Neural Computing and Applications
https://doi.org/10.1007/s00521-019-04419-y

ORIGINAL ARTICLE

Nonnegative discriminative encoded nearest points for image set classification

Zhenwen Ren[1,2] · Quansen Sun[2] · Chao Yang[1]

Received: 18 March 2019 / Accepted: 7 August 2019

Abstract
Image set classification has drawn much attention due to its rich set information. Recently, the most popular set-to-set distance-based representation methods have achieved interesting results by measuring the between-set distance. However, there are two intuitive assumptions, which are largely ignored: (1) The homogeneous samples should have positive contributions to approximate the nearest point in the probe set, while the heterogeneous samples should have no contributions and (2) the learned nearest points in each gallery set should have the lowest correlations. Therefore, this paper presents a novel method called nonnegative discriminative encoded nearest points for image set classification. Specifically, we use two explicit nonnegative constraints to ensure the coding coefficients sparse and discriminative simultaneously. Moreover, we additionally introduce a new class-wise discriminative term to further boost the discriminant ability of different sets. In this way, they can be boosted mutually so that the obtained coding coefficients are beneficial to the purpose of set classification. The results from extensive experiments and comparisons with some state-of-the-art methods on four challenging datasets demonstrate the superiority of our method.

Keywords Image set classification · Nonnegative coding · Sparse representation · Collaborative representation

1 Introduction

With the development of sensing and imaging technologies, multiple images can be easily achieved. Against this background, image set classification has became one of the most important tasks in computer vision [1–3], which aims to recognize a subject from a set of related images that form an image set. Accordingly, this technology has been widely applied in various real-world scenarios, including multi-view visual recognition, video-based surveillance, video retrieval, dynamic scene recognition, etc. Different from traditional single-shot image classification technology [4, 5], image set classification is more promising since it can effectively deal with a variety of appearance variations for improving the discrimination power and robustness. These variations could be caused by pose, illumination, non-rigid deformation, obscuration, misalignment, etc. However, image set shows large inter-class ambiguity and huge intra-class variability, which also brings tremendous challenges to make effective use of the potential supplemental information and faithfully measure the dissimilarity between image sets for accurate classification as well.

Over the past few years, lots of image set classification methods have been proposed [6, 7]. Among these methods, the set-to-set distance-based representation technologies (e.g., sparse representation classification (SRC) [2, 8–12], collaborative representation classification (CRC) [2, 9, 13], affine hull distance (AHD) [14], convex hull distance (CHD) [14], and regularized affine hull (RAH) [15]) have

✉ Zhenwen Ren
rzw@njust.edu.cn

✉ Quansen Sun
sunquansen@njust.edu.cn

Chao Yang
ychao1983@126.com

[1] Department of National Defence Science and Technology, Southwest University of Science and Technology, Room 606, Building 6A, Mianyang 621010, China

[2] Department of Computer Science, Nanjing University of Science and Technology, Room 3014, Nanjing 210094, China

Published online: 20 August 2019

been widely studied in the field of image set classification [2, 6, 14–22].

However, almost all of the previous set-to-set distance-based representation methods have largely ignored two intuitive assumptions: (1) The homogeneous samples should have positive contributions to approximation the nearest point of the probe set, while the heterogeneous samples should have no contributions, i.e., the coding coefficients of the gallery sets should be nonnegative. Moreover, the nearest point of the probe set more likely should be a real meaningful sample, i.e., the coding coefficient of the probe set also should be nonnegative. Otherwise, the coding coefficients have mixed signs and they may severely deviate from the ideal solution. (2) The learned nearest points of all the gallery sets should have the lowest correlations, which can explore the discrimination capability of coding coefficients between different image sets. Unlike the previous methods, this paper presents a novel method called nonnegative discriminative encoded nearest points (NDENP) for image set classification. Based on the proposed two assumptions, we therefore incorporate two nonnegative constraints on the coding coefficients and one class-wise discriminative term into our model so that the learned nearest points are more discriminative and reasonable. Specially, the nonnegative constrain can automatically enhance the representation ability of homogeneous samples while limiting the representation ability of heterogeneous ones. Therefore, it can encourage the coding coefficients sparse (as l_1 norm) and discriminative simultaneously. In addition, the class-wise discriminative term employs the label information and encourages mutual exclusion among the nearest points of the gallery sets to further boost the discriminant ability of different sets. In summary, the contributions of this paper are as follows.

1. We first introduce the nonnegative constraint to representation-based method for image set classification, which can implicitly generate sparse and discriminative coding coefficients.
2. We propose an effective class-wise discriminative term to further enhance the distinctiveness of different gallery sets. It forces the nearest points of different gallery sets to hold the lowest correlation, which is helpful for obtaining discriminative class-specific residuals. In doing so, it is very beneficial to produce higher classification accuracy when recognizing a probe set.
3. We develop an efficient and fast algorithm based on the alternating direction method of multipliers (ADMM) to solve the proposed model. It is noteworthy that our model only contains l_2 norm and does not contain l_1 norm, but it implies the l_1 norm. As we know, l_2 norm is efficient due to a closed-form solution.
4. Some set-based classification tasks on four datasets show that NDENP outperforms the state-of-the-art image set classification methods for the tasks of face recognition and object recognition.

The rest of this paper is arranged as follows. Section 3 presents the proposed NDENP method and its solution for image set classification. Section 4 presents extensive experimental results and mechanism analysis. Section 5 concludes this work.

2 Related works

Generally speaking, there are two major steps involved in image set classification, which are to effectively model an image set and to define an appropriate metric to compute the dissimilarity between two sets. According to the model type, relevant methods mainly fall into five categories [6]: statistical model-based methods [23], linear subspace-based methods [16, 24], nonlinear manifold-based methods [17], affine subspace-based methods [14, 15, 19, 22, 23], and compressed sensing-based methods [1, 19, 25]. Besides, deep learning has recently gained significant success in some tasks, but their applications to image set classification are few; the most recent articles are [26–28].

Among the above methods, along with the affine subspace and compressed sensing lines, the set-to-set distance-based representation technologies, SRC, CRC, AHD, CHD, and RAH, have been widely studied and gained significant successes. The similarity between two image sets is the distance between a pair of optimal nearest points belonging to either hull, respectively. In 2010, Cevikalp and Triggs [14] proposed a simple but efficient model called AHD and CHD hull to model image set, which tends to complement the unseen appearance variations that even do not appear in the image set via covering the affine combinations of sample images in this set. They used affine hull-based image set distance (AHISD) and convex hull-based image set distance (CHISD) to perform image set classification and achieved very interesting results on image set-based face and object recognition. However, the AHD and CHD are loose representation models, which bring challenges for computing the distance between two image sets. The reason is that the hulls of image sets are likely to be overlarge, which may results in the intersection of two hulls.

In order to address these challenges, Hu et al. [19] introduced sparse approximated nearest points (SANP) method for measuring the between-set dissimilarity, in which SANP of two image sets are defined as the nearest points of the sets that can be sparsely approximated by the sample images of the respective set individually. SANP achieves state-of-the-art performance compared to previous methods, but it takes more computing time and storage resources. To reduce the difficulty and complexity of solving SANP, by modeling an image set as a regularized affine hull (RAH), regularized nearest points (RNP) [15] and joint regularized nearest points (JRNP) [29] have been proposed. However, on the one hand, the challenge problem, intersection of two hulls, is a problem that has not been resolved yet. On the other hand, the discriminative information is ignored, where the hulls modeled based on original feature may not suffice to be discriminated linearly.

To address these problems, joint the hull model [14], prototype learning [30], and metric learning [31] methods have recently been proposed to utilize the set discriminative information. For example, set-to-set prototype and metric learning (SPML) framework [32] jointly learns prototypes and a Mahalanobis distance. Prototype discriminative learning (PDL) [22] simultaneously learns a set of prototypes for each image set and an orthometric discriminative projection to shrink the loose affine hull.

With the development of the theory of compressed sensing, Wright et al. [8] used sparse representation (SR) for face recognition and the performance is impressive; SRC method emphasizes much on the role of l_1-norm sparsity of representation coefficients, which sparsely represents a probe face image with a dictionary constructed from all gallery examples, and then classifies it into the subject with the smallest reconstruction residual. The so-called CRC with l_2-regularization brings about similar experimental results to SRC but with much less computation time. Based on SRC and CRC, Zhu et al. [2] introduced them into image set classification and showed favorable properties in face recognition task in terms of both recognition rate and computational efficiency.

In summary, the related works have the following disadvantages: (1) The hull model may suffer from the issue of intersection, which makes the subsequent distance computation incorrect. (2) The discriminative information is ignored. (3) The l_1 norm is widely used in these methods; however, the l_1 norm will lead to relatively higher storage burden and computational complexity than l_2 norm. To address these problems, in this paper, we propose two reasonable assumptions to strengthen the set discriminative ability and to accelerate the solving procedure (see Sect. 1).

3 Proposed method

A total of C gallery image sets $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_C]$ for training and a probe set $\mathbf{Y}$ for testing are given. The cth set can be expressed as a data matrix with n_i image samples, i.e., $\mathbf{X}_c = \{x_j\}_{j=1}^{n_i}$. Each column x_j of $\mathbf{X}_c$ could be a feature vector. The coding coefficients of the gallery sets and the probe set are $\mathbf{A}$ and $\mathbf{B}$, respectively. In addition, $\mathbf{XA}$ is the artificial virtual nearest point of the gallery sets $\mathbf{X}$, $\mathbf{X}_i\mathbf{A}_i$ is the artificial nearest point of the ith gallery set, and $\mathbf{YB}$ is the artificial virtual nearest point of the probe set. An illustration of these notations is shown in Fig. 1.

3.1 Problem formulation

The nearest points distance is a popular strategy used for measuring the dissimilarity between the probe set and the gallery ones, i.e., $\min_{\mathbf{A},\mathbf{B}} \|\mathbf{XA} - \mathbf{YB}\|_2^2$. Based on the discussion in the above section, we can obtain sparse and discriminative coding coefficients by optimizing the following objective function:

$$\min_{\mathbf{A},\mathbf{B}} \|\mathbf{XA} - \mathbf{YB}\|_2^2 + \alpha \sum_{i=1}^{C}\sum_{j=1}^{C} (\mathbf{X}_i\mathbf{A}_i)^T \mathbf{X}_j\mathbf{A}_j \qquad (1)$$
$$\text{s.t.} \quad \mathbf{A} \geq \mathbf{0}, \mathbf{B} \geq \mathbf{0}$$

where α is nonnegative trade-off parameter and $\mathbf{0}$ is zero vector. The first term in (1) measures the distance between the nearest points pairs of the probe set and the whole gallery sets, i.e., the dissimilarity between the probe set and the whole gallery sets. The second term is the proposed class-wise discriminative term, where the label information is incorporated to learn discriminative representation. We encourage the nearest points in different gallery sets to have the lowest correlations, i.e., $\min_{\mathbf{A}} (\mathbf{X}_i\mathbf{A}_i)^T \mathbf{X}_j\mathbf{A}_j$. In other words, it encourages the nearest points of the gallery sets to boost mutual exclusion. Moreover, there are two nonnegative coding constraints $\mathbf{A} \geq \mathbf{0}, \mathbf{B} \geq \mathbf{0}$ used for replacing the l_1 norm and ensuring the coding coefficients

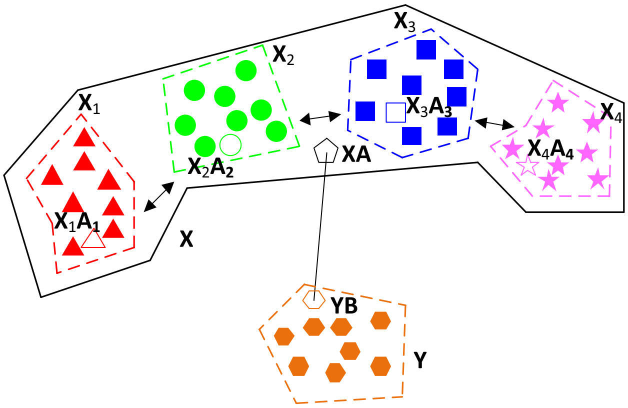


Fig. 1 Conceptual overview of the proposed method. There are four gallery sets (i.e., $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_4\}$) and one probe set (i.e., $\mathbf{Y}$), where hollow icons are artificial nearest points (i.e., $\mathbf{XA}, \mathbf{X}_1\mathbf{A}_1, \mathbf{X}_2\mathbf{A}_2, \mathbf{X}_3\mathbf{A}_3, \mathbf{X}_4\mathbf{A}_4$, and $\mathbf{YB}$), solid icons are given exemplars, and two-way arrows represent mutual exclusion between nearest points

sparse and discriminative simultaneously, which means the homogeneous samples should have positive contributions to approximation the nearest point in the probe set, while the heterogeneous samples should have no contributions. Figure 1 presents a conceptual overview of the proposed NDENP method.

3.2 Optimization

In this section, we show how to solve the optimization problem (1). We first introduce two auxiliary variables $\mathbf{Z}_1$ and $\mathbf{Z}_2$ in order to make the objective function separable. Therefore, it can be reformulated as follows:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B}} \|\mathbf{XA}-\mathbf{YB}\|_2^2 + \alpha\sum_{i=1}^{C}\sum_{j=1}^{C}(\mathbf{X}_i\mathbf{A}_i)^T\mathbf{X}_j\mathbf{A}_j \\ &\quad \text{s.t.} \quad \mathbf{A}\geq\mathbf{0}, \mathbf{B}\geq\mathbf{0}, \mathbf{Z}_1=\mathbf{A}, \mathbf{Z}_2=\mathbf{B} \end{aligned} \tag{2}$$

Then, we use the ADMM to obtain the optimal solutions of variables $\mathbf{A}$, $\mathbf{B}$, $\mathbf{Z}_1$, and $\mathbf{Z}_2$. To remove two linear constraints in problem (2), we introduce two Lagrange multipliers. Correspondingly, the augmented Lagrangian function of problem (2) is:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B},\mathbf{Z}_1,\mathbf{Z}_2} \|\mathbf{XA}-\mathbf{YB}\|_2^2 + \lambda\sum_{i=1}^{C}\sum_{j=1}^{C}(\mathbf{X}_i\mathbf{A}_i)^T\mathbf{X}_j\mathbf{A}_j \\ &\quad + \langle\mathbf{Y}_1, \mathbf{Z}_1-\mathbf{A}\rangle + \langle\mathbf{Y}_2, \mathbf{Z}_2-\mathbf{B}\rangle \\ &\quad + \frac{\mu}{2}\|\mathbf{Z}_1-\mathbf{A}\|_2^2 + \frac{\mu}{2}\|\mathbf{Z}_2-\mathbf{B}\|_2^2 \\ &\quad \text{s.t.} \quad \mathbf{Z}_1\geq 0, \mathbf{Z}_2\geq 0 \end{aligned} \tag{3}$$

where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are Lagrangian multipliers, and μ is a positive penalty parameter. Therefore, (3) can be minimized with respect to $\mathbf{A}$, $\mathbf{B}$, $\mathbf{Z}_1$, and $\mathbf{Z}_2$, respectively, by fixing the remaining variables and then updating the Lagrange multipliers $\mathbf{Y}_1$ and $\mathbf{Y}_2$. The solution scheme is as follows:

Step 1 Updating $\mathbf{A}$

By fixing all other variables, the optimal $\mathbf{A}$ is the solution to the following problem:

$$\begin{aligned} &\min_{\mathbf{A}} \|\mathbf{XA}-\mathbf{YB}\|_2^2 + \frac{\mu}{2}\left\|\mathbf{A}-\left(\mathbf{Z}_1+\frac{\mathbf{Y}_1}{\mu}\right)\right\|_2^2 \\ &\quad + \lambda\sum_{i=1}^{C}\sum_{j=1}^{C}(\mathbf{X}_i\mathbf{A}_i)^T\mathbf{X}_j\mathbf{A}_j \end{aligned} \tag{4}$$

For the third term in (4), we denote it as $\phi_1(\mathbf{A})$. Obviously, it does not explicitly contain $\mathbf{A}$; we first seek partial derivatives w.r.t $\mathbf{A}_k$, i.e., $\partial\phi_1/\partial\mathbf{A}_k$, and then extend it to $\partial\phi_1/\partial\mathbf{A}$. We rewrite $\phi_1(\mathbf{A})$ as follows:

$$\begin{aligned} \phi_1(\mathbf{A}) &= \sum_{i=1}^{C}\sum_{j=1}^{C}(\mathbf{X}_i\mathbf{A}_i)^T\mathbf{X}_j\mathbf{A}_j \\ &= \sum_{\substack{i=1,\ldots,C \\ i\neq k}}(\mathbf{X}_i\mathbf{A}_i)^T\mathbf{X}_k\mathbf{A}_k + \sum_{\substack{j=1,\ldots,C \\ j\neq k}}(\mathbf{X}_k\mathbf{A}_k)^T\mathbf{X}_j\mathbf{A}_j \\ &\quad + \sum_{\substack{i=1,\ldots,C \\ i\neq k}}\sum_{\substack{j=1,\ldots,C \\ j\neq k}}(\mathbf{X}_i\mathbf{A}_i)^T\mathbf{X}_j\mathbf{A}_j \\ &= 2\sum_{\substack{j=1,\ldots,C \\ j\neq k}}(\mathbf{X}_k\mathbf{A}_k)^T\mathbf{X}_j\mathbf{A}_j \\ &\quad + \sum_{\substack{i=1,\ldots,C \\ i\neq k}}\sum_{\substack{j=1,\ldots,C \\ j\neq k}}(\mathbf{X}_i\mathbf{A}_i)^T\mathbf{X}_j\mathbf{A}_j \end{aligned} \tag{5}$$

Then, we ignore some terms which are independent on $\mathbf{A}_k$, and the partial derivative over $\mathbf{A}_k$ of $\phi_1(\mathbf{A})$ is

$$\begin{aligned} \frac{\partial(\phi_1(\mathbf{A}))}{\partial\mathbf{A}_k} &= 2\frac{\partial}{\partial\mathbf{A}_k}\left(\sum_{\substack{j=1,\ldots,C \\ j\neq k}}(\mathbf{X}_k\mathbf{A}_k)^T\mathbf{X}_j\mathbf{A}_j\right) \\ &= 2\sum_{\substack{j=1,\ldots,C \\ j\neq k}}\mathbf{X}_k^T\mathbf{X}_j\mathbf{A}_j = 2\mathbf{X}_k^T(\mathbf{XA}-\mathbf{X}_k\mathbf{A}_k) \end{aligned} \tag{6}$$

Next, the derivative $(\frac{\partial\phi_1}{\partial\mathbf{A}})$ is

$$\begin{aligned} \frac{\partial\phi_1}{\partial\mathbf{A}} &= \begin{pmatrix} \frac{\partial\phi_1}{\partial\mathbf{A}_1} \\ \vdots \\ \frac{\partial\phi_1}{\partial\mathbf{A}_C} \end{pmatrix} = \begin{pmatrix} 2\mathbf{X}_1^T(\mathbf{XA}-\mathbf{X}_1\mathbf{A}_1) \\ \vdots \\ 2\mathbf{X}_C^T(\mathbf{XA}-\mathbf{X}_C\mathbf{A}_C) \end{pmatrix} \\ &= 2\mathbf{X}^T\mathbf{XA} - 2\begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{X}_C^T\mathbf{X}_C \end{pmatrix}\mathbf{A} \end{aligned} \tag{7}$$

and we have $\frac{\partial\phi_1}{\partial\mathbf{A}} = 2\mathbf{X}^T\mathbf{XA} - 2\mathbf{GA}$, where

$$\mathbf{G} = \begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{X}_C^T\mathbf{X}_C \end{pmatrix}$$

For the first two terms in (4), we denote it as $\phi_2(\mathbf{A})$, i.e.,

$$\phi_2(\mathbf{A}) = \|\mathbf{XA} - \mathbf{YB}\|_2^2 + \frac{\mu}{2}\|\mathbf{A} - \left(\mathbf{Z}_1 + \frac{\mathbf{Y}_1}{\mu}\right)\|_2^2 \tag{8}$$

Taking the derivative of the above two partial derivatives (7) and (8) w.r.t. $\mathbf{A}$ and setting it to zero, i.e., $\frac{\partial(\phi_1(\mathbf{A}))}{\partial\mathbf{A}} + \frac{\partial(\phi_2(\mathbf{A}))}{\partial\mathbf{A}} = 0$, we then can infer the optimal closed-form solution of $\mathbf{A}$ given by

$$\begin{aligned} \mathbf{G}_1 &= \left(\mathbf{X}^T\mathbf{X} + \frac{\mu}{2}\mathbf{I} + \lambda\mathbf{X}^T\mathbf{X} - \lambda\mathbf{G}\right) \\ \mathbf{G}_2 &= \left(\mathbf{X}^T\mathbf{YB} + \frac{\mu}{2}\mathbf{Z}_1 + \frac{1}{2}\mathbf{Y}_1\right) \\ \mathbf{A} &= \mathbf{G}_1^{-1}\mathbf{G}_2 \end{aligned} \tag{9}$$

where $\mathbf{I}$ is the identity matrix and $(\cdot)^{-1}$ is the matrix inverse transformation.

Step 2 Update $\mathbf{B}$

The optimization subproblem for updating $\mathbf{B}$ can be written as follows:

$$\min_{\mathbf{B}} \|\mathbf{XA} - \mathbf{YB}\|_2^2 + \frac{\mu}{2}\|\mathbf{B} - \left(\mathbf{Z}_2 + \frac{\mathbf{Y}_2}{\mu}\right)\|_2^2 \tag{10}$$

Note that the optimization procedure of problem (10) is similar to that of problem (8), so we can directly obtain its closed solution by

$$B = \left(\mathbf{Y}^T\mathbf{Y} + \frac{\mu}{2}\mathbf{I}\right)^{-1}\left(\mathbf{Y}^T\mathbf{XA} + \frac{\mu}{2}\mathbf{Z}_2 + \frac{1}{2}\mathbf{Y}_2\right) \tag{11}$$

Step 3 Solve $\mathbf{Z}_1$ and $\mathbf{Z}_2$

The solutions of the subproblems for updating $\mathbf{Z}_1$ and $\mathbf{Z}_2$ can be easily obtained by

$$\mathbf{Z}_1 = \max\left(\mathbf{A} - \frac{\mathbf{Y}_1}{\mu}, 0\right), \quad \mathbf{Z}_2 = \max\left(\mathbf{B} - \frac{\mathbf{Y}_2}{\mu}, 0\right) \tag{12}$$

Step 4 Check the convergence condition

These update steps (i.e., (9), (11), and (12)) will be repeated until the object function satisfies convergence condition or reaches the maximum number of iterations. The convergence condition used in our algorithm is verified at each iteration by checking the following constraints:

$$\|\mathbf{A} - \mathbf{Z}_1\|_\infty < \varepsilon, \quad \|\mathbf{B} - \mathbf{Z}_2\|_\infty < \varepsilon \tag{13}$$

where ε is the stop threshold.

Algorithm 1 Non-Negative Discriminative Encoded Nearest Points (NDENP) for image set classification.

```
Input: The gallery sets $\mathbf{X}$, the probe set $\mathbf{Y}$, the number of atoms in the dictionary or the number of all training samples of the gallery sets n, the number of samples in the probe set $m$, the trade-off parameters $\alpha$, $\mu$.
    Initialize: $\mathbf{A}^1 = 1/n$, $\mathbf{B}^1 = 1/m$, $\mathbf{Z}_1^1 = \mathbf{A}^1$, $\mathbf{Z}_2^1 = \mathbf{B}^1$, $\mathbf{Y}_1^1 = \mathbf{0}$, $\mathbf{Y}_2^1 = \mathbf{0}$, the stop threshold $\varepsilon = 10^{-5}$, $t = 1$, $\rho = 1.01$, $\mu_{max} = 10^5$ and $maxIter = 100$.
Learn the dictionary $\mathbf{D}$ by using dictionary learning method if we want to use compassed the gallery sets for training.
while convergence criterion (13) is not satisfied, and $t < maxIter$ do
    Update $\mathbf{A}^{t+1}$ by using the solution (9).
    Update $\mathbf{B}^{t+1}$ by using the solution (11).
    Update $\mathbf{Z}_1^{t+1}$ by using the solution (12).
    Update $\mathbf{Z}_2^{t+1}$ by using the solution (12).
    Update the multipliers:
    $\mathbf{Y}_1^{t+1} = \mathbf{Y}_1^t + \mu(\mathbf{Z}_1^{t+1} - \mathbf{A}^{t+1})$
    $\mathbf{Y}_2^{t+1} = \mathbf{Y}_2^t + \mu(\mathbf{Z}_2^{t+1} - \mathbf{B}^{t+1})$.
    Update $\mu$:
    $\mu^{k+1} = \min(\rho\mu^k, \mu_{\max})$
    Update $t$: $t = t + 1$.
end while
Perform image set classification by using (14) and (15).
Output: The class label of the probe set.
```

In order to speed up the learning process and obtain compacted data, research has demonstrated that learning a specific class dictionary from original data instead of using original data could lead to better performance [33]. Therefore, a specific class dictionary can adequately capture the main characteristics of an image set, i.e., compress $\mathbf{X}$ into a more compact set $\mathbf{D}$. Hence, we can replace $\mathbf{X}$ with $\mathbf{D}$ to speed up the learning process by using dictionary learning methods such as metaface learning [34] and KSVD [33]. Note that the gallery sets corresponding dictionaries $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_C]$ are learned off-line. Specifically, the samples themselves can be employed as the reconstruction dictionary.

The complete algorithm of the model (2) is outlined in Algorithm 1, in which the main time-consuming operation is solving inverse of square matrix, i.e., update $\mathbf{A}$ and $\mathbf{B}$. To be precise, the computational complexity of inverse transformation of a $n \times n$ square matrix is $\mathcal{O}(n^3)$. Hence, for updating $\mathbf{A}$, the computational complexity of $(\mathbf{X}^T\mathbf{X} + \frac{\mu}{2}\mathbf{I} + \lambda\mathbf{X}^T\mathbf{X} - \lambda\mathbf{G})^{-1}$ is $\mathcal{O}(N^3)$, where N is the column size of the compact dictionary. For updating $\mathbf{B}$, the computational complexity of $(\mathbf{Y}^T\mathbf{Y} + \frac{\mu}{2}\mathbf{I})^{-1}$ is $\mathcal{O}(M^3)$, where M is the column size of the probe set. In combination, the computational complexity of each iteration of NDNEP is about to $\mathcal{O}(N^3 + M^3)$, and the total time complexity of NDNEP can be loosely thought of as $\mathcal{O}(t(N^3 + M^3))$, where t is the iteration number.

3.3 Classification

Once the optimal coding coefficients $\mathbf{A}$ and $\mathbf{B}$ are learned, we define the distance between the probe set $\mathbf{Y}$ and the ith gallery set $\mathbf{X}_i$ as [11, 15], i.e.,

$$\mathbf{E}_i = (\|\mathbf{X}_i\|_* + \|\mathbf{Y}\|_*) \cdot \|\mathbf{X}_i\mathbf{A}_i - \mathbf{Y}\mathbf{B}\|_2^2, \qquad (14)$$

where $\|\cdot\|_*$ is the nuclear norm (i.e., the sum of the singular values), $\|\mathbf{X}_i\mathbf{A}_i - \mathbf{Y}\mathbf{B}\|_2^2$ represents the class-specific reconstruction residual between the nearest points of $\mathbf{Y}$ and $\mathbf{X}_i$, and $(\|\mathbf{X}_i\|_* + \|\mathbf{Y}\|_*)$ aims to alleviate the disturbance unrelated to the class information of image set.

The procedures of NDNEP are summarized as follows.

Step 1 Calculate the optimal coding coefficients $\mathbf{A}$ and $\mathbf{B}$ using Algorithm 1.

Step 2 Calculate residual $\mathbf{E}_i$ by Eq. (14), where $i = 1, 2, \ldots, C$.

Step 3 Classify the probe image set $\mathbf{Y}$ by selecting a minimal residual, i.e.,

$$\text{label } (\mathbf{Y}) = \arg\min_i \{\mathbf{E}_i\}. \qquad (15)$$

4 Experiments

4.1 Datasets and settings

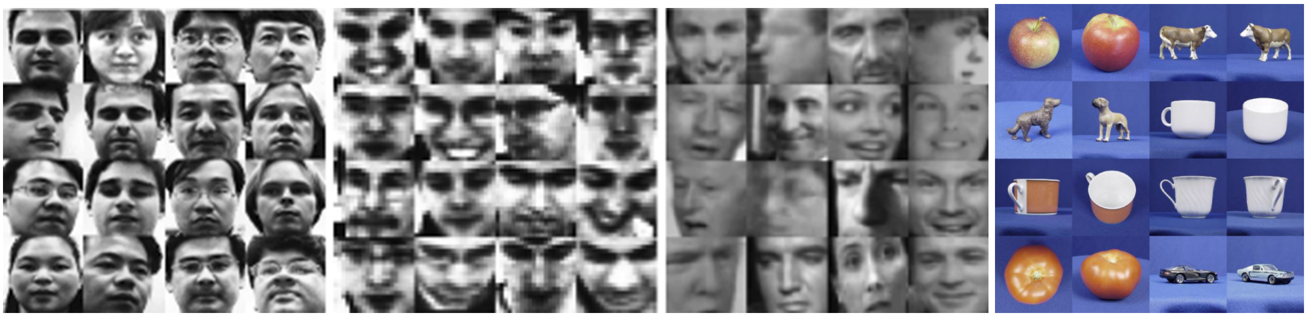
We evaluate the experimental performance of different methods on four widely used datasets, including Honda/UCSD (Honda) [14], CMU Mobo (Mobo) [14], YouTube Celebrities (YTC) [22], and ETH-80 [27]. Among these datasets, YTC is a large-scale video dataset collected under unconstrained real-life condition, whereas Honda, Mobo, and ETH-80 are relatively easy due to they are collected under controlled indoor lab environment. The brief descriptions and specific experimental settings of each of these datasets are presented in the following, and some samples from the four datasets are shown in Fig. 2.

The Honda/UCSD [14] dataset consists of 59 video sequences involving 20 different subjects. Each sequence contains approximately 12 to 645 frames covering large variations. We resize all images to 20 × 20; then, all video sequences are divided into two groups: 20 video sequences are used for the gallery (training) and the remaining 39 are used for the probe.

The CMU Mobo [14] dataset contains 96 video sequences of 24 individuals on a treadmill, which are captured from multiple cameras with four different walking situations, including slow, fast, inclined, and ball carrying. We resize all images to 30 × 30. Each video is further divided into 4 illumination sets: the first set for training and the remaining sets for testing.

The YouTube Celebrities [22] dataset contains 1910 YouTube videos of 47 celebrities (actors and politicians) collected from YouTube; most of the videos are of low resolution and highly compressed, which leads to noise, low-quality image frames. Each clip contains hundreds of frames. For this dataset, the obtained face images are resized to 30 × 30 grayscale images and then extract their LBP features. For each subject, we conduct experiments by randomly choosing 3 sets for training and 6 sets for testing.

The ETH-80 [27] dataset contains 8 categories, and each category contains 10 objects with 41 still RGB images of different views per object. For our experiments, we crop and resize the 128 × 128 grayscale images to 30 × 30. For each subject, we randomly choose 5 sets for training and the rest 5 sets for testing.



(a) Honda/UCSD dataset. **(b)** CMU Mobo dataset. **(c)** YTC dataset. **(d)** ETH-80 dataset.

Fig. 2 Sample images from **a** Honda, **b** Mobo, **c** YTC, **d** ETH-80. Note these datasets contain high intra-class variations in the form of different poses, illumination variations, expression deformations, and occlusions

For fairness and uniformity, in our experiments, we use the data matrix with MATLAB format provided by some scholars [15] instead of manual detecting face and extracting features.

4.2 Comparative methods

We compare our method against several excellent methods, including discriminant canonical correlation analysis (DCC) [16], manifold-to-manifold distance (MMD) [17], manifold discriminant analysis (MDA) [18], affine hull-based image set distance (AHISD) [14], convex hull-based image set distance (CHISD) [14], sparse approximated nearest points (SANP) [19], RH-ISCRC (i.e., ISCRC-l_1 and ISCRC-l_2) [2], regularized nearest points (RNP) [15], deep reconstruction models (DRM-MV) [27], pair-wise linear regression classification (PLRC) [21] and prototype discriminative learning (PDL) [22]. We adopt the implementations provided by the respective authors on their homepage for all compared methods, whose parameters are followed by the recommendations in the original references for best performance.

4.3 Experimental results

To evaluate the proposed method, we take face and object classification tasks as examples to evaluate the performances of different image set classification methods. For face biometric experiments, the size of image set N (i.e., the number of frames in a set) varies from $\{50, 100, 200\}$. Note that if the number of frames of a subject is less than the given set size N, all images are used for building a set. For object categorization experiments, the size of image set N is set to 41. In all experiments, we perform tenfold experiments by randomly selection of training and testing sets, and then report the average classification accuracies and standard deviations of all comparative methods. For all methods, the computing platform is Mac Mini 2018 with an Intel Core i7 (3.2GHz) CPU and 16-GB RAM, the operating system is macOS Mojave, and the simulation software is MATLAB 2016b. We set $\alpha = 0.001$ and $\alpha = 0.01$ for face biometric experiments and object categorization experiments, respectively.

In the face biometric experiments, some experiments with different set sizes N are conducted on Honda, Mobo, and YTC, respectively. For Honda, experimental results are shown in Table 1, from which we can see clearly that our NDENP performs best in all the 3 cases. Overall, all the previous methods achieve lower accuracies with small set size (i.e., $N = 50, 100$), while the frames in each image set are enough (i.e., $N = 200$), and good performances are achieved by all the methods. Delightedly, NDENP achieves 100% accuracy with all the set size (i.e., $N = 50, 100, 200$). Notably, compared to deep learning method DRM-MV (the second best), the improvements of NDENP are 3.1% and 0.7% when the set size is 50 and 100, respectively. For these methods, AHISD, CHISD, RNP, and SANP, which model image set as an affine hull or convex hull, show a relative enhancement compared with the linear subspace-based methods, such as DCC, MMD, and MDA. Moreover, our NDENP performs higher than ISCRC-l_1 (SR-based) and ISCRC-l_2 (CRC-based) by 10.3% and 2.6%, respectively. For Mobo, experimental

Table 1 Average classification accuracies (%) of NDENP versus different state-of-the-art methods on Honda/UCSD dataset

Methods	Honda/UCSD		
	50 frames	100 frames	200 frames
DCC	74.3	87.2	94.9
MMD	69.2	89.7	94.9
MDA	82.0	94.9	97.4
AHISD	82.0	84.6	89.7
CHISD	82.0	87.2	92.3
SANP	84.6	92.3	94.9
ISCRC-l_1	89.7	97.4	**100.0**
ISCRC-l_2	89.7	97.4	**100.0**
RNP	87.2	94.9	97.4
DRM-MV	96.9	99.3	**100.0**
PLRC	87.2	97.4	**100.0**
PDL	87.2	94.9	97.4
NDENP	**100.0**	**100.0**	**100.0**

The bold numbers are the highest accuracies

Table 2 Average classification accuracies (%) with standard deviation of NDENP versus different state-of-the-art methods on CMU Mobo dataset

Methods	CMU Mobo		
	50 frames	100 frames	200 frames
DCC	82.1 ± 2.7	85.5 ± 2.8	91.6 ± 2.5
MMD	86.2 ± 2.9	94.6 ± 1.9	96.4 ± 0.7
MDA	86.2 ± 2.9	93.2 ± 2.8	95.8 ± 2.3
AHISD	91.6 ± 2.8	94.1 ± 2.0	91.9 ± 2.6
CHISD	91.2 ± 3.1	93.8 ± 2.5	96.0 ± 1.3
SANP	91.9 ± 2.7	94.2 ± 2.1	97.3 ± 1.3
ISCRC-l_1	93.5 ± 2.8	96.5 ± 1.9	98.7 ± 1.7
ISCRC-l_2	93.5 ± 2.8	96.4 ± 1.9	98.4 ± 1.7
RNP	91.9 ± 2.5	94.7 ± 1.2	97.4 ± 1.5
DRM-MV	92.9 ± 1.7	96.2 ± 0.9	98.1 ± 0.8
PLRC	92.1 ± 1.6	94.6 ± 1.9	97.5 ± 1.8
PDL	92.5 ± 2.3	94.8 ± 1.9	96.6 ± 2.6
NDENP	**94.6** ± **1.1**	**97.3** ± **0.7**	**99.1** ± **0.9**

The bold numbers are the highest accuracies results are shown in Table 2, we achieved the maximum accuracy of 94.6%, 97.3%, and 99.1% when set size is 50, 100, and 200, respectively, which is the highest reported so far. Compared to ISCRC-l_1 and ISCRC-l_2, the average accuracy improvements of NDENP are 0.8% and 0.9%, respectively. When the set size is 50, the classification accuracies of DCC, MMD, and MDA are lower than 90%, which suggests that the discriminative information extraction and manifold analysis depend largely on large number of samples per image set. For YTC, experimental results are shown in Table 3, and similar conclusions to those methods on the previous two datasets could be reached. Overall, the recognition rates of all methods on this dataset are significantly lower than those on the previous two datasets. This is because the dataset is collected under unconstrained conditions with many tracking and detection errors. However, NDENP has better performance than all the comparative methods. In particular, the l_1-norm-based baseline method ISCRC-l_1 achieves 62.3%, 65.6%, and 66.7% when the set size is 50, 100, and 200, respectively, and the l_2-norm-based baseline method ISCRC-l_2 achieves 57.4%, 60.7%, and 61.4% when the set size is 50, 100, and 200, respectively. As a consequence, ISCRC-l_1 leads to much better results than ISCRC-l_2 on this challenging dataset because l_1-norm is very helpful to improve the discrimination and stability. By contrast, our method performs higher than ISCRC-l_1 by 4.1%, 3.7%, and 4.7% when the set size is 50, 100, and 200, respectively.

In the object categorization experiments, the experimental results are summarized in Table 4. As we know, ETH-80 is also more challenging since it has much less images per set, significant appearance differences across subjects of the same class, and larger view angle variations within each image set. On this dataset, our method again outperforms the vast majority of existing methods by exhibiting 90.5% accuracy and 2.1% standard deviation. The deep learning method DRM-MV has the highest 91.6% accuracy, and the PLRC method achieves 89.5% accuracy as the third best one. Nevertheless, the accuracy of ISCRC-l_1 and ISCRC-l_2 is lower than 80%, which are only 79.5% and 78.4%, respectively.

Table 3 Average classification accuracies (%) with standard deviation of NDENP versus different state-of-the-art methods on YouTube Celebrities dataset

Methods	YouTube Celebrities		
	50 frames	100 frames	200 frames
DCC	57.2 ± 7.1	61.8 ± 6.7	65.6 ± 6.4
MMD	57.4 ± 6.5	62.7 ± 8.7	64.1 ± 8.3
MDA	58.2 ± 8.7	63.1 ± 6.5	66.1 ± 6.2
AHISD	57.1 ± 8.1	59.7 ± 6.4	57.1 ± 8.1
CHISD	57.9 ± 6.8	62.7 ± 7.2	64.2 ± 7.5
SANP	56.7 ± 5.5	61.9 ± 8.1	65.4 ± 6.8
ISCRC-l_1	62.3 ± 6.2	65.6 ± 6.7	66.7 ± 6.4
ISCRC-l_2	57.4 ± 7.2	60.7 ± 6.5	61.4 ± 6.4
RNP	59.9 ± 7.3	63.3 ± 8.1	64.4 ± 7.8
DRM-MV	62.3 ± 5.5	68.2 ± 6.2	70.3 ± 4.8
PLRC	61.7 ± 8.2	65.6 ± 7.9	66.8 ± 7.5
PDL	63.9 ± 6.8	65.7 ± 7.7	67.1 ± 7.6
NDENP	**66.4** ± **6.5**	**69.3** ± **7.1**	**71.4** ± **6.9**

The bold numbers are the highest accuracies

Table 4 Average classification accuracies (%) with standard deviation of NDENP versus different state-of-the-art methods on ETH-80 dataset

Methods	DCC	MMD	MDA	AHISD
Rates	86.0 ± 6.5	77.5 ± 5.0	77.3 ± 5.5	78.8 ± 5.3
CHISD	SANP	ISCRC-l_1	ISCRC-l_2	RNP
79.5 ± 5.3	77.8 ± 7.3	79.5 ± 4.5	78.4 ± 3.8	81.0 ± 3.2
DRM-MV	PLRC	PDL	NDENP	
91.6 ± 3.7	89.5 ± 4.6	89.2 ± 5.8	90.5 ± 2.1	

The bold numbers are the highest accuracies

4.4 Computational time analysis

We can categorize the evaluated methods as online methods (which do all computations at run time) and off-line methods (which do the training component off-line and only testing is done at run time). The online methods contain MMD, AHISD, CHISD, SANP, ISCRC-l_1, ISCRC-l_2, RNP, PLRC, and NDENP. The off-line methods contain DCC, MDA, and DRM.

Then, we evaluate the main computational complexity (MCC) of the recently proposed online methods. Let n_a and n_b be the number of samples in query set and gallery sets, respectively, and let t be the iteration number. The MCC of sparse representation is $\mathcal{O}(d^2 n^{\varepsilon})$, where d is the feature dimension, n is the number of samples, and $\varepsilon \geq 1.2$. Specifically, the MCC of ISCRC-l_1 is $\mathcal{O}(td^2(n_a^{\varepsilon} + n_b^{\varepsilon}))$, the MCC of ISCRC-$l_2$ is $\mathcal{O}(t(n_a + n_b)^3)$, the MCC of SANP is $\mathcal{O}_{\mathrm{svd}} + \mathcal{O}(t(d^2(n_a + n_b)^{\varepsilon}))$ (where $\mathcal{O}_{\mathrm{svd}}$ denotes the time complexity of SVD), the MCC of RNP is $\mathcal{O}_{\mathrm{svd}} + \mathcal{O}(td(n_a + n_b))$, the MCC of PLRC is $\mathcal{O}((n_a + n_b)^3)$, the MCC of PDL is $\mathcal{O}(tn_b(2dn_b^2 + 2n_b + 2m + 2mn_b))$ (where m is the projection dimension), and the MCC of the proposed NDENP is $\mathcal{O}(t(n_a^3 + n_b^3))$. Therefore, the MCC of our method nearly equals ISCRC-l_2's when t is fixed.

Experimentally, we compare the run time of different methods on YTC dataset. Time in seconds required for one

Table 5 Time cost of different methods on YTC (50 images per set) for training and testing

Methods	DCC	MMD	MDA	AHISD
Time	13.17	14.42	5.31	0.92
CHISD	SANP	ISCRC-l_1	ISCRC-l_2	RNP
1.64	8.55	0.08	0.04	0.22
DRM-MV	PLRC	PDL	NDENP	
376.44	0.21	62.54	0.12	

subject (50 images per set), respectively, are listed in Table 5. The results in Table 5 suggest that our method is almost much faster than other methods except ISCRC-l_1 and ISCRC-l_2. Note that the computational time of the deep learning method DRM far exceeds other methods up to 376.44, but our method only requires 0.12 s in total.

4.5 Convergence and parameter sensitivity

It can be verified that the model of NDENP in (1) is convex and differentiable. Accordingly, we have derived closed-form solutions for all variables to efficiently solve our object function, and the convergence to the global optimum can be guaranteed. We show typical convergence curves in Fig. 3 on Honda, Mobo, YTC, and ETH-80 datasets, respectively. We can see that our method converges fast (within 15 iterations) with the primal residuals quickly reduced close to zero. To greatly accelerate convergence, we can use a relatively large μ in the ADMM, which gives rise to a large penalty parameter ρ in a few iterations.

Furthermore, we consider the influence of the parameter α on the classification performance. The relationship between recognition performance and parameter α is shown in Fig. 4. The candidate parameter λ varies from the coarse set $\{1, 0.1, 0.01, 0.001, 0\}$. We can see that the highest recognition rates are achieved at $\alpha = 0.001$ for face set classification, and $\alpha = 0.01$ for object set classification. This result verifies that the optimal α obtained by our method is able to boost classification performance. In addition, it is noticeable that when $\alpha = 0$, the recognition rates are always less than having an optimal α. This shows that our discriminative term has practical effects to further boost performance.

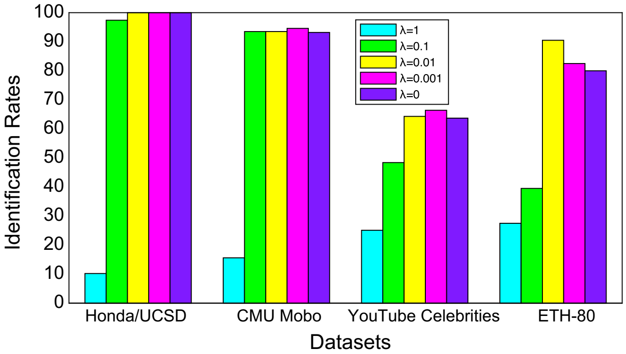
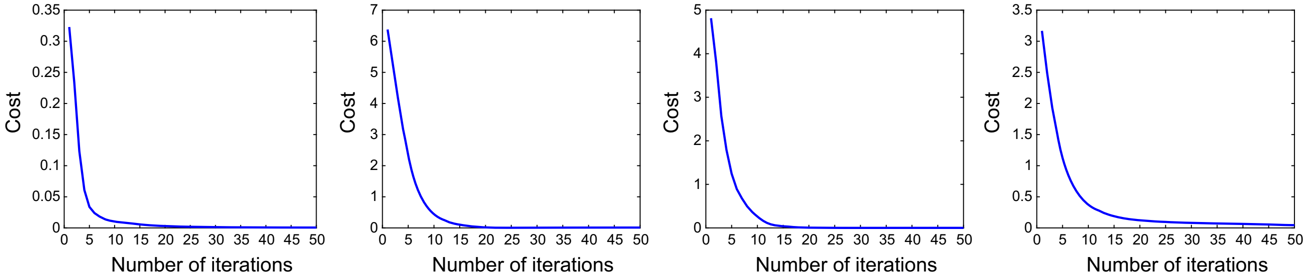


Fig. 4 Variations of recognition rates versus the parameter λ on the Honda, Mobo, YTC, and ETH-80 databases, respectively

4.6 Analysis and discussion

Taken together, the experiments demonstrate that our method achieves better performances than state-of-the-art methods. The main reasons are that: (1) the code coefficient of the gallery set is nonnegative, which can bring discriminability and sparsity simultaneously; (2) the code coefficient of the probe set is also nonnegative, which can bring sparsity and select the samples with high intra-class variation; (3) the class-wise discriminative term can further boost the discriminant ability of different gallery sets. In order to better demonstrate the power of the learned coding coefficients, we perform a small-scale experiment and then visually evaluate the coding coefficients and reconstructed image representations using some illustrations. In this experiment, we first select the first video clip of the first 20 individuals for training (i.e., gallery sets $\mathbf{X}$) and the second video clip of the 12th individual for testing (i.e., probe set $\mathbf{Y}$), and each clip has 50 random selected frames. We then compress the gallery sets $\mathbf{X}$ into a more compact dictionary



(a) Honda. **(b)** Mobo. **(c)** YTC. **(d)** ETH-80.

Fig. 3 Convergence of the optimization algorithm on **a** Honda/UCSD, **b** CMU Mobo, **c** YTC, **d** ETH-80

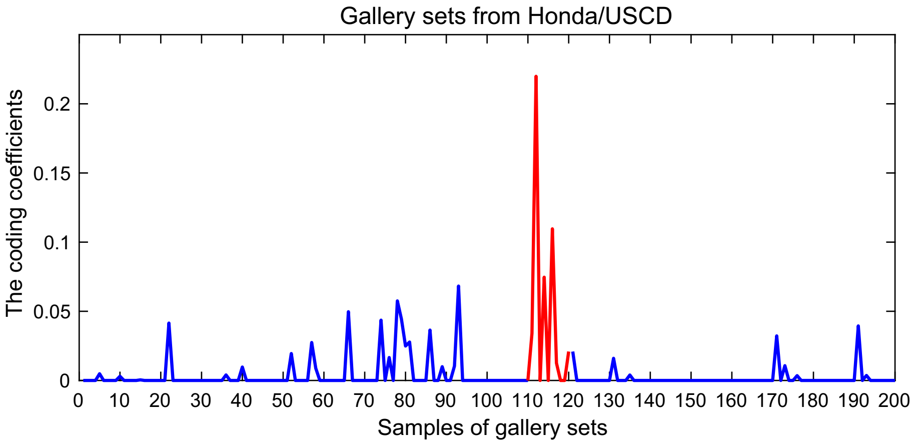


Fig. 5 Coding coefficient of the gallery sets. There are 20 classes in total, and the number of dictionary atoms for each class is set to 10. Red coefficients correspond to training images of the correct individual (color figure online)

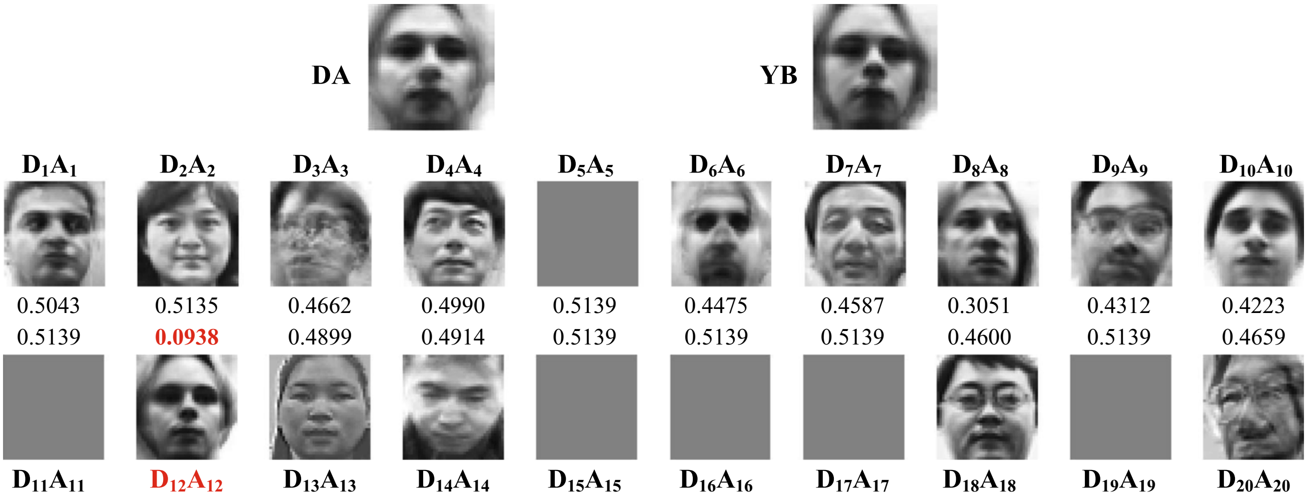
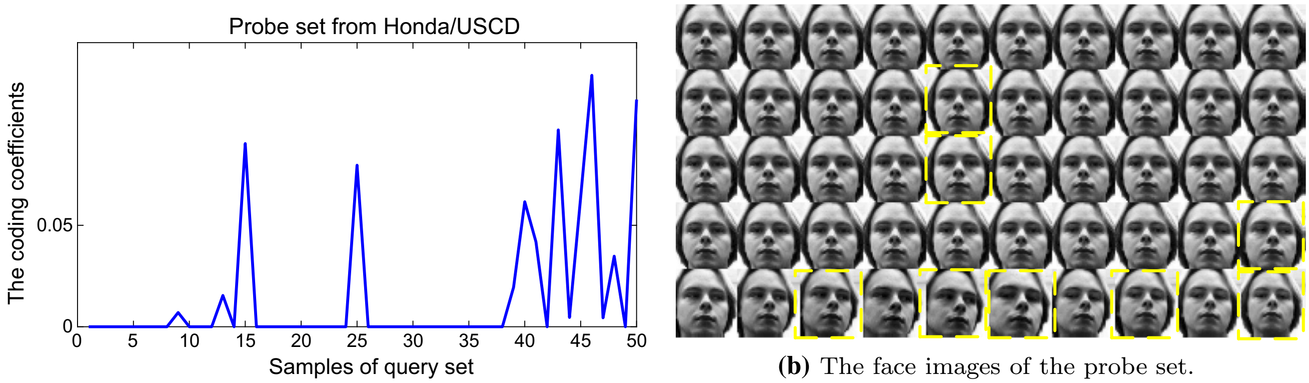


Fig. 6 Reconstructed face images and their corresponding reconstruction residuals. The first row shows the nearest point images of the whole gallery sets and the probe set, and the second and third rows show the nearest point images of each gallery set

D with 10 samples/atoms by using the metaface learning method [34]. The coding coefficient of the gallery sets and the class-wise reconstructed face images with reconstruction residuals are shown in Figs. 5 and 6, respectively. For simplicity, the reconstruction residual (14) is predigested as $\|\mathbf{X}_i\mathbf{A}_i - \mathbf{YB}\|_2^2$. It is obvious that the coefficients associated with the gallery set $\mathbf{D}_{12}$ (the red part) are larger than the other gallery sets, resulting in a smaller reconstruction residual 0.0938. In other words, the nearest points of the query set and the gallery sets belonging to the same individual are the most similar and have the minimum corresponding distance. Moreover, the coding coefficient in Fig. 5 is indeed sparse, involving only a small fraction of the overall training samples, which demonstrates that the nonnegative constrain can bring discriminability and sparsity simultaneously. The coding coefficient and samples of the probe set are shown in Fig. 7; it is easy to see that the learned coefficients are also sparse and have large magnitude only for images with high intra-class variations. The power of our proposed class-wise discriminative term is proved to be useful in boosting classification accuracy (i.e., when λ is set to 0, the accuracy is limited) (see Sect. 4.5).

5 Conclusion and future works

In this paper, we proposed a novel method NDENP for image set classification. The NDENP introduces two nonnegative constraints instead of the l_1 norm so that the learned coding coefficients can improve the discriminative ability between different image sets while maintaining sparsity. In addition, in order to use the label information to further boost the set discriminant ability, the NDENP integrates a class-wise discriminative term on the nearest points of the gallery sets, which can simultaneously minimize the nearest points distance between the whole gallery sets and the probe set, as well as the nearest points



(a) The coding coefficient of the probe set.

(b) The face images of the probe set.

Fig. 7 Face images of the gallery set from Honda and the corresponding coding coefficient. **a** Coding coefficient of the probe set, **b** probe faces (the selected face images are highlighted with a yellow bounding box). Note the only few samples are dominantly involved in representation, which are with high intra-class variations in the form of different head poses, illumination variations, and expression deformations (color figure online)

correlations of the gallery sets. Therefore, the learned coding coefficients bring better discriminability and sparsity simultaneously, which is benefit to set classification. Experiments on four widely used benchmark datasets prove that the proposed NDENP significantly outperforms the state-of-the-art image set classification methods in terms of both classification accuracy and computational efficiency.

Although the proposed method has achieved promising results, there are still some aspects that deserve study in the future. Firstly, we will introduce the affine hull or convex hull as our image set model. Further, we will explore the effect of learning the correlation of coefficients to further improve classification performance.

Acknowledgements We would like to thank Pengfei Zhu, who shared the source code of RH-ISCRC on their homepage, and to thank Meng Yang, who shared the RNP code. We would also like to thank all the anonymous reviewers who provided substantive suggestions for improving our work.

Funding This research was supported by the National Natural Science Foundation of China (Grant No. 61673220), the Department of Science and Technology of Sichuan Province (No. ZYF-2018-106), the Sichuan Province Science and Technology Support Program (No. 2018TZDZX0002), and the State Administration for Science, Technology and Industry for National Defense (Nos. JCKY2017209B010 and JCKY2018209B001).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Chen Z, Jiang B, Tang J, Luo B (2017) Image set representation and classification with attributed covariate-relation graph model and graph sparse representation classification. Neurocomputing 226:262–268
2. Zhu P, Zuo W, Zhang L, Shiu SCK, Zhang D (2014) Image set-based collaborative representation for face recognition. IEEE Trans Inf Forensics Secur 9(7):1120–1132
3. Shao M, Tang D, Liu Y, Kim TK (2016) A comparative study of video-based object recognition from an egocentric viewpoint. Neurocomputing 171:982–990
4. Moon HM, Seo CH, Pan SB (2017) A face recognition system based on convolution neural network using multiple distance face. Soft Comput 21(17):4995–5002
5. Wang G, Shi N (2019) Collaborative representation-based discriminant neighborhood projections for face recognition. Neural Comput Appl. https://doi.org/10.1007/s00521-019-04055-6
6. Chen L, Hassanpour N (2017) Survey: How good are the current advances in image set based face identification? Experiments on three popular benchmarks with a naïve approach. Comput Vis Image Underst 160:1–23
7. Ren Z, Wu B, Sun Q, Wu M (2019) Simultaneous learning of reduced prototypes and local metric for image set classification. Expert Syst Appl 134:102–111
8. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31(2):210–227
9. Xu J, An W, Zhang L, Zhang D (2019) Sparse, collaborative, or nonnegative representation: which helps pattern classification? Pattern Recognit 88:679–688
10. Tang D, Zhu N, Yu F, Chen W, Tang T (2014) A novel sparse representation method based on virtual samples for face recognition. Neural Comput Appl 24(3–4):513–519
11. Zeng S, Gou J, Yang X (2018) Improving sparsity of coefficients for robust sparse and collaborative representation-based image classification. Neural Comput Appl 30(10):2965–2978
12. Zhang H, Wang S, Xu X, Chow TW, Wu QJ (2018) Tree2vector: learning a vectorial representation for tree-structured data. IEEE Trans Neural Netw Learn Syst 99:1–15
13. Hua J, Wang H, Ren M, Huang H (2017) Collaborative representation analysis methods for feature extraction. Neural Comput Appl 28(1):225–231
14. Cevikalp H, Triggs B (2010) Face recognition based on image sets. In: IEEE computer society conference on computer vision and pattern recognition. IEEE, pp 2567–2573
15. Yang M, Zhu P, Van Gool L, Zhang L (2013) Face recognition based on regularized nearest points between image sets. In: 10th

IEEE international conference and workshops on automatic face and gesture recognition (FG). IEEE, pp 1–7

16. Kim TK, Kittler J, Cipolla R (2007) Discriminative learning and recognition of image set classes using canonical correlations. IEEE Trans Pattern Anal Mach Intell 29(6):1005–1018
17. Wang R, Shan S, Chen X, Gao W (2008) Manifold-manifold distance with application to face recognition based on image set. In: IEEE conference on computer vision and pattern recognition, CVPR 2008. IEEE, pp 1–8
18. Hamm J, Lee DD (2008) Grassmann discriminant analysis: a unifying view on subspace-based learning. In: Proceedings of the 25th international conference on machine learning. ACM, pp 376–383
19. Hu Y, Mian AS, Owens R (2012) Face recognition using sparse approximated nearest points between image sets. IEEE Trans Pattern Anal Mach Intell 34(10):1992–2004
20. Ren Z, Wu B, Zhang X, Sun Q (2019) Image set classification using candidate sets selection and improved reverse training. Neurocomputing 341:60–69
21. Feng Q, Zhou Y, Lan R (2016) Pairwise linear regression classification for image set retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4865–4872
22. Wang W, Wang R, Shan S, Chen X (2017) Prototype discriminative learning for image set classification. IEEE Signal Process Lett 24(9):1318–1322
23. Wang W, Wang R, Shan S, Chen X (2015) Probabilistic nearest neighbor search for robust classification of face image sets. In: 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), vol 1. IEEE, pp 1–7
24. Shah SA, Nadeem U, Bennamoun M, Sohel F, Togneri R (2017) Efficient image set classification using linear regression based image reconstruction. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 99–108
25. Song Z, Cui K, Cheng G (2019) Image set face recognition based on extended low rank recovery and collaborative representation. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-019-00941-6
26. Shah SAA, Bennamoun M, Boussaid F (2016) Iterative deep learning for image set based face and object recognition. Neurocomputing 174:866–874
27. Hayat M, Bennamoun M, An S (2015) Deep reconstruction models for image set classification. IEEE Trans Pattern Anal Mach Intell 37(4):713–727
28. Uzair M, Shafait F, Ghanem B, Mian A (2018) Representation learning with deep extreme learning machines for efficient image set classification. Neural Comput Appl 30(4):1211–1223
29. Yang M, Wang X, Liu W, Shen L (2017) Joint regularized nearest points for image set based face recognition. Image Vis Comput 58:47–60
30. Shao J, Huang F, Yang Q, Luo G (2018) Robust prototype-based learning on data streams. IEEE Trans Knowl Data Eng 30(5):978–991
31. Liu W, Xu D, Tsang IW, Zhang W (2019) Metric learning for multi-output tasks. IEEE Trans Pattern Anal Mach Intell 41(2):408–422
32. Leng M, Moutafis P, Kakadiaris IA (2015) Joint prototype and metric learning for set-to-set matching: application to biometrics. In: IEEE 7th international conference on biometrics theory, applications and systems (BTAS). IEEE, pp 1–8
33. Xu Y, Li Z, Yang J, Zhang D (2017) A survey of dictionary learning algorithms for face recognition. IEEE Access 5:8502–8514
34. Yang M, Zhang L, Yang J, Zhang D (2010) Metaface learning for sparse representation based face recognition. In: 2010 IEEE international conference on image processing. IEEE, pp 1601–1604

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.