Neural Computing and Applications (2023) 35:12283–12295
https://doi.org/10.1007/s00521-020-05479-1

S.I. : NEW TRENDS OF NEURAL COMPUTING FOR ADVANCED APPLICATIONS

Face image set classification with self-weighted latent sparse discriminative learning

Yuan Sun[1] · Zhenwen Ren[1,2] · Chao Yang[1] · Quansen Sun[2] · Liwan Chen[3] · Yanglong Ou[1]

Received: 1 September 2020 / Accepted: 26 October 2020 / Published online: 21 November 2020

Abstract
Since image set classification has strong power to overcome various variations in illumination, expression, pose, and so on, it has drawn extensive attention in recent years. Noteworthily, the point-to-point distance-based methods have achieved the promising performance, which aim to compute the similarity between each gallery set and the probe set for classification purpose. Nevertheless, these existing methods have to face the following problems: (1) they do not take full advantage of the between-set discrimination information; (2) they ideally presume that the importance of different gallery sets is equal, whereas this always violates objective facts and may degenerate algorithm performance in practice; (3) they tend to have high computational cost and several parameters, though explicit sparsity can enhance discrimination. To address these problems, we propose a novel method for face image set classification, namely *self-weighted latent sparse discriminative learning* (SLSDL). Specifically, a novel self-weighted strategy guided discrimination term is proposed to largely boost the discrimination of different gallery sets, such that the effect of true sets can be boosted while the effect of false sets can be weakened or removed. Moreover, we propose a latent sparse normalization to reduce computational complexity as well as the number of trade-off parameters. In addition, we propose an efficient optimization algorithm to solve the final SLSDL. Comprehensive experiments on four public benchmark datasets demonstrate that SLSDL is superior to the state-of-the-art competitors.

Keywords Image set classification · Face recognition · Video-based face classification · Sparse representation

1 Introduction

With the rapid improvement in computer vision, it is greatly easy to collect multiple faces images in real-life application [1]. Hence, image set model becomes one of the research hotspots in recent years, where each set consists of multiple images corresponding to a subject. Theoretically, image set can utilize more face information of the same subject to effectively decrease the impact of a mount of appearance variations about expression, pose, illumination, and so on [2, 3], thereby enhancing the robustness and discriminant ability for classification purpose. Accordingly, image set classification aims to model an image set by leveraging appropriate technologies and then calculate the similarity between the probe image set and each gallery set to accomplish classification using some distance criterion [2, 4–7]. Obviously, image set classification is very different from traditional single face image classification [1, 7–9]. What's more, since image set can provide us more comprehensive information to describe faces than single image, naturally, it has obtained more promising performance than the traditional classification methods under single face image in real-world applications. However, due to different variations, it is still

✉ Zhenwen Ren
rzw@njust.edu.cn

✉ Chao Yang
ychao1983@126.com

[1] Department of National Defence Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China

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

[3] Department of Electronic and Information Engineering, Chongqing Three Gorges University, Chongqing 404199, China

a great challenge to correctly classify each image set by effectively exploiting the discriminative information [10–14]. For instance, we choose four training sets from the YouTube Celebrities dataset and illustrate the projection map of them on the top three principal directions [6], which is shown in Fig. 1. As we can see, the data distribution does not follow the class labels and is incomplete concentrated. In this case, it is a challenge that image set classification effectively improves classification performance.

Over the last decade, plentiful image set classification methods [2, 4, 6, 15–23] have been proposed to address two key problems, which largely determine the performance of face recognition: constructing model of image set and defining appropriate distance criterion to calculate similarity between a pair of image sets. Based on plenty of literature review, it is well understood that the point-to-point distance-based methods have gained the most widely attention among all the existing methods [24, 25]. These methods utilize nearest points distance as image sets classifiers to obtain between-set similarity for image set classification. However, these existing methods have to face the following problems: (1) they do not make full use of the rich discriminant information between the whole gallery set and each gallery set to encode image sets; (2) most of the existing methods follow an ideal assumption that the importance of different gallery sets is equal, which is not adapted in practical applications; and (3) they usually introduce the sparse coding coefficients to enhance the representation capability of the homogeneous images and reduce that of inhomogeneous sets images. In short, the first two ones hinder the performance breakthrough, and the latter brings out computational burden and multi-parameters.

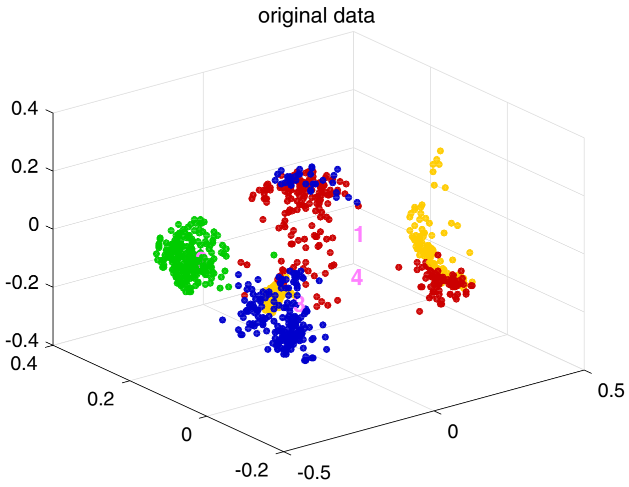


Fig. 1 Choosing four training sets from the YouTube Celebrities dataset. Projection map of them on the top three principal directions is drawn in four different colors dots

To address the above drawbacks of point-to-point distance-based methods, this paper presents an efficient and effective face image set classification method, termed *self-weighted latent sparse discriminative learning* (SLSDL). In summary, our SLSDL has the following strengths and contributions:

- To make full use of the encoding information between different gallery sets and improve discrimination of the representation model, it simultaneously minimizes the nearest distance between the probe set and the whole gallery set, and the distances between each independent gallery set and the whole gallery set.
- To find a more precise nearest point in the gallery sets, it proposes a self-weighted strategy to control the contribution of each gallery set, such that each gallery set can be treated differently to improve its discriminant ability.
- It proposes a latent sparse normalization with capped simplex constraint to approximate the sparse constraint term for reducing the involved trade-off parameters and computational complexity; meanwhile, it can preserve the robustness of sparse representation.
- It obtains excellent performance by experimenting on four public benchmark datasets.

2 Related work

Recently, an excellent point-to-point distance model is attracted extensive attention from scholars, which calculates distance of the between-set points to obtain between-set similarity. Up to now, numbers of typical methods have been proposed one after another [2, 3, 6, 15–23, 26]. For instance, affine hull-based image set distance (AHISD) [15] and convex hull-based image set distance (CHISD) [15] are proposed to compute distance between two affine/convex hulls of image sets and then obtain the similarity of image sets. However, this can lead to performance degradation because of the overlarge of affine/convex hulls. For getting rid of the problem, sparse approximated nearest points (SANP) [16] put forward the sparsity constraint on sample coefficients and achieve better accuracies. However, SANP can bring into high computational complexity; hence, regularized nearest points (RNP) [17] are proposed to model the image set as regularized affine hull for more concise and lower time consumption, whereas these methods [27, 28] only consider between-set relation in the training phase and ignore discriminative information between a probe set and the whole gallery set. Nowadays, more and more methods based on collaborative representation [29] have been proposed for overcoming this issue as follows: regularized hull based collaborative representation (RH-ISCRC) [18], collaborative regularized nearest points (CNRP) [19], and joint regularized nearest points (JRNP) [22]. Furthermore, prototype discriminative learning (PDL) [20] also achieves nice performance by simultaneously learning the prototype set and a linear projection matrix, which keeps a prototype set from a same class closer and different classes farther. Based on this idea, dual linear regression classification (DLRC) [30] just focuses on the related class subspace. As a development of DLRC, pairwise linear regression model (PLRC) [21] considers a new unrelated subspace as well. On the basis of PLRC, the recently proposed discriminative residual analysis (DAR) [31] obtains discriminant features and then projects the gallery set and probe set into the discriminant subspace for improving the classification performance. Recently, deep learning (e.g., DRM-MV) [32] is also gradually applied to image set classification tasks. Though there are some breakthroughs in these approaches, they have still some defects to defeat (see Sect. 1).

In general, the relationship between our SLSDL and the related methods is that both of them are point-to-point distance model. The difference is that the existing point-to-point methods unwittingly ignore the important discrimination information between each gallery set and whole gallery set, and usually introduce explicitly several sparse coding coefficients to enhance discrimination and robustness. Moreover, these methods ideally assume that the importance of different gallery sets is equal, which is not adapted in practical applications. Conversely, our SLSDL utilizes the weighted value to adjust precisely the coding coefficient errors and further enhance the correlation between the whole gallery set and the ground truth gallery set and the repellency between the heterogeneous sets. For SANP [16], RNP [17], and CRNP [19], in a broad sense, the weighted value between the probe set and each gallery set is set to 1, and they consider each gallery to be equally important. Meanwhile, SLSDL also proposes a latent sparse normalization for representation sparsity, thereby reducing the number of trade-off parameters and running time. Note here that *latent sparse discriminative learning* (LSDL) [33] is a conference version of this paper, which treats each different gallery set indistinguishably, i.e., all weighted values of gallery sets are fixed to 1.

3 Proposed method

In this section, we present the objective function, the optimization algorithm, and the classification criterion of the proposed SLSDL method.

3.1 Problem formulation

Point-to-point image set classification method models each image set as virtual point, and calculate the similarity between each gallery set and the probe set[1] by appropriate distance criterion. Mathematically, the whole gallery sets are denoted as $\boldsymbol{X} = [\boldsymbol{X_1}, \ldots, \boldsymbol{X_m}]$, which consisted of m different classes, and the probe set is denoted as $\boldsymbol{Y}$. The size of each image set is presented as d. Inspired by [18, 34], it can thus be known that sparse representation can produce sparsity to enhance robustness, thereby resisting noise and even outliers. According to the above idea, we assume that the virtual point $\boldsymbol{X\alpha}$ of whole gallery sets is not only close to the virtual point $\boldsymbol{Y\beta}$ of the probe set, but also close to that of ground truth gallery set (i.e., $\{\boldsymbol{X}_i\gamma_i\}_{i=1}^m$). $\boldsymbol{Y\beta}$ can produce a better approximation of $\boldsymbol{X\alpha}$ by collaborative representation and adjust coding errors. However, the idea is too ideal because the ground truth gallery set corresponded to label is unknown. In summary, our ideas can be summarized as follows: (1) To measure the similarity between the probe set and the gallery ones, we model the query set as a regularized hull collaboratively over all the gallery sets. (2) In order to enhance the discriminative power of all gallery sets, we propose a discrimination term to minimize the distance between the whole gallery set and an unknown ground truth gallery set. (3) To obtain sparse solutions to enhance robustness, we introduce the sparsity constraint on coding coefficients to suppress the contributions of the unnecessary samples. Further, sparsity constraint can enhance the representation capability of the homogeneous image sets and reduce that of inhomogeneous image sets. Regarding the mentioned above, our model can be written as follows

$$\begin{aligned}\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}} & \|\boldsymbol{X\alpha} - \boldsymbol{Y\beta}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_1 \\ & + \lambda\sum_{i=1}^{m}\|\boldsymbol{X\alpha} - \boldsymbol{X}_i\gamma_i\|_2^2 + \lambda_3\|\gamma_i\|_1\end{aligned} \tag{1}$$

where λ, λ_1, λ_2, and λ_3 are the trade-off parameters, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the sparse coefficients, and $\gamma_i = [\gamma_{i1}^T, \gamma_{i2}^T, \ldots, \gamma_{im}^T]^T$ is the sparse coefficient correspond to the ith gallery set. In summary, the first term with collaborative representation minimizes the distance between the virtual points of a pair of sets, which are the probe set and the whole gallery set. The fourth term aims to enhance the discriminative power of all gallery sets. By minimizing the distance between the

[1] For image sets classification tasks [16], image set can be divided into two parts, including the gallery set and the probe set, where the gallery set with label is used to train and the probe set is used to test. Usually, each set may contain large variations in pose, illumination, and scale. Images in each set contained a same subject are collected from video-based face recognition systems, multiple cameras, or personal photo album.

virtual points of the whole gallery set and an unknown ground truth gallery set to adjust coding errors, the probe set can obtain a better approximation. In addition, the rest terms are constrained by the l_1 norm to obtain sparse solutions (Fig. 2).

However, problem (1) has the following disadvantages. Firstly, the explicit sparsity inevitably results in high computational complexity; moreover, three parameters, λ_1, λ_2, and λ_3, need to be tuned, whereas it is very not user-friendly for tuning these parameters. To overcome this disadvantage, we normalize $\boldsymbol{1}^T\boldsymbol{\alpha} = 1$, $\boldsymbol{1}^T\boldsymbol{\beta} = 1$, and $\boldsymbol{1}^T\boldsymbol{\gamma}_i = 1$ to make the sparsity terms to ablate. Naturally, it can produce the sparsity on the coefficients because the normalizations can choose a few samples in probe set to approximate the samples of probe set. That is, the normalizations $\boldsymbol{1}^T\boldsymbol{\alpha} = 1$, $\boldsymbol{1}^T\boldsymbol{\beta} = 1$, and $\boldsymbol{1}^T\boldsymbol{\gamma}_i = 1$ are equivalent to the latent sparse constraints on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}_i$, respectively; moreover, the normalizations can enhance the discriminability by finding the homogeneous set in the gallery sets to represent the probe set. By doing so, we can obtain the sparse and the discriminated solutions simultaneously. Secondly, we can see that the used spare constraints may produce some negative coefficients because of no non-negative constraint on the coding coefficient vectors. Obviously, the non-negative coefficients are mathematically explainable and physically meaningful, and the constructed virtual face images are more likely to be real faces. Therefore, we introduce some non-negative constraints on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}_i$ simultaneously, i.e., $\boldsymbol{\alpha} \geq 0$, $\boldsymbol{\beta} \geq 0$, and $\boldsymbol{\gamma}_i \geq 0$. Finally, it is not reasonable that each gallery set is treated equally. To overcome this disadvantage, a new self-weighted strategy is introduced to update the weight value between the whole gallery set and each gallery set iteratively. Therefore, the weight value can control the contribution of each gallery set, which makes explicitly the whole gallery virtual point closer to the nearest point of estimated correct class. Ideally, the correct gallery set has a larger weighted value, and the incorrect sets have the lower weighted values. Regarding the discussions mentioned above, problem (1) can be upgraded to

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{w}} & \|\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{Y}\boldsymbol{\beta}\|_2^2 \\ & + \lambda \sum_{i=1}^{m} w_i \|\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{X}_i\boldsymbol{\gamma}_i\|_2^2 \\ & \text{s.t. } \boldsymbol{\alpha} \geq 0, \boldsymbol{1}^T\boldsymbol{\alpha} = 1, \boldsymbol{\beta} \geq 0, \boldsymbol{1}^T\boldsymbol{\beta} = 1, \\ & \quad \gamma_i \geq 0, \boldsymbol{1}^T\gamma_i = 1, \boldsymbol{w} \geq 0, \boldsymbol{1}^T\boldsymbol{w} = 1 \end{aligned} \tag{2}$$

where $\boldsymbol{w} = [w_1, \ldots, w_m]$ is the weighted vector and λ is the trade-off parameter. Note that each set $\boldsymbol{X}_i$ could be replaced with a dictionary with k atoms, such as KSVD [18].

Compared to the existing methods, our method has a concise formulation and latent sparsity. Note here that only one parameter λ. Concretely, our SLSDL method has the following benefits over the regular ones: (1) it minimizes the distance between the probe set and the whole gallery set by using collaborative representation; meanwhile, it proposes a novel discrimination term to minimize the distance between the whole gallery set and an unknown ground truth gallery set. In the objective function of SLSDL, the collaborative representation term plays the main role, while the discrimination term assists with enhancing discrimination by adjusting the coding coefficient errors. Hence, SLDSL can make the best use of the comprehensive information between different gallery set to improve the discriminability of face image sets. (2) It proposes a self-weighted strategy, which can control the contribution of each gallery set to enhance the correlation between the whole gallery set and the ground truth gallery set, such that the nearest point can be obtained exactly. Based on the self-weighted strategy, a novel self-weighted strategy guided discrimination term is proposed to largely boost the discrimination of different gallery sets, such that the effects of true sets can be boosted while the effects of false sets can be weakened or removed. (3) Instead of using explicit sparse constraint, it proposes a latent sparse normalization with capped simplex constraint to produce sparsity solution, thereby reducing the number of involved trade-off parameters and computational complexity. Meanwhile, our SLSDL can obtain sparse solutions to enhance robustness, as the traditional sparse constraint.

3.2 Optimization

Due to only several ℓ_2-norm regularization constraints in problem (2), the convergence is proved easily. Correspondingly, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ can be solved alternatively.

Step 1: Update $\boldsymbol{\alpha}$, *we fix* $\boldsymbol{\beta}$, *and* $\boldsymbol{\gamma}$ First, $\boldsymbol{\alpha}$ can be solved by using the following problem when fixing the other coefficients, i.e.,

$$\begin{aligned} & \min_{\boldsymbol{\alpha}} \|\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{Y}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{i=1}^{m} \|\boldsymbol{X}\alpha - \boldsymbol{X}_i\boldsymbol{\gamma}_i\|_2^2 \\ & \quad \text{s.t. } \boldsymbol{\alpha} \geq 0, \boldsymbol{1}^T\boldsymbol{\alpha} = 1 \end{aligned} \tag{3}$$

Then, this problem can be rewritten as

$$\begin{aligned} & \min_{\boldsymbol{\alpha}} (\|\boldsymbol{X}\boldsymbol{\alpha}\|_2^2 + \|\boldsymbol{X}_i\boldsymbol{\gamma}_i\|_2^2) \\ & \quad - 2\mathrm{tr}\left(\boldsymbol{\alpha}^T\boldsymbol{X}^T\left(\frac{\boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{\lambda}\sum_{i=1}^{m}\boldsymbol{X}_i\boldsymbol{\gamma}_i}{m\lambda + 1}\right)\right) \end{aligned} \tag{4}$$

For simplicity, we introduce an auxiliary variable $\boldsymbol{P} = \frac{1}{m\lambda+1}(\boldsymbol{Y}\boldsymbol{\beta} + \lambda\sum_{i=1}^{m}\boldsymbol{X}_i\boldsymbol{\gamma}_i)$, and transform (4) to

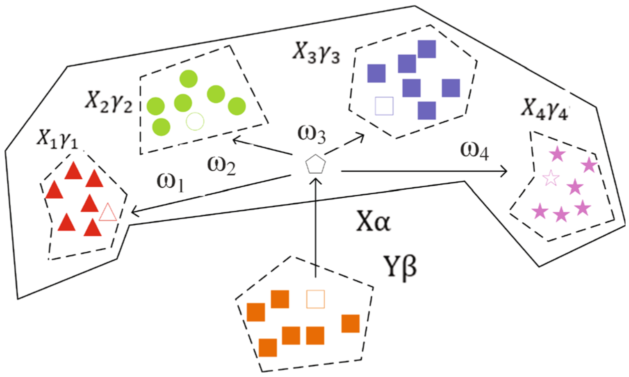


Fig. 2 Illustration of our SLSDL. For instance, five dashed boxes represent four gallery sets $\{X_i\}_{i=1}^4$ and a probe set Y. The solid patterns and hollow patterns substitute the true images and the virtual points, respectively. The arrows substitute distance between the virtual points of image sets. The weighted value w_i stands for the contribution of the i-th gallery set

$$\hat{\alpha} = \arg\min_{\alpha} \|X\alpha - P\|_2^2 \quad \text{s.t. } \alpha \geq 0, 1^T\alpha = 1 \tag{5}$$

Step 2: Update β, we fix γ, and α β can be solved by using the following problem when fixing other coefficients, i.e.,

$$\hat{\beta} = \arg\min_{\beta} \|X\alpha - Y\beta\|_2^2 \quad \text{s.t. } \beta \geq 0, 1^T\beta = 1 \tag{6}$$

Step 3: Update γ, we fix β, and α Similar to problem (6), while fixing other variables, each γ_i can be directly updated via problem (7), i.e.,

$$\hat{\gamma}_i = \arg\min_{\gamma_i} \|X\alpha - X_i\gamma_i\|_2^2 \quad \text{s.t. } \gamma_i \geq 0, 1^T\gamma_i = 1 \tag{7}$$

Interestingly, problems (5), (6), and (7) derive the same optimization manner. We call such the problem *capped simplex projection problem* (CSPP), which adopt affined constraint and the non-negative to output a narrow solution. The solution process of CSPP is presented in Sect. 3.4 for better understand and description.

Step 4: Update w, we fix α, β, and γ We can obtain w via

$$\min_{w} \sum_{i=1}^{m} w_i \|X\alpha - X_i\gamma_i\|_2^2 \quad \text{s.t. } w \geq 0, 1^T w = 1 \tag{8}$$

Then, the values of the weighted vector w can be calculated by Theorem 1, i.e.,

$$w_i = \frac{1}{2\|X\alpha - X_i\gamma_i\|_2 + \zeta} \tag{9}$$

where ζ is infinitely close to zero. The detail of Theorem 1 is shown in Sect. 3.3.

Therefore, according to the above theoretical analysis, we can update α, β, γ, and w alternately. Hereto, the pseudo-code is depicted in Algorithm 1. The demo code will be released on our Github homepage https://github.com/renzhenwen.

Algorithm 1 The pseudocode of our proposed SLSDL

```
Input: Gallery sets $X = [X_1, \cdots, X_m]$ with set size $d$, probe set $Y$, and param $\lambda$.
Compress $\{X_i\}_{i=1}^m$ using KSVD [18] algorithm with $k$ dictionary atoms;
Initialize $\alpha = \frac{1}{mk}$, $\beta = \frac{1}{n}$, $\gamma_i = \frac{1}{k}$, $\epsilon = 10^{-5}$, $iter = 1$, and $mit = 10^3$;
repeat
    Solve $\alpha$ via (5);
    Solve $\beta$ via (6);
    Solve $\{\gamma_i\}_{i=1}^m$ via (7);
    Solve $\{w_i\}_{i=1}^m$ via (9);
    Compute $iter{+}{+}$, and compute the difference between two successive iterators;
until The difference is lower than $\epsilon$ or $iter < mit$;
Output: Perform image set classification via (24).
```

3.3 Determine the value of weight *w*

Recall that the optimization problem (8) can be simplified as below

$$\min_{w} \sum_{i=1}^{m} w_i \|z - g^{(i)}\|_2^2 \quad \text{s.t. } w \geq 0, 1^T w = 1 \tag{10}$$

where $z = X\alpha$ and $\{g^{(i)}\}_{i=1}^m = \{X_i\gamma_i\}_{i=1}^m$. The value of vector w can be calculated via Theorem 1.

Theorem 1 *Motivated by iteratively re-weighted technique, a transition problem without* w *is redefined as follows*

$$\begin{aligned} &\min_{z} \sum_{i=1}^{m} \sqrt{\|z - g^{(i)}\|_2^2} \\ &\text{s.t. } w \geq 0, 1^T w = 1, z \in \mathbf{W} \end{aligned} \tag{11}$$

The Lagrange function of (11) is $\sum_{i=1}^m \sqrt{\|z - g^{(i)}\|_2^2} + \Phi(\xi, z)$, where ξ is the Lagrange multiplier, and $\Phi(\xi, z)$ indicates the indicator function of z from the domain $\mathbf{W}$ (i.e., $z \in \mathbf{W}$). By setting this Lagrange function *w.r.t.* z as zero, we can obtain

$$\sum_{i=1}^{m} \widehat{w}_i \frac{\partial \|z - g^{(i)}\|_2^2}{\partial z} + \frac{\partial \Phi(z)}{\partial z} = 0 \tag{12}$$

where $\widehat{w}_i = 1/(2\|z - g^{(i)}\|_2)$. It is easy to see that Eq. (12) is the same as the derivation of the Lagrange function of problem (10). Thus, $\widehat{w}_i$ can be viewed as the w_i in (10). To avoid the denominator being zero in theory, $\widehat{w}_i$ can be rewritten as

$$w_i = \frac{1}{2\|z - g^{(i)}\|_2 + \zeta} \tag{13}$$

where ζ is infinitely close to zero. The proof is completed. □

3.4 Solve capped simplex projection problem

We adopt the accelerated projected gradient (APG) method to optimize the resultant CSPP. For the convenience of notations, problem (6) and $\mathbf{W}$ are defined as $\varphi(\boldsymbol{\beta})$ and the capped simplex constraint domain, respectively. Without loss of generality, by taking the $\boldsymbol{\beta}$-problem as an example, we define CSPP as below

$$\min_{\boldsymbol{\beta}\in\mathbf{W}} \varphi(\boldsymbol{\beta}) = \|\boldsymbol{x} - \boldsymbol{Y}\boldsymbol{\beta}\|_2^2 \tag{14}$$

For obtaining easier paradigm, an auxiliary variable $\boldsymbol{\eta}$ is introduced to solve problem (14). We can approximatively solve $\boldsymbol{\beta}$ by alternately iterating of $\boldsymbol{\eta}$.

$\boldsymbol{\beta}^0$ is solved from problem (1) when setting $\lambda_2 = 1$ and fixing $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}_i$ without considering the capped simplex constraint. And then starting the alternative optimization by initializing $\boldsymbol{\eta}^0 = \boldsymbol{\beta}^0$. Now, we define the initial value of Newton acceleration coefficient as c and $\boldsymbol{\eta}$ at iteration t as $\boldsymbol{\eta}^t$, and it can be updated at each iteration.

When the t-th iteration, Taylor expansion up to second order is used to approximate $\boldsymbol{\beta}$. i.e.,

$$\begin{aligned}\boldsymbol{\beta}^t &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} f(\boldsymbol{\eta}^{t-1}) \\ &\quad + (\boldsymbol{\beta} - \boldsymbol{\eta}^{t-1})^T \boldsymbol{\varphi}'(\boldsymbol{\beta}^{t-1}) + \frac{\boldsymbol{L}}{2}\|\boldsymbol{\beta} - \boldsymbol{\eta}^{t-1}\|_F^2\end{aligned} \tag{15}$$

Problem (15) can be replaced with a more compact form after ignoring the irrelevant terms. As a consequence, (15) can be rewritten as

$$\boldsymbol{\beta}^t = \arg\min_{\boldsymbol{\beta}\in\mathbf{W}} \frac{\boldsymbol{L}}{2}\|\boldsymbol{\beta} - (\boldsymbol{\eta}^{t-1} - \frac{1}{\boldsymbol{L}}\boldsymbol{\varphi}'(\boldsymbol{\eta}^{t-1}))\|_2^2 \tag{16}$$

Let $\boldsymbol{e} = (\boldsymbol{\eta}^{t-1} - \frac{1}{\boldsymbol{L}}\boldsymbol{\varphi}'(\boldsymbol{\eta}^{t-1}))$. Problem (16) can be abbreviated to

$$\boldsymbol{\beta}^t = \arg\min \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{e}\|_2^2 \quad \text{s.t. } \boldsymbol{\beta} \geq 0, \boldsymbol{1}^T\boldsymbol{\beta} = 1 \tag{17}$$

Analogously to [35, 36], the Lagrangian function of problem (17) is rewritten as

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{w}, \pi) = \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{e}\|_2^2 - \pi(\boldsymbol{1}^T\boldsymbol{z} - 1) - \boldsymbol{w}^T\boldsymbol{z} \tag{18}$$

where $\boldsymbol{w}$ is a Lagrangian coefficient vector and π is a scalar.

Suppose the Lagrange coefficient vectors are $\boldsymbol{w}^*$ and π^*, respectively, and the optimal solution of (18) is $\boldsymbol{z}^*$. According to the Karush–Kuhn–Tucker (KKT) conditions [37, 38], for $\forall j$, we have $\sum_{m=1}^r z_j^* - \sum_{m=1}^r e_j^m - w_j^* - \pi^* = 0$, $z_j^* \geq 0$, $w_j^* \geq 0$ and $z_j^* w_j^* = 0$. Evidently, we also have $r\boldsymbol{z}^* - \sum_{m=1}^r \boldsymbol{e}^m - \boldsymbol{w}^* - \pi^*\boldsymbol{1} = \boldsymbol{0}$.

We obtain $\pi^* = (r - \sum_{m=1}^r \boldsymbol{1}^T\boldsymbol{e}^m - \boldsymbol{1}^T\boldsymbol{w}^*)/n$ according to $\boldsymbol{1}^T\boldsymbol{z}^* = 1$. At this time, we formulate the optimal solution $\boldsymbol{z}^*$ as

$$\frac{\sum_{m=1}^r \boldsymbol{e}^m}{r} + \frac{\boldsymbol{1}}{n} - \frac{\sum_{m=1}^r \boldsymbol{1}^T\boldsymbol{e}^m\boldsymbol{1}}{rn} - \frac{\boldsymbol{1}^T\boldsymbol{w}^*\boldsymbol{1}}{rn} + \frac{\boldsymbol{w}^*}{r} \tag{19}$$

Defining $\boldsymbol{g} = \sum_{m=1}^r \boldsymbol{e}^m/r + \boldsymbol{1}/n - \sum_{m=1}^r \boldsymbol{1}^T\boldsymbol{e}^m\boldsymbol{1}/(rn)$ and $\hat{w}^* = \boldsymbol{1}^T\boldsymbol{w}^*/(rn)$, Eq. (19) can then be simplified to $\boldsymbol{z}^* = \boldsymbol{g} - \hat{w}^*\boldsymbol{1} + \frac{\boldsymbol{w}^*}{r}$. For $\forall j$, the result can be written as

$$z_j^* = g_j - \hat{w}^* + \frac{w_j^*}{r} = \lfloor g_j - \hat{w}^* \rfloor_+ \tag{20}$$

Similarly, $w_j^* = r\lfloor \hat{w}^* - g_j \rfloor_+$ is derived in the same way. Because of $\hat{w}^* = \boldsymbol{1}^T\boldsymbol{w}^*/(rn)$, $\hat{w}^*$ can be written as $\hat{w}^* = \frac{1}{n}\sum_{j=1}^n \lfloor \hat{w}^* - g_j \rfloor_+$. Then, an auxiliary function is defined as Eq. (21) to solve the self-dependent $\hat{w}^*$, i.e.,

$$\Theta(\hat{w}) = \frac{1}{n}\sum_{j=1}^n \lfloor \hat{w}^* - g_j \rfloor_+ - \hat{w} \tag{21}$$

Note that $\hat{w} \geq 0$ and $\Theta'(\hat{w}) \leq 0$, and $\Theta'(\hat{w})$ is a piecewise linear and convex function, we can use Newton method to find the root $\hat{w}^*$ when $\Theta(\hat{w}) = 0$, i.e.,

$$\hat{w}_{t+1} = \hat{w}_t - \frac{\Theta(\hat{w}_t)}{\Theta'(\hat{w}_t)} \tag{22}$$

3.5 Classification

The latent sparse coding coefficients $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$ and $\{\hat{\boldsymbol{\gamma}}_i\}_{i=1}^m$ can be obtained by the optimal solving (2). Then, we compute the similarity between each gallery set and the probe set by distance criterion, which can be represented as follows

$$r_i = (\|\boldsymbol{X}_i\|_* + \|\boldsymbol{Y}\|_*) * \|\boldsymbol{X}_i\hat{\boldsymbol{\alpha}}_i - \boldsymbol{Y}\hat{\boldsymbol{\beta}}\|_2^2 / \|\hat{\boldsymbol{\alpha}}_i\|_2^2 \tag{23}$$

where $\|\boldsymbol{X}_i\|_*$ and $\|\boldsymbol{Y}\|_*$ are the nuclear norm of $\boldsymbol{X}_i$ and $\boldsymbol{Y}$, respectively. And the nuclear norm can be obtained by the sum of the singular values. $\|\boldsymbol{X}_i\hat{\boldsymbol{\alpha}}_i - \boldsymbol{Y}\hat{\boldsymbol{\beta}}\|_2^2$ represents the point-to-point distance of each gallery set and the probe set. $\|\boldsymbol{X}_i\|_* + \|\boldsymbol{Y}\|_*$ can avoid the influence of irrelevant class information between image sets. In the end, we can recognize the identity of the probe image set $\boldsymbol{Y}$ to that of minimum residual distance via

$$\text{identity } (\boldsymbol{Y}) = \arg\min_i \{\boldsymbol{r}_i\} \tag{24}$$

3.6 Computational complexity

Obviously, the computational complexity of the proposed SLSDL mainly depends on the complexity of Algorithm 1, which consists of four steps (i.e., updating $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{w}$).

Note here that the optimization problems of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ have the same optimization manner, i.e., the proposed CSPP. As present in Sect. 3.4, the computational complexity of solving CSPP is $\mathcal{O}(ld)$, where l is the iterators of Newton method for solving CSPP, and d is the dimension of sample. Therefore, the updating problems of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ cost $\mathcal{O}(lmd)$, $\mathcal{O}(ld)$, and $\mathcal{O}(lmd)$, respectively. Moreover, the updating of $\boldsymbol{w}$ costs $\mathcal{O}(m)$. In summary, the computational complexity of Algorithm 1 is $\mathcal{O}(t(2 * lmd + ld + m))$, where t is the number of iterations of Algorithm 1. Since l and m are very small numbers, the computational complexity of our SLSDL can be approximately complexity $\mathcal{O}(td)$.

4 Experiments

To verify the performance of our proposed method, we make the comparison with some baseline methods on four benchmark image set datasets. All comparison experiments are performed by MATLAB 2017b on the PC with an Intel Core i7 (2.9GHz) CPU and 32GB RAM. Numerous experimental results of this section demonstrate the superiority of our proposed method.

4.1 Experimental setup

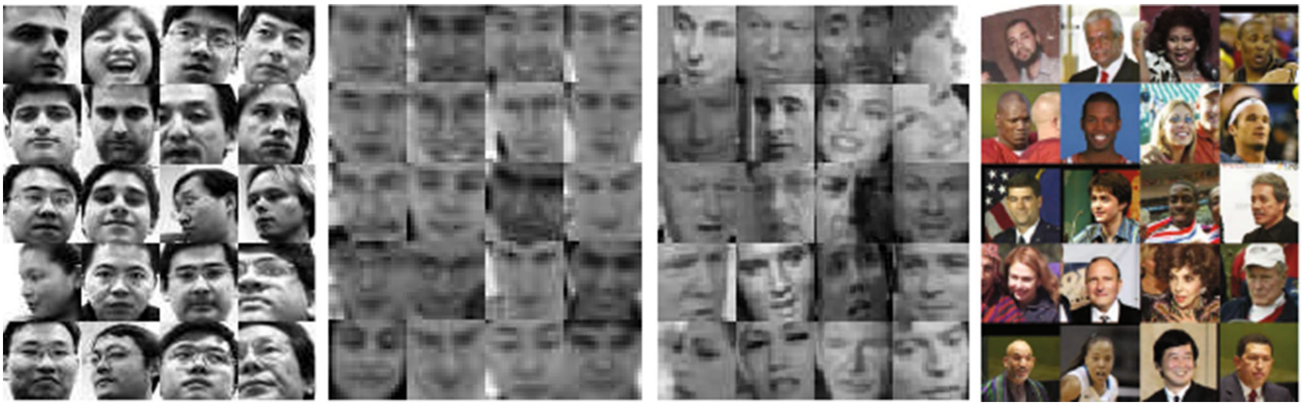
We perform experiments on the four common benchmark face set datasets, which are the Honda/UCSD (Honda) [17, 18], CMU Mobo (Mobo) [18, 39], YouTube Celebrities (YTC) [18, 39, 40] datasets, and Labeled Faces in the Wild Dataset (LFW) [41, 42], as shown in Fig. 3. Note here that we extract the traditional features and deep features successively for comprehensive evaluation of the proposed method. In our experiment, we set the numbers of dictionary atoms as $k = 20$ if there are no special instructions. According to the following parameter sensitivity analysis, the trade-off parameter λ is set as $\lambda = 1$ for Mobo dataset and $\lambda = 0.01$ in three remaining benchmark datasets for simplicity.

Our SLSDL is compared to several representative classification methods for face image set. They were, respectively, AHISD [15], CHISD [15], SANP [16], ISCRC-ℓ_1 [18], ISCRC-ℓ_2 [18], RNP [17], DRM-WV [32], PLRC [21], PDL [20], and LSDL [33]. To obtain the average classification accuracies and standard deviations, we adopt classical approaches which set the size [50, 100, 200] in the first three experiments, respectively. The latest experiment performs with the various resolutions and the deep feature. To observe the effect by using the deep features, we add several comparison methods, such as DLRC [30] and DAR [31]. All comparison methods are performed adopted the source codes, which are given by the authors' homepage. And we use the recommended parameters provided in their original papers for best performance.

4.2 Results on the Honda/UCSD dataset

Honda/UCSD dataset possesses 59 video sequences involving 20 different subjects. Each sequence about 12–645 frames constitutes an image set, which has large variations. We use histogram equalization to eliminate the influences of the illumination and then resize all face images to 20 × 20. Randomly, dividing all image sets into two parts: 20 image sets are used to train and remaining 39 ones are used to test.

The experimental results are listed in Table 1 when the gallery image sets have different number of frames. The best results are shown in boldface. As can be seen, the proposed SLSDL has the highest performance than other



(a) Honda/UCSD **(b)** CMU Mobo **(c)** YouTube Celebrities **(d)** Labled Faces in the Wild

Fig. 3 Sample images of the four datasets. These faces contain high intra-class variations in the form of different poses, illumination variations, expression deformations, and occlusions

Table 1 Classification performance (%) of different methods on the Honda/UCSD and CMU Mobo datasets

Method	Honda/UCSD			CMU Mobo		
	Set size 50	Set size 100	Set size 200	Set size 50	Set size 100	Set size 200
AHISD [15]	82.0	84.6	89.4	91.6 ± 2.8	94.1 ± 2.0	91.9 ± 2.6
CHISD [15]	82.0	87.2	92.3	91.2 ± 3.1	93.8 ± 2.5	96.0 ± 1.3
SANP [16]	84.6	92.3	94.9	91.9 ± 2.7	94.2 ± 2.1	97.3 ± 1.3
RNP [17]	87.2	94.9	97.4	91.9 ± 2.5	94.7 ± 1.2	97.4 ± 1.5
ISCRC-ℓ_1 [18]	89.7	97.4	100.0	93.5 ± 2.8	96.5 ± 1.9	**98.7 ± 1.7**
ISCRC-ℓ_2 [18]	89.7	94.8	100.0	93.5 ± 2.8	96.4 ± 1.9	98.4 ± 1.7
PLRC [21]	87.2	97.4	100.0	92.1 ± 1.6	94.6 ± 1.9	97.5 ± 1.8
PDL [20]	87.2	94.9	97.4	92.5 ± 2.3	94.8 ± 1.9	96.6 ± 2.6
DRM-MV [32]	96.9	99.3	100.0	92.9 ± 1.7	96.2 ± 0.9	98.1 ± 0.8
LSDL [33]	**100.0**	**100.0**	**100.0**	**93.5 ± 2.2**	**96.5 ± 1.8**	**98.5 ± 0.6**
Our SLSDL	**100.0**	**100.0**	**100.0**	**94.1 ± 2.1**	**96.8 ± 1.8**	98.4 ± 1.4

The best results are shown in bold

comparison methods. Both methods achieve much more stable results in all instances and achieve the most perfect classification accuracies, delightfully. However, the comparison methods can achieve better classification effect only if the number of frames is 200. Notably, our methods outperform the second best, DRM-MV, by over 3.1% and 0.7% when the number of frames is 50 and 100%, respectively. The results are excellent demonstrations of the superiority of our method, and it shows that our method further enhances the robustness and discrimination.

4.3 Results on the CMU Mobo dataset

Mobo dataset contains 96 video sequences involving 25 persons, which walk on a treadmill. They implement various kinds of patterns, such as inclined, slow, fast, or carrying a ball. Histogram equalization also is adopted, and then, all face images are resized to 30 × 30. We randomly choose an image set to train and the rest three ones to test from each subject. For the sake of fairness, we perform tenfold cross-validation experiments.

The experimental results of all methods are revealed in Table 1, and the best results are shown in boldface. Notice that overall the proposed SLSDL has slightly higher accuracies than ISCRC and LSDL. More importantly, SLSDL performs higher than ISCRC by 0.6% and 0.4% when set size is 50 and 100, respectively, which is difficult.

4.4 Results on the YouTube Celebrities dataset

In YTC dataset, there are 1910 videos involving 47 actors and politicians. Notice that it was collected from YouTube Web site under uncontrolled real-life circumstance, and these face image set tend to have large pose or expression variation, and motion blur. Due to low quality of images and highly compressed, apparently, the face recognition task will be more challenging. Each face grayscale image is resized to 30 × 30. Randomly, choose three image sets to train and six image sets to test, and conduct fivefold cross-validation experiments.

The comparisons between all methods are summarized in Table 2, and the best results are shown in boldface. Notice that overall the LSDL and SLSDL both outperform other methods. Even though LSDL already achieves very high performance, SLSDL further improves performance about accuracy and robustness. It is worth mentioning that our SLSDL method gets the best performance reported so far and achieves the accuracies up to 74.0%, 78.1%, and 76.0% in different size length. Specially, even though image frames are only 50, our SLSDL has more perfect performance than all comparison methods with different numbers of image frames. This further demonstrates that our method is superior to the comparison methods and has great discriminability and robustness. Clearly, our method can resist noise and even outliers for face image set classification under complex conditions. Compared to DRM-MV, 11.7%, 9.9%, and 5.7% improvements are achieved when frame is 50, 100, and 200, respectively, which is a very surprising result. Meanwhile, we notice that our average classification accuracies fluctuate with the increase in the frame number, probably because more extra information of faces description interferences when the set size is 200.

4.5 Results on the labeled faces in the wild dataset

LFW dataset covers more than 13,000 face images gathered from the Internet. It uses two or more different images of 1680 people as the image set. In LFW database, there are

Table 2 Classification performance (%) and running time of the compared methods on the YTC dataset

Method	Set size 50	Set size 100	Set size 200	Running time (Set size 50)
AHISD [15]	57.1 ± 8.1	59.7 ± 6.4	57.1 ± 8.1	0.92 ± 0.01
CHISD [15]	57.9 ± 6.8	62.7 ± 7.2	64.2 ± 7.5	1.64 ± 0.02
SANP [16]	56.7 ± 5.5	61.9 ± 8.1	65.4 ± 6.8	8.55 ± 0.94
RNP [17]	58.4 ± 6.9	63.2 ± 8.4	65.4 ± 7.2	0.22 ± 0.00
ISCRC-ℓ_1 [18]	62.3 ± 6.2	65.6 ± 6.7	66.7 ± 6.4	0.08 ± 0.00
ISCRC-ℓ_2 [18]	57.4 ± 7.2	60.7 ± 6.5	61.4 ± 6.4	0.04 ± 0.00
PLRC [21]	61.7 ± 8.2	65.6 ± 7.9	66.8 ± 7.5	6.21 ± 0.88
PDL [20]	63.9 ± 6.8	65.7 ± 7.7	67.1 ± 7.6	62.54 ± 5.26
DRM-MV [32]	62.3 ± 5.5	68.2 ± 6.2	70.3 ± 4.8	376.44 ± 15.79
LSDL [33]	**72.2 ± 8.7**	**77.2 ± 9.0**	**75.7 ± 10.5**	0.54 ± 0.01
Our SLSDL	**74.0 ± 9.0**	**78.1 ± 9.2**	**76.0 ± 10.6**	1.05 ± 0.01

The best results are shown in bold

faces images with various illumination, poses and partially obscured. We perform the alignment version LFW-a [31, 42]. Similar to way [30], we resize all face images to 90 × 78. The subset of LFW-a is composed of the classes with more than 20 pictures, and it has 3032 images of 62 subjects. In addition, we want to see how effectively it performs on the recent deep features. Hence, we select the VggFace2 pretrained version of ResNet-50 (VggFace2-ResNet-50) [31] to capture deep feature, and its feature dimensions is 2048. We resize the raw grey-scale images of the LFW-a dataset to 10 × 10, 15 × 10 and 30 × 15, and we utilize VggFace2-ResNet-50 to experiment. The raw images and deep feature are used to evaluate the performance. All images are randomly divided into gallery set and probe set. we perform experiments by setting $d = 5$ and $k = 3$. And thirty-fold cross-validation experiments are conducted.

As shown in Table 3, our SLSDL gains the best scores in accuracy and standard deviation than other comparison methods in LFW-a dataset. We also discover the accuracies improvement when deep features are used. Clearly, the resolution the can improve recognition accuracies from Table 3. However, it is limited to the increase in accuracy compared with that of deep features. These indicate that deep features have more discrimination and robustness for noisy variations than raw pixels. Excitedly, the accuracies of our SLSDL can even be up to 100%. We also notice that our SLSDL gains best recognition accuracies than all other methods in low resolution. Compared to DAR-PE, 6.3%, 3.4%, and 2.1% improvements are achieved under different resolutions, respectively. In addition, our extension SLSDL method is further improved on the basis of LSDL. It indicates that our SLSDL method can enhance effectiveness and robustness.

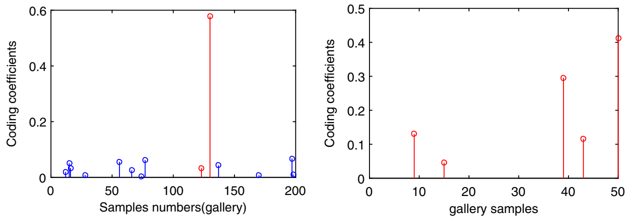
4.6 Running time comparison

As a real-world application, time cost is another noteworthy. Whereupon, we set the number of images per set to 50 and perform experiments about running time on YTC dataset. As we can see from the last column of Table 2, the running time of LSDL is slightly lower and is only 0.54 s.

Table 3 Classification performance (%) of the compared methods on the LFW-a dataset

Method	10 × 10	15 × 10	30 × 15	Deep Features
AHISD [15]	39.8 ± 1.1	42.2 ± 1.1	44.6 ± 1.0	95.6 ± 0.5
CHISD [15]	39.5 ± 1.0	42.0 ± 1.1	44.2 ± 1.0	95.6 ± 0.5
DLRC [30]	39.3 ± 1.0	42.0 ± 1.1	44.3 ± 1.0	97.3 ± 0.4
RH-ISCRC [18]	61.8 ± 1.1	66.2 ± 1.0	68.7 ± 1.0	98.1 ± 0.4
KCH-ISCRC [18]	44.1 ± 1.0	37.8 ± 1.2	56.0 ± 1.2	97.4 ± 0.5
PLRC [21]	42.4 ± 1.1	46.1 ± 1.1	48.1 ± 0.9	96.7 ± 0.3
PDL [20]	58.3 ± 1.3	63.2 ± 1.1	61.7 ± 1.3	98.2 ± 0.3
DAR-TE [31]	55.0 ± 1.1	63.8 ± 1.1	71.3 ± 1.0	98.6 ± 0.3
DAR-PE [31]	63.4 ± 1.1	69.0 ± 1.0	73.8 ± 1.0	**99.7 ± 0.1**
LSDL [33]	**67.5 ± 1.2**	**71.6 ± 1.0**	**75.8 ± 1.2**	99.4 ± 0.2
Our SLSDL	**69.7 ± 1.0**	**72.4 ± 1.1**	**75.9 ± 1.1**	**100.0 ± 0.0**

The best results are shown in bold



(a) α **(b)** β

Fig. 4 Coding coefficients of the gallery sets (i.e., $\boldsymbol{\alpha}$) are on the left and that of the probe set (i.e., $\boldsymbol{\beta}$) are on the right, produced by the proposed SLSDL method. Honda dataset contains 20 subjects, and $k = 10$

And our SLSDL is slightly higher time cost about 1.05 s. Note here that the running time of our SLSDL has same order of magnitude as that of ISCRC and RNP about 1 s. Clearly, we perform a self-weighted strategy for improving image set classification performance without sacrificing too much computation time.

4.7 Latent sparse property analysis

As we make a point in Sect. 3.1, SLSDL has the latent sparse property. Here proving this point from experiments. We conduct experiments by setting $k = 50$ and $d = 10$ on Honda dataset. Coding coefficients of the gallery sets $\boldsymbol{\alpha}$ are on the left and that of the probe set $\boldsymbol{\beta}$ are on the right from Fig. 4. It is clearly that red coefficients correspond to the correct class, and the coefficients solutions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of our method have the sparsity. The experimental results of reconstructed face images and residuals are shown in Fig. 5. All $\boldsymbol{X}_i\boldsymbol{\alpha}_i$ denote the virtual point of each gallery image set. The reconstruction residual is simplified to $\|\boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{Y}\boldsymbol{\beta}\|_2^2$. The smallest residual between the probe image set and the gallery image sets is highlighted with a red color, which represents that they are the same subject. This demonstrates that our SLSDL can not only have sparse solutions to enhance robustness but also reduce computational complexity.

4.8 Convergence analysis

By the previous description, our SLSDL is clearly convex, and it has closed-form solutions on the basis of the mentioned in Sect. 3.2. By evaluating on the Honda, Mobo, YTC, and LFW-a datasets, the convergence curves of the proposed SLSDL method are shown in Fig. 6. It shows that SLSDL converges fast and the objective function value achieves stable state about 15–20 iterations, even as long as 5 iterations on the first two databases.

4.9 Parameter sensitivity

Pleasantly, our SLSDL has only one parameter, which needs to be adjusted. For obtaining its optimum performance of parameter λ, we set adjustment range varied from $\{10^{-7},\ldots,10^{1}\}$, and fix $d = 50$. For the sake of simplicity, we perform experiment as representativeness in Mobo and YTC. The performance curves are revealed in Fig. 7. We observe that the classification accuracy fluctuates with λ, and the best effects are achieved at the appropriate values. The recognition accuracies decrease rapidly when $\lambda = 10$, because it plays a supporting role. Obviously, SLSDL can obtain the more perfect classification accuracies when $\lambda = 1$ and $\lambda = 0.01$ on two benchmark datasets, respectively. This further indicates that the latent sparse normalization and a self-weighted term can enhance the discrimination. Overall, we suggest to set λ as a relatively small value within the range of $\{10^{-3},\ldots,1\}$ for different datasets.

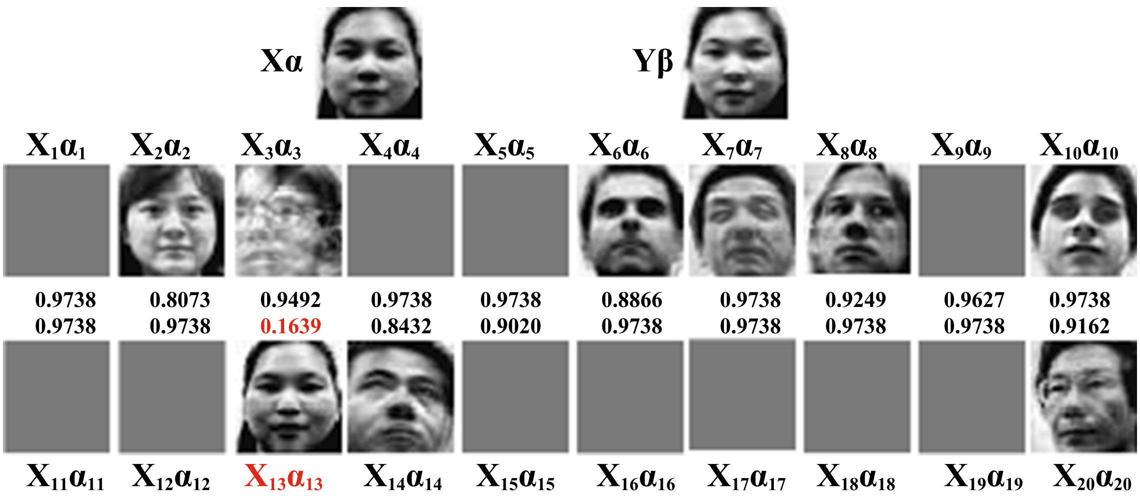
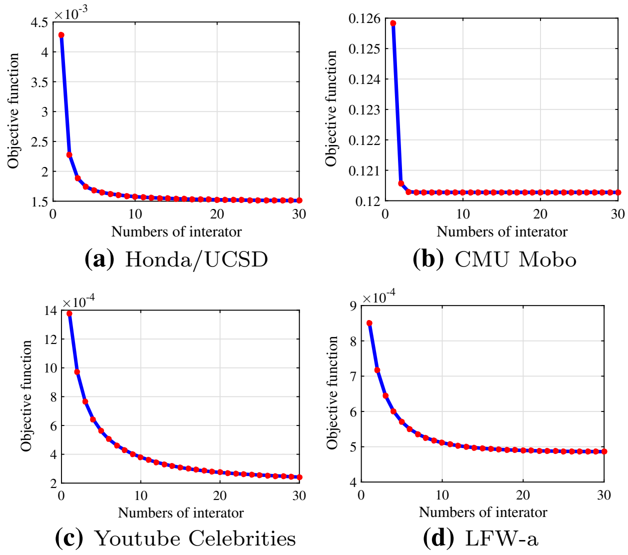


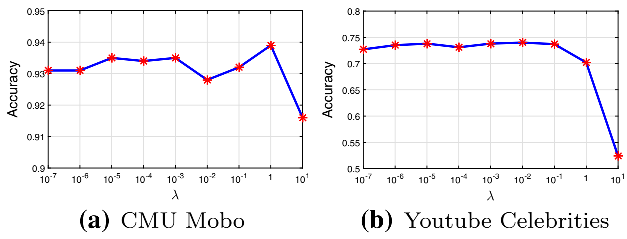
Fig. 5 Virtual faces produced by our SLSDL method on the Honda/UCSD dataset. $\boldsymbol{X}\boldsymbol{\alpha}$, $\boldsymbol{Y}\boldsymbol{\beta}$, and $\boldsymbol{X}_i\boldsymbol{\alpha}_i$ represent the virtual face point of the whole gallery sets $\boldsymbol{X}$, the probe set $\boldsymbol{Y}$, and the i-th gallery set $\boldsymbol{X}_i$, respectively



(a) Honda/UCSD **(b)** CMU Mobo

(c) Youtube Celebrities **(d)** LFW-a

Fig. 6 Convergence of the proposed SLSDL method on the four datasets



(a) CMU Mobo **(b)** Youtube Celebrities

Fig. 7 Classification performance of the proposed SLSDL method *w.r.t.* λ on the Mobo and YTC datasets

Table 4 Classification performance (%) of LSR, LSDL and SLSDL on the four datasets

Dataset	LSR	LSDL	SLSDL
Honda/UCSD	**100**	**100**	**100**
CMU Mobo	93.0	93.5	**94.1**
YouTube Celebrities	72.0	72.2	**74.0**
LFW-a (10×10)	67.3	67.5	**69.7**

The best results are shown in bold

4.10 Ablation analysis

In our method, a new discriminative strategy is proposed to control the contribution of each gallery set, such that each gallery set can be treated differently. It can adjust the coding coefficient errors to enhance the set discrimination. We dub it as *latent sparse representation* (LSR) when LSDL does not have a discriminative term, i.e., $\lambda = 0$. In order to demonstrate this, we perform experiments by setting $\lambda = 0$ and $d = 50$ in the first three datasets, and setting $k = 3$ in the LFW-a. Table 4 shows that performance decreases without the discriminative term on Mobo, YTC, and LFW-a. Hence, it clearly shows that the discriminative term can boost the discrimination power. In the addition, we can see easily that the correct gallery (the red mark) has a larger weight value, and the incorrect galleries have lower weight values from Fig. 8. Clearly, the weighted value can boost the correlation power of homogeneous sets while limiting the correlation power of heterogeneous sets.

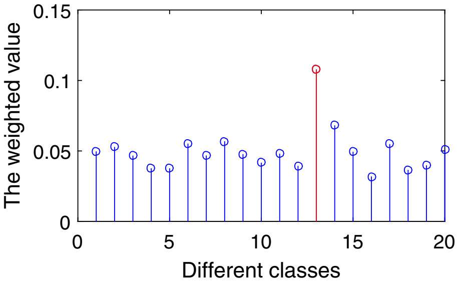


Fig. 8 Weight vector $\boldsymbol{w}$ produced by the proposed SLSDL method on the YTC dataset. Note here that the red color stem stands for the ground-true class

5 Conclusion

This paper proposes a novel SLSDL method for image set classification. SLSDL considers the differences of contributions between the whole gallery sets and each gallery set and proposes a self-weighted strategy to reflect the correlation between the whole gallery set and each gallery set, which is different from the existing point-to-point distance-based methods; moreover, SLDSL proposes a latent sparse normalization with capped simplex constraint to avoid high running time as well as enhancing robustness. A large number of experiments on four public benchmark face set datasets demonstrate that SLSDL is superior to the competitors about computational speed and classification performance.

Although SLSDL has achieved promising performance, some aspects still deserve to study in the future. (1) We will integrate hash learning and deep learning [43] for large-scale image set classification. (2) We will introduce multiple kernel learning [37] for handling the problem of nonlinear image set classification.

Acknowledgements This research was supported by the Sichuan Science and Technology Program (Grant Nos. 2019ZDZX0043 and 2020ZDZX0014), the Key Lab of Film and TV Media Technology of Zhejiang Province (Grant No. 2020E10015), the Natural Science Foundation Project of CQ CSTC (Grant No. cstc2020jcyj-msxmX0473), the Major Cultivation Research Projects of Chongqing

Three Gorges University (Grant No. 18ZDPY07), the Scientific Research Fund of Sichuan Provincial Education Department (Grant No. 17ZB0441), the Scientific Research Fund of Southwest University of Science and Technology (Grant No. 17zx7137), the Postgraduate Innovation Fund Project by Southwest University of Science and Technology (Grant No. 20ycx0001), and the Undergraduate Innovation and Entrepreneurship Training Program of Sichuan (Grant No. S202010619014).

Compliance with ethical standards

Conflict of interest There is no conflicts of interest.

References

1. Yang J, Liu Y (2019) Undersampled face recognition based on virtual samples and representation classification. Neural Comput Appl 31(7):2447–2453
2. Wei D, Shen X, Sun Q, Gao X, Yan W (2020) Prototype learning and collaborative representation using Grassmann manifolds for image set classification. Pattern Recogn 100:107123
3. Liu B, Jing L, Li J, Yu J, Gittens A, Mahoney MW (2019) Group collaborative representation for image set classification. Int J Comput Vis 127(2):181–206
4. Gao X, Sun Q, Xu H, Wei D, Gao J (2019) Multi-model fusion metric learning for image set classification. Knowl Based Syst 164:253–264
5. 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
6. Ren Z, Sun Q, Yang C (2020) Nonnegative discriminative encoded nearest points for image set classification. Neural Comput Appl 32(13):9081–9092
7. Cevikalp H, Yavuz HS, Triggs B (2019) Face recognition based on videos by using convex hulls. IEEE Trans Circuits Syst Video Technol PP(99):1
8. Lei D, Jiang Z, Wu Y (2020) Weighted huber constrained sparse face recognition. Neural Comput Appl 32(9):5235–5253
9. Huang C, Li Y, Chen CL, Tang X (2019) Deep imbalanced learning for face recognition and attribute prediction. IEEE Trans Pattern Anal Mach Intell 42(11):2781–2794
10. Zhang Z, Jiang W, Qin J, Zhang L, Li F, Zhang M, Yan S (2017) Jointly learning structured analysis discriminative dictionary and analysis multiclass classifier. IEEE Trans Neural Netw Learn Syst 29(8):3798–3814
11. Zhang Z, Sun Y, Wang Y, Zhang Z, Zhang H, Liu G, Wang M (2020) Twin-incoherent self-expressive locality-adaptive latent dictionary pair learning for classification. IEEE Trans Neural Netw Learn Syst 99:1–15
12. Zhang Z, Jiang W, Zhang Z, Li S, Liu G, Qin J (2019a) Scalable block-diagonal locality-constrained projective dictionary learning. In: Proceedings of the 28th international joint conference on artificial intelligence, AAAI Press, pp 4376–4382
13. Zhang Z, Ren J, Jiang W, Zhang Z, Hong R, Yan S, Wang M (2019b) Joint subspace recovery and enhanced locality driven robust flexible discriminative dictionary learning. IEEE Trans Circuits Syst Video Technol 30:42
14. Sun Y, Zhang Z, Jiang W, Zhang Z, Zhang L, Yan S, Wang M (2020) Discriminative local sparse representation by robust adaptive dictionary pair learning. IEEE Trans Neural Netw Learn Syst 31(10):4303–4317
15. Cevikalp H, Triggs B (2010) Face recognition based on image sets. In: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE, pp 2567–2573
16. 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
17. Yang M, Zhu P, Van Gool L, Zhang L (2013) Face recognition based on regularized nearest points between image sets. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), IEEE, pp 1–7
18. 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
19. Zhang L, Yang M, Feng X, Ma Y, Zhang D (2012) Collaborative representation based classification for face recognition. arXiv preprint arXiv:12042358
20. Wang W, Wang R, Shan S, Chen X (2016) Prototype discriminative learning for face image set classification. In: Asian conference on computer vision, Springer, pp 344–360
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. Yang M, Wang X, Liu W, Shen L (2017) Joint regularized nearest points for image set based face recognition. Image Vision Comput 58:47–60
23. Zheng P, Zhao ZQ, Gao J, Wu X (2017) Image set classification based on cooperative sparse representation. Pattern Recogn 63:206–217
24. Huang Z, Shan S, Wang R, Zhang H, Lao S, Kuerban A, Chen X (2015) A benchmark and comparative study of video-based face recognition on cox face database. IEEE Trans Image Process 24(12):5967–5981
25. 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 Vision Image Underst 160:1–23
26. Liu X, Guo Z, You J, Kumar BV (2019) Dependency-aware attention control for image set-based face recognition. IEEE Trans Inf Forensics Secur 15:1501–1512
27. Song Z, Cui K, Cheng G (2020) Image set face recognition based on extended low rank recovery and collaborative representation. Int J Mach Learn Cybern 11(1):71–80
28. Mian A, Hu Y, Hartley R, Owens R (2013) Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning. IEEE Trans Image Process 22(12):5252–5262
29. Wang G, Shi N (2020) Collaborative representation-based discriminant neighborhood projections for face recognition. Neural Comput Appl 32(10):5815–5832
30. Chen L (2014) Dual linear regression based classification for face cluster recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2673–2680
31. Ren CX, Luo YW, Xu XL, Dai DQ, Yan H (2019) Discriminative residual analysis for image set classification with posture and age variations. IEEE Trans Image Process 29:2875–2888
32. Hayat M, Bennamoun M, An S (2014) Deep reconstruction models for image set classification. IEEE Trans Pattern Anal Mach Intell 37(4):713–727
33. Sun Y, Ren Z, Yang C, Lei H (2020) Latent sparse discriminative learning for face image set classification. In: International conference on neural computing for advanced applications, Springer, pp 144–156
34. Ren Z, Sun Q (2020) Simultaneous global and local graph structure preserving for multiple kernel clustering. IEEE Trans Neural Netw Learn Syst PP:2

35. Nie F, Wang X, Jordan MI, Huang H (2016) The constrained Laplacian rank algorithm for graph-based clustering. In: Proceedings of the Thirtieth AAAI conference on artificial intelligence, pp 1969–1976
36. Wang H, Yang Y, Liu B, Fujita H (2019) A study of graph-based system for multi-view clustering. Knowl Based Syst 163:1009–1019
37. Ren Z, Yang SX, Sun Q, Wang T (2020a) Consensus affinity graph learning for multiple kernel clustering. IEEE Trans Cybern PP:1
38. Ren Z, Mukherjee M, Lloret J, Venu P (2020b) Multiple kernel driven clustering with locally consistent and selfish graph in industrial IoT. IEEE Trans Ind Inf PP:3
39. Ren Z, Wu B, Zhang X, Sun Q (2019a) Image set classification using candidate sets selection and improved reverse training. Neurocomputing 341:60–69
40. Ren Z, Sun Q, Wu B, Zhang X, Yan W (2019b) Learning latent low-rank and sparse embedding for robust image feature extraction. IEEE Trans Image Process 29(1):2094–2107
41. Learned-Miller E, Huang GB, RoyChowdhury A, Li H, Hua G (2016) Labeled faces in the wild: a survey. Advances in face detection and facial image analysis. Springer, Cham, pp 189–248
42. Taigman Y, Wolf L, Hassner T et al (2009) Multiple one-shots for utilizing class label information. BMVC 2:1–12
43. Cui H, Zhu L, Li J, Yang Y, Nie L (2019) Scalable deep hashing for large-scale social image retrieval. IEEE Trans Image Process 29:1271–1284

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