised dictionary learning and supervised dictionary learning. For the unsupervised dictionary learning methods, such as method of optimal directions (MOD)²⁵⁾ and K-SVD,²⁶⁾ the learned dictionary may not be best for classification because the class information of a training dataset isn't utilized. In recent years, a large number of researches about supervised or discriminative dictionary learning have been proposed, such as sparse representation based classification (SRC).²²⁾ A reconstructive and discriminative dictionary, which has a finer adaptation for classification, can be learned. The state-of-the-art classification performance^{27–35)} suggests that sparse representation and dictionary learning is a promising method for classifying surface defects of steel sheet.

The surface defects of steel sheet as shown in **Fig. 1**

* Corresponding author: E-mail: mnizhang@mail.hust.edu.cn
DOI: <http://dx.doi.org/10.2355/isijinternational.ISIJINT-2016-478>

Fig. 1. Examples of practical surface defects of steel sheet: (a) Folding; (b) Line; (c) Patch; (d) Scratch; (e) Non-defective.

can be regarded as the local anomaly against the relatively homogeneous background. The defect area of an image (discriminative pattern) occupies a small portion of an image. The background texture of an image (shared pattern), which is useful for reconstruction rather than discrimination, is nearly the same. Because the class-specific information of defects is only a small proportion in the total information of an image, the feature representation of different defect images aren't discriminative. If a traditional discriminative dictionary learning method is adopted, most of atoms of the learned dictionary are used to represent the background information that contributes nothing to classify different defects, and only a small part of atoms are used to represent the class-specific information of defects. Therefore, the discriminative ability of a coding vector between different defect images will reduce, and the poor classification performance may be achieved.

Based on the above analysis, a class-specific and shared dictionary learning (CSDSL) algorithm, in which some class-specific sub-dictionaries and only one shared sub-dictionary can be learned simultaneously, is proposed. Each class-specific sub-dictionary encodes the exclusive pattern for corresponding class, and the shared sub-dictionary encodes the explicit and hidden common pattern that shared by all the classes. With these sub-dictionaries, the exclusive pattern and the shared pattern of an image can be explicitly separated. Because the atoms of these sub-dictionaries can capture the inherent structure of a defect image, the effective feature representation of the image for classification can be obtained.

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, sparse representation and dictionary learning based classification. In Section 3, the proposed CSDSL algorithm, including formulation and optimization, is introduced. In Section 4, experimental results are presented. Finally, conclusions are given.

2. Related Work

In the past decades, a lot of works about classification of surface defects of steel sheet have been studied by some researchers. Caleb *et al.* presented a classification method of surface defects of hot rolled steel based on the classifier of multi-layer perceptron (MLP) and self-organising maps network (SOM).¹²⁾ Medina *et al.* introduced an approach to classify six different kinds of defects of flat steel coils by combining an ANN, a k -nearest neighbor (KNN) and a naive Bayesian.¹³⁾ Ghorai *et al.* exploited an approach of automatic defect detection of hot rolled flat steel based on three-level Haar feature set.¹⁷⁾ Hu *et al.* established a SVM model to classify five types of surface defects of

stripe steel.⁹⁾ Furthermore, they optimized feature vector of image and kernel function of SVM by hybrid chromosome genetic algorithm and had better classification performance.¹⁰⁾ Moreover, Chu *et al.* constructed an invariant feature extraction approach based on smoothed LBP.¹⁹⁾ However, these methods largely depend on an appropriate set of hand-crafted features, which need the special design for the special task.

Recently, some sparse representation and dictionary learning based classification methods, which are mainly divided into two categories, have also been paid more attention. One category aims to improve the discriminative capability of a dictionary.^{27,28)} Ramirez *et al.* introduced a structured incoherence term for promoting the independence of sub-dictionary associated with different classes.²⁷⁾ Gao *et al.* learned both some category-specific sub-dictionaries and a shared dictionary by imposing cross-incoherence constraint between different sub-dictionaries and self-incoherence constraint in each sub-dictionary.²⁸⁾ For these methods, the reconstructive error associated with different classes can be used for classification, but a coding vector, which isn't discriminative, isn't suitable for classification. Another category aims to improve the discriminative capability of a coding vector based on a discriminative dictionary.^{29–32)} Huang *et al.* presented a method that quantified the discriminative ability of a coding vector by Fisher's discrimination criterion.²⁹⁾ Zhang *et al.* introduced a classification error term of a linear classifier.³⁰⁾ Furthermore, Jiang *et al.* exploited a label consistency regularization term.³¹⁾ These methods have learned a shared dictionary by all the classes and treat a coding vector as a new feature representation of an original image. Although the coding vector is discriminative, the classification based on the class-specific reconstructive error doesn't work. The main difference between these two kinds of method is the discriminative capability of a dictionary or a coding vector. If the dictionary discrimination and the coding vector discrimination can be combined, better classification performance may be gained.

3. Class-specific and Shared Dictionary Learning (CSDSL)

Given a training dataset of c classes, the CSDSL algorithm constructs c discriminative class-specific sub-dictionaries for each class and a shared sub-dictionary for all classes, respectively. Some structural incoherence terms are added in the objective function to promote the incoherence for the sub-dictionary and the coding vector. The coding vector over the learned dictionary, which can be directly treated as a new feature representation of an image, contains crucial information for reconstruction and discriminative information for classification. The coding vector of a defect image of the same class is similar, while the coding vector of different class is dissimilar. Therefore, both the reconstructive error associated with different classes and the coding vector can be used for classification. In addition, we can amplify the discriminative part and suppress the shared part for feature representation of a defect image by controlling the number of atoms of sub-dictionary. The illustration of the CSDSL algorithm is shown in **Fig. 2**.

Fig. 2. An illustration of the CDSDL algorithm. Each image can be sparsely represented by a few atoms from corresponding class-specific sub-dictionary D_i (purple) and shared sub-dictionary D_{c+1} (red).

3.1. Formulation of CDSDL

Let $Y = [Y_1, \dots, Y_i, \dots, Y_c] \in \mathbb{R}^{d \times N}$ denote a training dataset of c classes, $Y_i \in \mathbb{R}^{d \times n_i}$ is a matrix constructed by stacking n_i samples of length d , $\sum_{i=1}^c n_i = N$, where d is the dimension of a sample, n_i is the number of sample of the i -th class, N is the total number of the whole training dataset. Let $D = [D_1, \dots, D_i, \dots, D_c, D_{c+1}] \in \mathbb{R}^{d \times K}$ denote the learned dictionary of K atoms, $D_i \in \mathbb{R}^{d \times k_i}$, $i = 1, 2, \dots, c$, is the class-specific sub-dictionary that trained from a corresponding training dataset Y_i , $D_{c+1} \in \mathbb{R}^{d \times k_{c+1}}$ is the shared sub-dictionary that trained from the whole training dataset Y , where $K = \sum_{i=1}^{c+1} k_i$, k_i is the number of atoms of the i -th sub-dictionary. Therefore, the objective function of the CDSDL algorithm is shown as follows.

$$J = \min_{D, X} F(D, X; Y) + \lambda_1 \|X\|_1 + \lambda_2 G(X) + H(D) \quad \dots (2)$$

$$\text{s.t. } \|D_i(:, j)\|_2 = 1, \forall i, j$$

where

$$F(D, X; Y) = \sum_{i=1}^c \left(\|Y_i - DX_i\|_F^2 + \|Y_i - (D\tilde{Q}_i)((\tilde{Q}_i)^T X_i)\|_F^2 + \lambda_1 \|X_i\|_1 \right) \quad \dots (3)$$

$$G(X) = \left[\sum_{i=1}^c \|X_i - M_i\|_F^2 - \sum_{i=1}^c \|M_i - M\|_F^2 + \sum_{i=1}^c \|X_i\|_F^2 \right] \quad \dots (4)$$

$$H(D) = \sum_{i=1}^{c+1} \left(w_1 \frac{n_i}{k_i^2} \|D_i^T D_i - I_{k_i}\|_F^2 + w_2 \frac{n_i}{k_i(K-k_i)} \|D_i^T (DQ_{-i})\|_F^2 \right) \quad \dots (5)$$

$F(D, X; Y)$ is a reconstructive error term, $G(X)$ is a Fisher discriminative term imposed on the coding vector matrix X , $H(D)$ is a structural incoherence term imposed on the dictionary D , each atom d_j of D is constrained to unit l_2 -norm to avoid trivial solution X .

$\|\cdot\|_F$ denotes Frobenius-norm, $\|X\|_1$ is a sparsity regularizer, I_{k_i} is an identity matrix, λ_1 , λ_2 , w_1 and w_2 are the scalar

weight parameters, $n_{c+1} = N$.

$X_i = [X_i^1; \dots; X_i^i; \dots; X_i^c; X_i^{c+1}] \in \mathbb{R}^{K \times n_i}$ is a coding vector matrix corresponding to a training dataset Y_i , each column x_i^j is the coding vector corresponding to the j -th training sample of the i -th class, X_i^i is the coding vector matrix of Y_i corresponding to a sub-dictionary D_i .

$Q_i \in \mathbb{R}^{K \times k_i}$ is a selective operator, each column has only one nonzero element 1 that the row index equals the column index of an atom d_j , $\tilde{Q}_i = [Q_i, Q_{c+1}] \in \mathbb{R}^{K \times (k_i + k_{c+1})}$, $Q_{-i} = [Q_i, \dots, Q_{i-1}, Q_{i+1}, \dots, Q_c, Q_{c+1}] \in \mathbb{R}^{K \times (K - k_i)}$.

3.2. Analysis of CDSDL

The term $\|Y_i - DX_i\|_F^2$ indicates that the dictionary D should represent Y_i well.

Because D_i is a class-specific sub-dictionary and D_{c+1} is a shared sub-dictionary, the term $\|Y_i - (D\tilde{Q}_i)((\tilde{Q}_i)^T X_i)\|_F^2$ indicates that Y_i should be well represented by $[D_i, D_{c+1}]$, but not by $[D_j, D_{c+1}]$, $j \neq i$, which forces the samples from the same class to have similar feature representation.

The term $G(X)$ directly operates the coding vector and promotes them to more discriminative between different classes. Based on Fisher's linear discriminant,⁴³⁾ which maximizes the ratio of between-class scatter matrix to within-class scatter matrix, we can jointly minimize the within-class scatter matrix $S_W(X)$ and maximum the between-class scatter matrix $S_B(X)$. $S_W(X)$ and $S_B(X)$ are defined as follows.

$$S_W(X) = \sum_{i=1}^c \sum_{j=1}^{n_i} (x_i^{(j)} - m_i)(x_i^{(j)} - m_i)^T \quad \dots (6)$$

$$S_B(X) = \sum_{i=1}^c n_i (m_i - m)(m_i - m)^T \quad \dots (7)$$

where $x_i^{(j)}$ denotes the j -th training sample of the i -th class, m_i and m are mean vector of X_i and X respectively, $m_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_i^{(j)} \in \mathbb{R}^K$, $m = \frac{1}{N} \sum_{i=1}^c \sum_{j=1}^{n_i} x_i^{(j)} \in \mathbb{R}^K$.

Therefore

$$\text{tr}(S_W) = \sum_{i=1}^c \|X_i - M_i\|_F^2 \quad \dots (8)$$

$$\text{tr}(S_B) = \sum_{i=1}^c n_i \|m_i - m\|_2^2 = \sum_{i=1}^c \|M_i - M\|_F^2 \quad \dots (9)$$

where $M_i \in \mathbb{R}^{K \times n_i}$, each column equals m_i . For corresponding M_i , $M \in \mathbb{R}^{K \times N}$, each column equals m .

The combination of $\|X\|_1$ and $\|X\|_F^2$ can make the solution X more stable.³⁸⁾

The term $\|D_i^T D_i - I_{k_i}\|_F^2$ shows that each sub-dictionary is self-incoherent, which enhances incoherence for each sub-dictionary and has a direct impact on the speed of computation.^{27,44)}

The term $\|D_i^T D_{-i}\|_F^2$ denotes a cross-incoherent constraint between different sub-dictionaries, which promotes atoms of different sub-dictionary to be as independent as possible.^{27,44)}

The weight $w_1 \frac{n_i}{k_i^2}$ and $w_2 \frac{n_i}{k_i(K-k_i)}$ alleviate the effect of imbalance between the number of samples and atoms.²⁸⁾

An incoherent constraint on sub-dictionary was proposed in,^{27,28,35)} a Fisher discrimination constraint on coding vector was presented in,^{33,34)} and a class-specific and shared dictionary was adopted in.^{28,34,35)} However, there are mainly two differences between the proposed CDSDL algorithm

and these algorithms. Firstly, Ramirez²⁷⁾ and Yang³³⁾ only learn some class-specific sub-dictionaries, which can't separate the class-specific features and the shared features. Secondly, the constraints of these algorithms are imposed on either sub-dictionary^{27,28,35)} or coding vector,^{33,34)} while the proposed CSDSL algorithm explicitly imposes the constraints on both sub-dictionary and coding vector.

3.3. Optimization of CSDSL

The optimization of the CSDSL algorithm is not jointly convex, but is convex with respect to each variable when the others are fixed. The optimization problem can be divided into three sub-problems: (1) Solving X with fixed D ; (2) Solving D_i with fixed X and D_{-j} , $j \neq i$, where D_{-j} is the sub-matrix by removing D_j from D ; (3) Solving D_{c+1} with fixed X and $D_{-(c+1)}$. An effective iterative algorithm that alternatively optimizes objective function J with respect to coding vector matrix X , class-specific sub-dictionary D_i and shared sub-dictionary D_{c+1} is introduced as follows. The CSDSL algorithm is summarized in **Table 1**.

3.3.1. Update of X

With fixed D , J can be rewritten as follows.

$$\min_{X_i} \|Y_i - DX_i\|_F^2 + \left\| Y_i - D\tilde{Q}_i (\tilde{Q}_i)^T X_i \right\|_F^2 + \lambda_2 \left(\|X_i - M_i\|_F^2 - \sum_{i=1}^c \|M_i - M\|_F^2 + \|X_i\|_F^2 \right) + \lambda_1 \|X_i\|_1 \quad (10)$$

Equation (10) can be rewritten as follows.

$$\min_{X_i} F(X_i) + 2\tau \|X_i\|_1 \quad (11)$$

where $F(X_i) = \|Y_i - DX_i\|_F^2 + \|Y_i - D\tilde{Q}_i (\tilde{Q}_i)^T X_i\|_F^2 + \lambda_2 \left(\|X_i - M_i\|_F^2 - \sum_{i=1}^c \|M_i - M\|_F^2 + \|X_i\|_F^2 \right)$, $\tau = \lambda_1/2$.

Because $F(X_i)$ is convex and differentiable to X_i , a fast iterative shrinkage thresholding algorithm (FISTA)⁴²⁾ and a two-step iterative shrinkage/thresholding (TwIST) algorithm⁴⁵⁾ can be employed to solve Eq. (11).

The first derivative of $F(X_i)$ with respect to X_i is calculated, some detailed steps are listed in Appendix 1, then

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

Algorithm 1: Class-specific and Shared Dictionary Learning	
Input:	Training dataset $Y = \{Y_i\}_{i=1,2,\dots,c}$; size of sub-dictionary k_i , $i = 1, 2, \dots, c$, $c+1$; parameters $\lambda_1, \lambda_2, w_1, w_2$, and T .
Output:	The learned dictionary $D = \{D_i\}_{i=1,2,\dots,c,c+1}$, D_i is a class-specific sub-dictionary, $i \neq c+1$, D_{c+1} is a shared sub-dictionary.
Initialize:	The class-specific sub-dictionary D_i is initialized by K-SVD in Y_i , the shared sub-dictionary D_{c+1} is initialized by K-SVD in Y .
While stop criterion isn't reached	
Update of X	Equation (11) is solved by FISTA and TwIST
Update of D_i	Equation (18) is solved by MOCOD
Update of D_{c+1}	Equation (22) is solved by MOCOD
End While	

we have

$$\begin{aligned} \nabla_{X_i} F(X_i) &= 2D^T (DX_i - Y_i) \\ &+ 2\tilde{Q}_i (\tilde{Q}_i)^T D^T \left[D\tilde{Q}_i (\tilde{Q}_i)^T X_i - Y_i \right] \\ &+ 2\lambda_2 \left[X_i O_i O_i^T + X_i P_i P_i^T - R P_i^T + X_i Q_i^j (Q_i^j)^T - T_j (Q_i^j)^T \right] \end{aligned} \quad (12)$$

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

$$\frac{E_i^i}{n_i} - Q_i^j, \quad R = \sum_{j=1}^c X_j Q_j^j, \quad T_j = X_j \frac{E_j^j}{n_j} - \sum_{l=1}^c X_l Q_l^j.$$

According to FISTA and TwIST algorithm, we have

$$\begin{aligned} \{X_i\}_{t+1} &= (1-\alpha)\{X_i\}_{t-1} + (\alpha-\beta)\{X_i\}_t \\ &+ \beta S_{\frac{\tau}{\sigma}} \left[\{X_i\}_t - \frac{1}{2\sigma} \nabla_{X_i} F(X_i) \right] \end{aligned} \quad (13)$$

where $S_{\frac{\tau}{\sigma}}(\cdot)$ is a soft-threshold shrinkage operator, $\{X_i\}_{t-1}$ is the previous value, $\{X_i\}_t$ is the current value, $\{X_i\}_{t+1}$ is the next value, $\alpha > 0$, $\beta > 0$, $\sigma > 0$.

3.3.2. Update of D_i

With fixed X and D_{-j} , $j \neq i$, J can be rewritten as follows.

$$\begin{aligned} \min_{D_i} \sum_{i=1}^c \left(\|Y_i - DX_i\|_F^2 + \left\| Y_i - D\tilde{Q}_i (\tilde{Q}_i)^T X_i \right\|_F^2 \right) \\ + w_1 \frac{n_i}{k_i^2} \|D_i^T D_i - I_{k_i}\|_F^2 + w_2 \frac{n_i}{k_i(K-k_i)} \|D_i^T (DQ_{-i})\|_F^2 \end{aligned} \quad (14)$$

Then, we have

$$\begin{aligned} \min_{D_i} \|Y - D(X_1, \dots, X_c)\|_F^2 \\ + \left\| Y - D \left[\tilde{Q}_1 (\tilde{Q}_1)^T X_1, \dots, D\tilde{Q}_c (\tilde{Q}_c)^T X_c \right] \right\|_F^2 \\ + w_1 \frac{n_i}{k_i^2} \|D_i^T D_i - I_{k_i}\|_F^2 + w_2 \frac{n_i}{k_i(K-k_i)} \|D_i^T (DQ_{-i})\|_F^2 \end{aligned} \quad (15)$$

Denote $A = [Y, Y] \in \mathbb{R}^{d \times 2N}$, $B = [X_1, \dots, X_c, \tilde{Q}_1 (\tilde{Q}_1)^T X_1, \dots, D\tilde{Q}_c (\tilde{Q}_c)^T X_c] \in \mathbb{R}^{K \times 2N}$, $C = DQ_{-i} \in \mathbb{R}^{d \times (K-k_i)}$, we have

$$\begin{aligned} \min_{D_i} \|A - DB\|_F^2 + w_1 \frac{n_i}{k_i^2} \|D_i^T D_i - I_{k_i}\|_F^2 \\ + w_2 \frac{n_i}{k_i(K-k_i)} \|D_i^T C\|_F^2 \end{aligned} \quad (16)$$

Therefore

$$\begin{aligned} \min_{D_i} \left\| A - \sum_{j=1}^{c+1} D_j B^j - D_i B^i \right\|_F^2 + w_1 \frac{n_i}{k_i^2} \|D_i^T D_i - I_{k_i}\|_F^2 \\ + w_2 \frac{n_i}{k_i(K-k_i)} \|D_i^T C\|_F^2 \end{aligned} \quad (17)$$

Denote $\tilde{A} = A - \sum_{j=1}^{c+1} D_j B^j \in \mathbb{R}^{d \times 2N}$, Eq. (17) is equivalent to the following problem.

$$\min_{D_i} \left\| \tilde{A} - D_i B^i \right\|_F^2 + w_1 \frac{n_i}{k_i^2} \left\| D_i^T D_i - I_{k_i} \right\|_F^2 + w_2 \frac{n_i}{k_i(K-k_i)} \left\| D_i^T C \right\|_F^2 \quad (18)$$

where B^i denotes k_i rows of B corresponding to the sub-dictionary D_i .

Equation (18) can be solved by method of optimal coherence-constrained directions (MOCOD)⁴⁴⁾ algorithm.

3.3.3. Update of D_{c+1}

With fixed X and $D_{-(c+1)}$, J can be rewritten as follows.

$$\min_{D_{c+1}} \sum_{i=1}^c \left\| Y_i - D X_i \right\|_F^2 + \sum_{i=1}^c \left\| Y_i - D_i X_i^i - D_{c+1} X_i^{c+1} \right\|_F^2 + w_1 \frac{n_{c+1}}{k_{c+1}^2} \left\| D_{c+1}^T D_{c+1} - I_{k_{c+1}} \right\|_F^2 + w_2 \frac{n_{c+1}}{k_{c+1}(K-k_{c+1})} \left\| D_{c+1}^T (D Q_{-(c+1)}) \right\|_F^2 \quad (19)$$

Because $\sum_{i=1}^c \left\| Y_i - D X_i \right\|_F^2$ denotes the global reconstructive error, therefore

$$\min_{D_{c+1}} \left\| Y - \sum_{j=1}^c D_j X^j - D_{c+1} X^{c+1} \right\|_F^2 + \sum_{i=1}^c \left\| Y_i - D_i X_i^i - D_{c+1} X_i^{c+1} \right\|_F^2 + w_1 \frac{n_{c+1}}{k_{c+1}^2} \left\| D_{c+1}^T D_{c+1} - I_{k_{c+1}} \right\|_F^2 + w_2 \frac{n_{c+1}}{k_{c+1}(K-k_{c+1})} \left\| D_{c+1}^T (D Q_{-(c+1)}) \right\|_F^2 \quad (20)$$

where $X^{c+1} = (X_1^{c+1}, X_2^{c+1}, \dots, X_c^{c+1})$.

Denote $A = Y - D Q_{-(c+1)} (Q_{-(c+1)})^T X$, $B = (Y_1 - D_1 X_1^1, Y_2 - D_2 X_2^2, \dots, Y_c - D_c X_c^c)$, $C = D Q_{-(c+1)}$, then we have

$$\min_{D_{c+1}} \left\| A - D_{c+1} X^{c+1} \right\|_F^2 + \sum_{i=1}^c \left\| B - D_{c+1} X^{c+1} \right\|_F^2 + w_1 \frac{n_{c+1}}{k_{c+1}^2} \left\| D_{c+1}^T D_{c+1} - I_{k_{c+1}} \right\|_F^2 + w_2 \frac{n_{c+1}}{k_{c+1}(K-k_{c+1})} \left\| D_{c+1}^T C \right\|_F^2 \quad (21)$$

Denote $U = [A, B]$, $V = [X^{c+1}, X^{c+1}]$, Eq. (21) is equivalent to the following problem.

$$\min_{D_{c+1}} \left\| U - D_{c+1} V \right\|_F^2 + w_1 \frac{n_{c+1}}{k_{c+1}^2} \left\| D_{c+1}^T D_{c+1} - I_{k_{c+1}} \right\|_F^2 + w_2 \frac{n_{c+1}}{k_{c+1}(K-k_{c+1})} \left\| D_{c+1}^T C \right\|_F^2 \quad (22)$$

Similar to Eqs. (18), (22) can be solved by MOCOD algorithm.

3.4. Classification

For the CSDL algorithm, both the reconstructive error associated with different class and the coding vector can achieve good classification performance. Based on the coding vector, there are two kinds of classifier as follows.

For a global classifier, a test sample $\hat{y}$ is computed over

the whole dictionary D as follows.

$$\hat{x} = \arg \min_x \left\| \hat{y} - D x \right\|_2^2 + \lambda \left\| x \right\|_1 \quad (23)$$

According to $\arg \min_i \left\| \hat{y} - D_i \hat{x}_i \right\|_2^2$, the class of $\hat{y}$ is the index i of the smallest reconstruction error $\left\| \hat{y} - D_i \hat{x}_i \right\|_2^2$, where $\hat{x} = [\hat{x}_1; \dots; \hat{x}_i; \dots; \hat{x}_c; \hat{x}_{c+1}] \in \mathbb{R}^K$, $i = 1, 2, \dots, c$, $\hat{x}_i$ is the sub-vector of $\hat{x}$ that corresponding to the sub-dictionary D_i , the global coding vector $\hat{x}^G$ equals $\hat{x}$.

For a local classifier, we stack a class-specific sub-dictionary D_i and a shared sub-dictionary D_{c+1} to construct c compositional dictionaries. A test sample $\hat{y}$ is computed over each compositional dictionary as follows.

$$\hat{x}_i = \arg \min_x \left\| \hat{y} - (D_i, D_{c+1}) x \right\|_2^2 + \lambda \left\| x \right\|_1 \quad (24)$$

According to $\arg \min_i \left\| \hat{y} - (D_i, D_{c+1}) \hat{x}_i \right\|_2^2$, the class of $\hat{y}$ is the index i of the smallest reconstruction error $\left\| \hat{y} - (D_i, D_{c+1}) \hat{x}_i \right\|_2^2$, where $\hat{x}_i = [\hat{x}_i^i; \hat{x}_i^{c+1}] \in \mathbb{R}^{k_i+k_{c+1}}$, $i = 1, 2, \dots, c$, $\hat{x}_i^i$ is the sub-vector corresponding to the class-specific sub-dictionary D_i , $\hat{x}_i^{c+1}$ is the sub-vector corresponding to the shared sub-dictionary D_{c+1} .

In fact, $\hat{x}_i^{c+1}$, which just works for reconstruction, can be omitted. Therefore, we stack c remaining $\hat{x}_i^i$ together, and denote the local coding vector $\hat{x}^L = [\hat{x}_1^i; \dots; \hat{x}_c^i; \dots; \hat{x}_c^i] \in \mathbb{R}^{K-k_{c+1}}$.

The $\hat{x}^G$ and $\hat{x}^L$ can be directly treated as the input feature vector of linear one versus rest (OVR) SVM, respectively.

4. Experimental Results

The dataset of steel surface images as shown in Fig. 1 was collected from practical steel sheet production line. The dataset comprises four kinds of typical surface defects (Folding, Line, Patch, and Scratch) and one kind of non-defective (Original). There are 300 8-bit grayscale images per class, whose image size is 200×200 pixels. Each image is resized from 200×200 pixels to 40×40 pixels to avoid over-fitting. Because of non-uniform brightness of the images, each image is normalized to have zero mean and unit l_2 -norm. We randomly choose 50% images per class for a training dataset, and the rest is a testing dataset. The classification accuracy is defined as follows.

$$Acc = \frac{N_R}{N} \quad (25)$$

where N_R is the number of test samples that are correctly classified, N is the total number of test samples. To deal with random influence, we do each experiment 10 times and have the mean and standard deviation of results. The weight parameters λ_1 , λ_2 , w_1 , w_2 and sparsity T are tuned with 5-fold cross validation. Finally, we set $\lambda_1 = 0.1$, $\lambda_2 = 0.001$, $w_1 = 0.01$, $w_2 = 0.1$, $T = 5$, respectively. We consider further fine-tuning may further improve the classification performance.

The relationship between classification accuracy and size of sub-dictionary is shown in Fig. 3. Figure 3(a), in which the size of class-specific sub-dictionary is fixed, shows that classification performance decreases with increasing the size of shared sub-dictionary. The possible reasons are that smaller shared sub-dictionary is enough to capture the shared features of the defect images, and that larger shared

Fig. 3. The classification accuracy versus the number of atoms in sub-dictionary, where the legend denotes the number of atoms in the class-specific sub-dictionary (top) and the shared sub-dictionary (bottom), respectively.

sub-dictionary may cause redundancy of the feature vector. Figure 3(b), in which the size of shared sub-dictionary is fixed, shows the classification performance improves with increasing the size of class-specific sub-dictionary. The possible reason is that more discriminative information can be captured by a larger class-specific sub-dictionary. 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. When the class-specific sub-dictionary reaches to a certain size, classification performance won't develop. Therefore, we should make a tradeoff between classification performance and computational efficiency. As shown in Fig. 3, when the size of shared sub-dictionary $k_s = 2$ and the size of class-specific sub-dictionary $k_c = 25$, CSDSL can still have higher classification accuracy 94.25%. From $k_c = 50$ to $k_c = 25$, classification accuracy only drops by 1.58%. Therefore, the CSDSL algorithm is simple and efficient to learn a compact, reconstructive and discriminative dictionary, which can boost classification performance and reduce computational load.

Figure 4 shows that objective function value of the CSDSL algorithm decreases fast and converges within about 5–8 iterations.

Figure 5 shows that the global coding vector matrix of training dataset and testing dataset are nearly diagonal block. Each block indicates the main part of coding matrix for the corresponding class-specific sub-dictionary. Therefore, the discrimination of the coding vector which can improve classification performance is enhanced by the CSDSL algorithm.

The CSDSL algorithm is compared with several state-of-the-art methods, including COPAR,³⁵⁾ FDDL,³³⁾ LCKSVD³¹⁾ and SRC.²²⁾ As shown in **Table 2**, CSDSL with a local clas-

Fig. 4. The convergence curve of the CSDSL algorithm.

Fig. 5. The visualization image of global coding vector matrix (Folding, Line, Non-defect, Patch, Scratch): training dataset (top); testing dataset (bottom).

Table 2. The comparison of classification accuracy of different methods.

Method	Accuracy (%)
COPAR	91.85 ± 1.02
FDDL	51.89 ± 4.31
LCKSVD	66.99 ± 1.42
SRC	56.96 ± 1.20
CSDSL	94.25 ± 0.42

sifier achieves 94.25% classification accuracy, compared to 91.85% for COPAR, 51.89% for FDDL, 66.99% for LCKSVD and 56.96% for SRC. Compared to SRC, which is the baseline method in the experiment, CSDSL improves the classification accuracy with a margin of more than 37%. CSDSL outperforms FDDL by about 42.36%. CSDSL is competitive with 2.4% improvement to 91.85% achieved by COPAR. Therefore, CSDSL has better classification performance than these methods. Besides, a local classifier (94.25%) of CSDSL achieves much better performance than a global classifier (83.51%) with about 10% advancement. The probable reason is that the number of training samples

	Folding	Line	Patch	Scratch	Non-defect
Folding	149	0	0	9	0
Line	0	148	0	10	0
Patch	0	0	147	13	0
Scratch	1	0	3	117	0
Non-defect	0	2	0	1	150

Fig. 6. The confusion matrix, where the element located in the i -th row and the j -th column indicates the number of the j -th class sample is misclassified to the i -th class, $j \neq i$.

Fig. 7. Some scratch samples are misclassified to other defects: (a) Scratch→Folding; (b) Scratch→Line; (c) Scratch→Patch.

Fig. 8. The same image is polluted by Gaussian noise: (a) 15 dB; (b) 20 dB; (c) 25 dB; (d) Original.

of each class is relatively large, or the learned class-specific sub-dictionary has large size.

According to **Fig. 6**, error rate of scratch is bigger than other defects. There are 33 images of scratch are misclassified to 9 folding, 10 line and 13 patch. The possible reasons for this phenomenon are listed as follows. (1) These defects are similar with scratch in size, direction and illumination. (2) The CDSDL algorithm is sensitive to scratch defect, and thus generates the improper feature representation. We will make a further research for this phenomenon. **Figure 7** shows that some scratch samples are misclassified to other defects.

To evaluate robustness of the CDSDL algorithm, additive Gaussian noise was added to some surface images. We randomly choose 50% images per class as a training dataset, and the rest is a testing dataset. All the training samples are

Table 3. The classification accuracy of the CDSDL algorithm for different Gaussian noise.

SNR (dB)	Accuracy (%)
15	65.77 ± 2.50
20	85.88 ± 1.63
25	92.45 ± 0.29

added to Gaussian noise with different signal to noise ratio (SNR), including 15 dB, 20 dB and 25 dB. As shown in **Fig. 8**, the same image is polluted by Gaussian noise with different SNR. According to **Table 3**, CDSDL can achieve 92.45% classification accuracy even at 25 dB noise.

5. Conclusion

A classification method based on class-specific and shared sub-dictionary learning is presented in the paper. The class-specific sub-dictionary learning and shared sub-dictionary learning are jointly formulated with reconstructive error, sparse and discriminative promotion constraints. Each class-specific sub-dictionary has good performance to represent the sample of corresponding class, and the shared sub-dictionary contributes to reconstruction of all the samples. Therefore, the associated feature representation based on class-specific sub-dictionary and shared sub-dictionary not only faithfully represent the sample from anyone class, but also are more compact and discriminative for classification. Experimental results show that the CDSDL algorithm can achieve better accuracy in classification of surface defects of steel sheet.

Acknowledgements

This work was supported by the Natural Science Foundation of China, under grant No. 51174151, and the Specialized Research Fund for the Key Science and Technology Innovation Plan of Hubei Province, under grant No. 2013AAA011. The authors would like to thank Kechen Song and Yungang Tan for providing the defect images. MATLAB procedure was revised and optimized from.^{27,33,35,44)}

REFERENCES

- 1) T. S. Newman and A. K. Jain: *Comput. Vis. Image Underst.*, **61** (1995), 231.
- 2) D. Brzakovic and N. Vujovic: *Pattern Recogn.*, **29** (1996), 1401.
- 3) D. M. Tsai and T. Y. Huang: *Image Vis. Comput.*, **21** (2003), 307.
- 4) E. N. Malamas, E. G. M. Petrakis, M. Zervakis, L. Petit and J. D. Legat: *Image Vis. Comput.*, **21** (2003), 171.
- 5) X. H. Xie: *ELCVIA*, **7** (2008), 1.
- 6) R. Shanmugamani, M. Sadique and B. Ramamoorthy: *Measurement*, **60** (2015), 222.
- 7) B. K. Kwon, J. S. Won and D. J. Kang: *Int. J. Precis. Eng. Manuf.*, **16** (2015), 965.
- 8) N. Neogi, D. K. Mohanta and P. K. Dutta: *EURASIP J. Image. Video Process.*, **1** (2014), 1.
- 9) H. J. Hu, Y. X. Li, M. F. Liu and W. H. Liang: *Multimed. Tools Appl.*, **69** (2014), 199.
- 10) H. J. Hu, Y. Liu, M. F. Liu and L. Q. Nie: *Neurocomputing*, **181** (2015), 86.
- 11) M. F. Liu, Y. Liu, H. J. Hu and L. Q. Nie: *J. Vis. Commun. Image. R.*, **37** (2016), 70.
- 12) P. Caleb and M. Steuer: *Proc. IEEE KES*, IEEE, Piscataway, NJ, (2000), 103.
- 13) R. Medina, F. Gayubo, L. M. González-Rodrigo, D. Olmedo, J. Gómez-García-Bermejo, E. Zalama and J. R. Perán: *Int. J. Adv. Manuf. Tech.*, **57** (2011), 1087.

- 14) X. W. Zhang, Y. Q. Ding, Y. Y. Lv, A. Y. Shi and R. Y. Liang: *Expert Syst. Appl.*, **38** (2011), 5930.
- 15) C. S. Lee, C. H. Choi, J. Y. Choi, Y. K. Kim and S. H. Choi: *Proc. IEEE ICIP*, IEEE, Piscataway, NJ, (1996), 673.
- 16) F. Pernkopf: *Pattern Anal. Appl.*, **7** (2004), 333.
- 17) S. Ghorai, A. Mukherjee, M. Gangadharan and P. K. Dutta: *IEEE Trans. Instrum. Meas.*, **62** (2013), 612.
- 18) K. C. Song and Y. H. Yan: *Appl. Surf. Sci.*, **285** (2013), 858.
- 19) M. X. Chu and R. F. Gong: *ISIJ Int.*, **55** (2015), 1956.
- 20) K. Xu, S. H. Liu and Y. H. Ai: *Image Vis. Comput.*, **35** (2015), 23.
- 21) J. Wright, Y. Ma and J. Mairal: *Proc. IEEE*, **98** (2015), 1031.
- 22) J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma: *IEEE Trans. Pattern Anal.*, **31** (2008), 210.
- 23) I. Tošić and P. Frossard: *IEEE Signal Proc. Mag.*, **28** (2011), 27.
- 24) R. Rubinstein, A. M. Bruckstein and M. Elad: *Proc. IEEE*, **98** (2010), 1045.
- 25) K. Engan, S. O. Aase and J. H. Husøy: *Signal Process.*, **80** (2000), 2121.
- 26) M. Aharon, M. Elad and A. K. Bruckstein: *IEEE Trans. Signal Process.*, **54** (2006), 4311.
- 27) I. Ramirez, P. Sprechmann and G. Sapiro: *Proc. IEEE CVPR*, IEEE, Piscataway, NJ, (2010), 3501.
- 28) S. H. Gao, I. W. H. Tsang and Y. Ma: *IEEE Trans. Image Process.*, **23** (2014), 623.
- 29) K. Huang and S. Aviyente: *Proc. NIPS*, Vol. 19, IEEE, Piscataway, NJ, (2006), 609.
- 30) Q. Zhang and B. X. Li: *Proc. IEEE CVPR*, IEEE, Piscataway, NJ, (2010), 2691.
- 31) Z. L. Jiang, Z. Lin and L. S. Davis: *IEEE Trans. Pattern Anal.*, **35** (2013), 2651.
- 32) J. Mairal, F. Bach and J. Ponce: *IEEE Trans. Pattern Anal.*, **34** (2012), 791.
- 33) M. Yang, L. Zhang, X. C. Feng and D. Zhang: *Int. J. Comput. Vis.*, **109** (2014), 209.
- 34) N. Zhou and J. P. Fan: *IEEE Trans. Pattern Anal.*, **36** (2013), 715.
- 35) D. H. Wang and S. Kong: *Pattern Recogn.*, **47** (2014), 885.
- 36) M. Elad, M. A. T. Figueiredo and Y. Ma: *Proc. IEEE*, **98** (2010), 972.
- 37) A. M. Bruckstein, D. L. Donoho and M. Elad: *SIAM Rev.*, **51** (2009), 34.
- 38) H. Zou and T. Hastie: *J. R. Stat. Soc. B*, **67** (2005), 301.
- 39) A. Y. Yang, A. Genesh and Z. H. Zhou: *IEEE T. Image Process.*, **22** (2013), 3234.
- 40) Z. Zhang, Y. Xu, J. Yang, X. L. Li and D. Zhang: *IEEE Access*, **3** (2015), 490.
- 41) Y. C. Pati, R. Rezaifar and P. S. Krishnaprasad: *Proc. of 27th Asilomar Conf. on Signals, Systems and Computers*, IEEE, Piscataway, NJ, (1993), 40.
- 42) A. Beck and M. Teboulle: *SIAM J. Imaging Sci.*, **2** (2009), 183.
- 43) R. O. Duda, P. E. Hart and D. G. Stork: *Pattern Classification*, 2nd ed., John Wiley&Sons, New York, (2001), 30.
- 44) I. Ramirez, F. Lecumberry and G. Sapiro: *Technical Report 2279*, Institute for Mathematics and its Application, University of Minnesota, MN, (2009), 1.
- 45) J. M. Bioucas-Dias and M. A. T. Figueiredo: *IEEE Trans. Image Process.*, **16** (2007), 2992.

Appendix

Appendix 1: Derivation of $F(X_i)$ in Eq. (12)

$$\text{Denote } E_i^j = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}_{n_i \times n_j}, \quad I_{n_i \times n_i} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}_{n_i \times n_i},$$

$$O_i = I_{n_i \times n_i} - \frac{E_i^i}{n_i}, \text{ then, we have}$$

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

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

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

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

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

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

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

Separate X_i , we have $\left\| X_j \frac{E_j^j}{n_j} - \left(X_i Q_i^j + \sum_{l=1, l \neq i}^c X_l Q_l^j \right) \right\|_F^2$

$$= \left\| X_j \frac{E_j^j}{n_j} - \sum_{l=1, l \neq i}^c X_l Q_l^j - X_i Q_i^j \right\|_F^2 = \|T_j - X_i Q_i^j\|_F^2$$

Expand the anyone non i -th class, $\sum_{j=1, j \neq i}^c \|T_j - X_i Q_i^j\|_F^2$,

then, we have

$$\begin{aligned} \|X_i - M_i\|_F^2 + \sum_{i=1}^c \|M_i - M\|_F^2 &= \|X_i O_i\|_F^2 - \|X_i P_i - R\|_F^2 \\ &\quad - \sum_{j=1, j \neq i}^c \|T_j - X_i Q_i^j\|_F^2 \end{aligned}$$

Calculate the first derivative with respect to X_i , we have

$$\begin{aligned} \frac{\partial \|X_i O_i\|_F^2}{\partial X_i} &= 2 X_i O_i O_i^T \\ \frac{\partial \|X_i P_i - R\|_F^2}{\partial X_i} &= 2 (X_i P_i P_i^T - R P_i^T) \\ \frac{\partial \sum_{j=1, j \neq i}^c \|T_j - X_i Q_i^j\|_F^2}{\partial X_i} &= 2 \sum_{j=1, j \neq i}^c [X_i Q_i^j (Q_i^j)^T - T_j (Q_i^j)^T] \end{aligned}$$