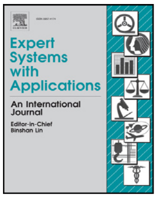
Expert Systems With Applications 134 (2019) 102–111

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Simultaneous learning of reduced prototypes and local metric for image set classification

Zhenwen Ren[a,b,*], Bin Wu[c], Quansen Sun[b], Mingna Wu[a]

[a] School of National Defence Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China
[b] Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, China
[c] School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, China

ARTICLE INFO

Article history:
Received 24 July 2018
Revised 19 May 2019
Accepted 19 May 2019
Available online 24 May 2019

Keywords:
Image set classification
Prototype learning
Metric learning
Face recognition
Expert system

ABSTRACT

Classification based on image set is recently a competitive technique, where each set contains multiple images of a person or an object. As a widely used model, affine hull has shown its power in modeling image set. However, due to the existence of noise and outliers, the over-large affine hull usually matches fails when two hulls overlapped. Aiming at alleviating this handicap, this paper proposes a novel method for image set classification, namely Learning of Reduced Prototypes and Local Metric (LRPLM). Specifically, for each gallery image set, a reduced set of prototypes and an optimal local feature-wise metric are simultaneously learned, which jointly minimize the loss function involved the estimation of classification error probability. In doing so, LRPLM inherits the merits of affine hull with better representation to account for the unseen appearances and makes use of the powerful discriminative ability improved by the local metric. It looks like that LRPLM pulls similar image sets with the same class label "closer" to each other, while pushing dissimilar ones "far away". Extensive experiments illustrate the considerable effectiveness of LRPLM on three widely used datasets. As we know, classification is a research hotspot in expert and intelligent systems. Different from the previous classification methods, LRPLM focuses on image set-based classification technology, while most of them are single-shot classification technology. Thus, the proposed method can be considered as an expert system technology for medical diagnosis, security monitoring, object categorization, and biometrics recognition applications.

1. Introduction

With the significant progress of chip and imaging technologies in recent years, multiple images of a person or an object are readily available in many real-life applications, including multi-view visual recognition, video-based surveillance, video retrieval, dynamic scene recognition, and so on. Generally, an image set contains multiple still images that are captured under various circumstances, or frames cropped from a video clip. The goal of image set classification is to assign the class label to a probe image set, where the similarity between a gallery and a probe is obtained by calculating the set-to-set distance.

Different from traditional single-shot image recognition tasks (Behera, Dogra, & Roy, 2018; Hanmandlu & Mamta, 2014; Zhang et al., 2016), image set-based classification methods are more promising and challenging (Chen, Jiang, Tang, & Luo, 2017; Zhu, Zuo, Zhang, Shiu, & Zhang, 2014). The main attraction is that it can effectively deal with a variety of appearance variations to improve the effectiveness and the robustness to image variations. However, the major difficulty is that image sets exhibit huge intra-class variability and large inter-class ambiguity, which poses great challenge to make effective use of the set information of the data, and to faithfully measure the similarity between two image sets for accurate classification. As we know, expert system has been used widely in many areas and industries, e.g, fault diagnosis (Glowacz, 2018a; 2018b), repair, instruction, interpretation, classification (Soleymani, Granger, & Fumera, 2018; Zeng, Gou, & Deng, 2017; Zhang et al., 2016), monitoring and others many (Tan et al., 2016). Image set classification technology is greatly prominent compared to the other classification technologies, which can enrich the theory of expert system. We present some typical application scenarios of image set classification in Fig. 1.

Among the literatures we surveyed, the efficient affine/convex hull model is proposed in Cevikalp and Triggs (2010), which

[*] Corresponding author.
E-mail addresses: rzw@njust.edu.cn (Z. Ren), wb@swust.edu.cn (B. Wu), sqs@njust.edu.cn (Q. Sun).

https://doi.org/10.1016/j.eswa.2019.05.025
0957-4174/© 2019 Elsevier Ltd. All rights reserved.

Security and surveillance systems
Biological recognition
Gesture recognition
Object recognition

Fig. 1. Typical application scenarios of image set classification, e.g., security and surveillance systems, biological recognition, gesture recognition, and object recognition.

provides better localization to account for the unseen appearances of image set, since it attempts to reduce sensitivity of within-class variations by artificially generating samples within the set. In Cevikalp and Triggs (2010), the similarity between two sets is defined as the Euclidean distance between a pair of nearest points from two hulls respectively, which also can be considered as a nearest neighbour (NN) classifier (Behera et al., 2018). However, the images of one image set are often corrupted by strong noise and outliers in real-life application, the affine hulls of different sets are more likely to intersect each other due to the over-large geometric region structure, which will result in dramatic deterioration of classification performance of NN classifier.

To address this deadly drawback of affine hull model, inspired by the weighted metric learning algorithm (Paredes & Vidal, 2006) and Prototype Discriminative Learning (PDL) method (Wang, Wang, Shan, & Chen, 2017), we propose a novel method called Learning of Reduced Prototypes and Local Metric (LRPLM), which simultaneously learns both a reduced set of prototypes and an optimal local feature-wise metric for each gallery set. In LRPLM, prototype is actually virtual sample, i.e., it does not certainly appear in a set but is only assumed to belong to the corresponding affine hull. Therefore, each gallery image set is naturally modeled as a hull which is spanned by those prototypes to account for the unseen appearances and reduce the computational and storage costs. The local metric is a feature-dependent weight matrix for the corresponding prototype set, which can significantly improve the performance of the traditional NN classifier. The object function is derived by minimizing a specific criterion, which is an approximation of the estimate of the classification error probability. More specifically, for each image set, starting with an initial selection of a few representatives as prototypes and an all-ones matrix as local metric, LRPLM iteratively adjusts both the positions (i.e., affine coefficients) of these prototypes and their corresponding feature-dependent weights (i.e., local metric) simultaneously. In doing so, they can be boosted mutually so that the obtained prototypes and metric matrixes can significantly improve the robustness and effectiveness of between-set similarity. From another point of view, LRPLM brings similar image sets “closer” to each other, while pushing dissimilar ones “far away”. Therefore, LRPLM not only inherits the advantage of affine hull with better representation to account for the unseen appearances, but also significantly improves the discrimination ability of gallery sets. However, the weakness of LRPLM is that although we give its convergence property from experimental perspective, it is difficult to prove mathematically. In summary, the main contributions of this work can be summarized by the following:

- To reduce the computational and storage costs while significantly increasing the robustness of the hull model, we propose to learn a few prototypes to efficiently represent an image set.
- To significantly improve the classification accuracy of the NN classifier, we propose to tune a local feature-wise metric for each prototype set. Most of the previous methods use the simple NN classifier directly, but the proposed method uses the local metric weighted enhanced NN classifier.
- An effective optimization strategy of alternately gradient descent procedure is developed to solve the resulting optimization problem, and the convergence analysis of the designed optimization problem is presented from experimental perspective. In this way, the prototype sets and the corresponding local metric matrices can be boosted mutually so that the obtained results are optimal for image set classification.
- Extensive experiments show that the proposed LRPLM performs better than the previous methods on three widely used datasets.

The paper is outlined as follows: In Section 2, we briefly review some related works. In Section 3, the proposed method and its optimization scheme are presented. In Section 4, the experimental results and discussion are represented. In Section 5, the conclusion is offered.

2. Related work

Existing image set-based classification methods mainly focused on set modeling and similarity computation, i.e., 1) an effective model has to be carefully designed to express the potential set relationship between images; 2) a suitable distance metric has to be defined for computing the similarity between such set models. Once the model is chosen, the corresponding similarity computation strategy is roughly fixed.

From the view of set modeling, existing methods can be divided into five categories: 1) statistical model-based methods (Wang, Guo, Davis, & Dai, 2012a; Wang, Wang, Shan, & Chen, 2015); 2) linear subspace-based methods (Kim, Kittler, & Cipolla, 2007); 3) nonlinear manifold-based methods (Harandi, Sanderson, Shirazi, & Lovell, 2011; Wang, Shan, Chen, Dai, & Gao, 2012b); 4) affine/convex hull subspace-based methods (Cevikalp & Triggs, 2010; Hu, Mian, & Owens, 2012; Wang et al., 2015; 2017; Yang, Zhu, Van Gool, & Zhang, 2013); 5) compressed sensing-based methods (Chen et al., 2017; Hu et al., 2012; Xuan, Wang, Zhao, & Liu, 2013; Zeng et al., 2017). Besides, deep learning (DL) techniques have recently gained significant successes in some computer vision and pattern recognition tasks (Affonso, Rossi, Vieira, & de Leon Ferreira, 2017), but its applications to image set classification are few, the most recent articles are Shah, Bennamoun, and Boussaid (2016) and Hayat, Bennamoun, and An (2015).

In the above methods, the hull based methods have gained the most attention, which usually rely on a distance metric to measure the between-set similarity. The similarity between two image sets is the distance between a pair of optimal nearest points belonging to either hull respectively. In general, there are two widely used distance metrics, including the simple Euclidean distance metric and the Mahalanobis distance metric. For example, Affine Hull-based Image Set Distance (AHISD) (Cevikalp & Triggs, 2010), Convex Hull-based Image Set Distance (CHISD)

(Cevikalp & Triggs, 2010), and Sparse Approximated Nearest Points (SANP) (Hu et al., 2012) model an image set as an affine/convex hull, and search a pair of nearest points without any additional constraints. However, the structure of image set is largely ignored by them. With consideration of the structure of image set, Regularized Nearest Points (RNP) (Yang et al., 2013), Joint Regularized Nearest Points (JRNP) (Yang, Wang, Liu, & Shen, 2017) and Collaboratively Regularized Nearest Points (CRNP) (Wu, Minoh, & Mukunoki, 2013) has been proposed and achieved excellent performance compared with previous affine/convex hull-based methods by modeling an image set as a regularized affine hull (RAH). The reason is that RAH can avoid containing the meaningless points which are too far from the sample mean. These methods always use the Euclidean distance metric for measuring the between-set similarity. To learn a more discriminative and robust distance metric for preserving the similarity relationships among data, the Mahalanobis distance metric is used. For example, Self-Regularized Non-Negative Coding and Adaptive Distance Metric Learning (SRN-ADML+DW) (Mian, Hu, Hartley, & Owens, 2013) learns a more discriminative Mahalanobis distance for robust face recognition. DMPL (Köstinger, Wohlhart, Roth, & Bischof, 2013) jointly chooses the positioning of prototypes and also optimizes the Mahalanobis distance metric. However, most of these hull based methods are online ones (i.e., doing all computations at run time), which suffer high computational complexity. In addition, they may have inferior robustness when the data contains noise and outliers.

To address these problems, joint the hull model (Cevikalp & Triggs, 2010; Karanam, Wu, & Radke, 2018), metric learning (Liu, Xu, Tsang, & Zhang, 2019; Zhang et al., 2018) and prototype learning (Shao, Huang, Yang, & Luo, 2018) methods have recently been proposed in the literatures. For example, Set-to-set Prototype and Metric Learning framework (SPML) (Leng, Moutafis, & Kakadiaris, 2015) jointly learns prototypes and a Mahalanobis distance to reduce the time and storage requirements while the classification accuracy is maintained or improved. Set-to-set Distance Metric Learning (SSDML) (Zhu, Zhang, Zuo, & Zhang, 2013) learns point-to-set distance (PSD) and even set-to-set distance (SSD), then tackles SSDML problems by using standard support vector machine solvers, to make the metric learning very efficient for multi-class visual classification tasks. Prototype Discriminative Learning (PDL) (Wang et al., 2017) simultaneously learns a set of prototypes for each image set and a global orthometric discriminative projection to shrink the loose affine hull, whose projection matrix can be seen as a global metric. However, the learned projection matrix of PDL is a global metric rather than a local one. In theory, learning a local metric is more conducive to further improve performance. This is also the main strength of this work.

3. Proposed method

We first review the basic concept of the reduced affine hull model. Then we present the definition of prototype and local metric. Next, we describe the proposed LRPLM method and the optimization scheme. Finally, we give the testing process of LRPLM for image set classification.

3.1. Reduced affine hull model

Suppose that we have a collection of labeled image sets $\mathcal{X} = \{X_1, X_2, \cdots, X_C\}$, each of which denotes a gallery image set, and $X_c = \{x_{c,1}, x_{c,2}, \cdots, x_{c,n_c}\}$, where n_c is the size of cth image set, and $x_{c,k} \subset \mathbb{R}^d$ $(1 \le c \le C, 1 \le k \le n_c)$ is a d-dimensional features vector (i.e., the kth sample of the cth class).

As mentioned above, an image set can be approximated as an affine hull which provides better localization to account for the unseen appearances and implicitly selects few representative prototypes instead of all the samples. To be more general, the affine hull of image set X_c can be written as:

$$H_c^{aff} = \{x\}, \forall x = \textstyle\sum_{k=1}^{n_c} \omega_{c,k} x_{c,k}, \sum_{k=1}^{n_c} \omega_{c,k} = 1. \tag{1}$$

By using the sample mean $\mu_c = \frac{1}{n_c}\sum_{k=1}^{n_c} x_{ck}$, the unconstrained representation of (1) can be defined as follow:

$$H_c^{aff} = \{x\}, \forall x = \mu_c + U_c v_c, v_c \in \mathbb{R}^{l_c}, c = [1, 2 \cdots C], \tag{2}$$

where U_c is an orthonormal basis matrix which is obtained by applying the Singular Value Decomposition (SVD) to centered data $[x_{c1} - \mu_c, x_{c2} - \mu_c, \cdots, x_{cn_c} - \mu_c]$ with holding the expected energy rate, v_c is a vector of free parameters (i.e., affine coefficient) w.r.t. U_c, and l_c $(l_c < < n_c)$ is the number of the reserved singular values of U_c. The remaining singular values are called leading singular values. Based on Eq. (2), the similarity between two image sets can be obtained by computing the Euclid distance between the closest points from the two affine geometric regions (Cevikalp & Triggs, 2010).

Although the affine hull model provides better localization ability to account for the unseen appearances of image set and a simple distance metric to compute the similarity between two image sets, the affine hulls are likely to be over-large due to the existence of noise and outliers. As we will show in the next section, we propose a novel LRPLM method which simultaneously learns a reduced set of prototypes and a local metric matrix for each image set to eliminate the performance deterioration caused by the ambiguity of the overlarge geometry of the affine hull.

3.2. Definitions of prototype and local metric

Let $\mathcal{P} = \{P_1, P_2, \cdots, P_C\}$ be a collection of prototype sets of the given gallery sets $\mathcal{X}$. Among them, for the cth gallery image set X_c, its corresponding prototype set $P_c = \{y_{c,1}, y_{c,2}, \cdots, y_{c,m_c}\} \subseteq H_c^{aff}$ is a reduced set of prototype points belonging to the affine hull of X_c, where m_c $(m_c = l_c, m_c << n_c)$ is the size of P_c, and $y_{c,i} \subset \mathbb{R}^d$ $(1 \le c \le C, 1 \le i \le m_c)$ is the ith prototype of P_c. As defined in the previous section, a prototype can be written as

$$y_{c,i} = \mu_c + U_c v_{c,i}, v_{c,i} \in \mathbb{R}^{l_c}. \tag{3}$$

Note that y may not appear in the set, i.e., the prototype y is actually "virtual" image.

Moreover, let $\mathcal{W} = \{W_1, W_2, \cdots, W_C\}$ be a collection of local feature-dependent metric matrix. For each prototype set P_c, whose corresponding local metric can be represented as a weight matrix $W_c = \{w_{c,ij}, 1 \leqslant c \leqslant C, 1 \leqslant i \leqslant m_c, 1 \leqslant j \leqslant d\}$.

To better describe the proposed method, some concise presentations and functional functions will be needed throughout the paper:

1) The local feature-dependent weight $w_{c,\,ij}$ is abbreviated as w_{ij}^c.

2) The jth feature of the prototype $y_{c,i}$ and the orthonormal base $U_{c,i}$ are denoted as $y_{c,ij}$ and $U_{c,ij}$, respectively.

3) The class label of a prototype $y \in P_c$ is denoted as $class(x)$, i.e., $class(x) = c$.

4) The index of a prototype $y \in P_c$ is denoted as $index(y)$, i.e., $index(y) = i$ iff $y = y_{c,i}$.

By using $\mathcal{W}$, the local feature-wise weighted distance from an arbitrary image x to a prototype y_i can be seen as a weighted Euclidean distance, i.e.,

$$d_W(x, y_i) = \sqrt{\sum_{j=1}^{d} (w_{ij}^c)^2 (x_j - y_{ij})^2}, \tag{4}$$

where $c = class(y_i)$, w_{ij}^c is the local metric associated with the jth dimension feature of the prototype y_i. Note that if $w_{ij}^c = 1$, Eq. (4) is just the most commonly used ℓ_2 distance, if the weights are the inverse of the variances in each dimension, Eq. (4) is the Mahalanobis distance (MD), which is unitless and scale-invariant, and takes into account the correlations of the data set. For brevity, $d_W(x, y_i)$ is denoted as $d(x, y_i)$ in this paper.

3.3. Learning reduced prototype set and local metric

In this paper, we aim to simultaneously learn a collection of reduced prototype sets $\mathcal{P}$ and their corresponding local feature-dependent metric matrix set $\mathcal{W}$ by minimizing a specific criterion, which is an approximation to the leaving-one-out (LOO) NN estimating of the probability of classification error. Specifically, it makes points belonging to different classes discriminative by using a similarity measure based on the affine hull model, i.e., the weighted distance should be low for the prototypes belonging to the same class while high for the prototypes laying in different class. This idea suggests the following loss function to be minimized:

$$J(\mathcal{P}, \mathcal{W}) = \sum_{c=1}^{C} \sum_{\forall x \in X_c} \frac{1}{n_c} S_\beta(Q(x)), \tag{5}$$

where S_β is the step function, $Q(x)$ is a discrimination function. Since $Q(x)$ is used for estimating of the probability of classification error, it can be defined as

$$Q(x) = \frac{d(x, y_x^{=})}{d(x, y_x^{\neq})}, \tag{6}$$

where $d(x, \cdot)$ is the local weighted distance between x and a found prototype, $y_x^{=}$ is the NN prototype of x in the same class, and $y_x^{\neq}$ is the NN prototype of x in a different class. They are respectively defined as

$$y_x^{=} = \hat{a}, \hat{a} = \underset{\substack{a \in \mathcal{P} \\ class(a) = class(x)}}{\operatorname{argmin}} d(x, a),$$

$$y_x^{\neq} = \hat{b}, \hat{b} = \underset{\substack{b \in \mathcal{P} \\ class(b) \neq class(x)}}{\operatorname{argmin}} d(x, b). \tag{7}$$

For any sample $x \in X_c$, if $d(x, y_x^{=}) > d(x, y_x^{\neq})$, the value of loss function for classifying x is approximately equal to 0, i.e., x is correctly classified. On the contrary, a large loss of 1 means that x is misclassified. Therefore, the loss function J can be approximated as the estimate of classification error in the range of 0 to C.

Obviously, J must be differentiable with respect to all the involved parameters to be optimized via gradient descent method. As we know, the differentiable sigmoid function is reasonable as an approximation of the step function S_β. Therefore, we employ the sigmoid function with slope β centered at $z = 1$ to replace S_β in (8). Similarly, if $Q(x)$ is less than 1,x is accurately classified. We have

$$S_\beta(z) = \frac{1}{1 + e^{\beta(1-z)}}. \tag{8}$$

The derivative of $S_\beta(z)$ is given by

$$S'_\beta(z) = \frac{dS_\beta(z)}{dz} = \frac{\beta e^{\beta(1-z)}}{(1 + e^{\beta(1-z)})^2}. \tag{9}$$

Note that $S'_\beta(z)$ is a "windowing" function, which has its maximum for $z = 1$ and vanishes for $|z - 1| > 0$.

The block diagram of LRPLM is shown in Fig. 2. First, we construct the gallery image sets and extract their features (see Section 4.3). Second, the prototype sets and local metric matrices are initialized (see Section 4.2). Third, LRPLM learns the optimal prototype sets and local metric matrices simultaneously (see Section 3.4). Fourth, the label of the given probe set is obtained (see Section 3.5).

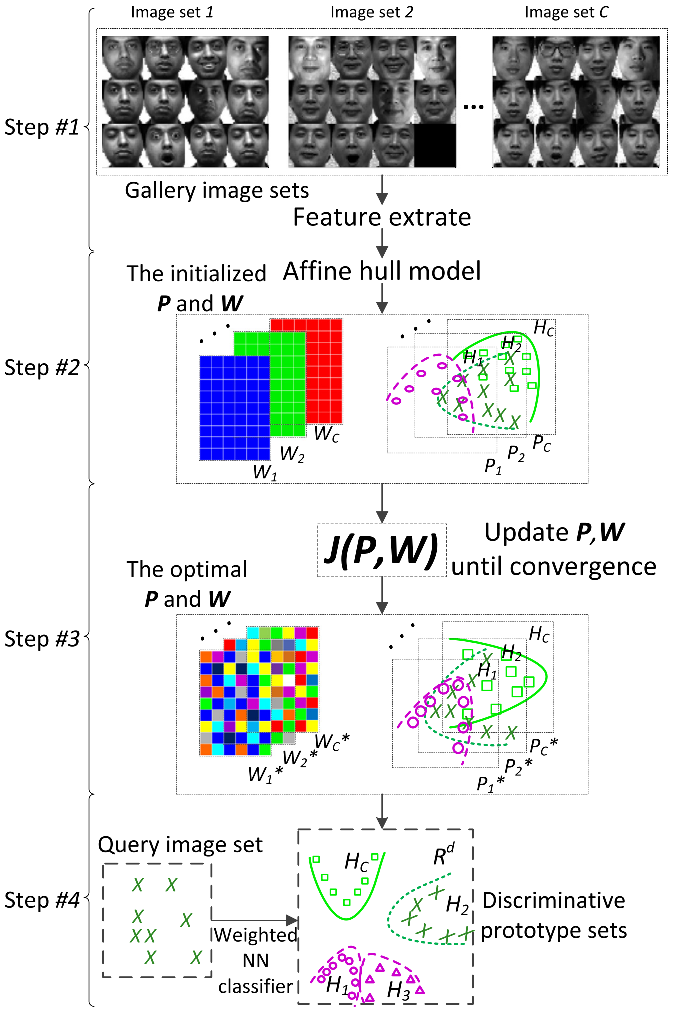


Fig. 2. Block diagram of LRPLM. It tunes both affine coefficients (positions) of prototypes and their corresponding local metric to obtain the optimal $\mathcal{P}^*$ and $\mathcal{W}^*$. By combining $\mathcal{P}^*$ and $\mathcal{W}^*$ together, the intersected image sets become more discriminative and descriptive so that the performance of image set classification can be boosted.

3.4. Solving and optimization

For learning the optimal prototype sets $\mathcal{P}^* = \{P_1^*, P_2^*, \cdots, P_C^*\}$ and their corresponding local metric matrices $\mathcal{W}^* = \{W_1^*, W_2^*, \cdots, W_C^*\}$, we need to optimize the following optimization problem:

$$J\begin{pmatrix} P_1^*, P_2^*, \ldots, P_C^* \\ W_1^*, W_2^*, \cdots, W_C^* \end{pmatrix} = \underset{\substack{P_1, P_2, \cdots, P_C \\ W_1, W_2, \cdots, W_C}}{\operatorname{argmin}} J(\mathcal{P}, \mathcal{W}) \tag{10}$$

To effectively solve problem (10) via gradient descent method, we first induce the partial derivatives of J with respect to $\mathcal{P}$ and $\mathcal{W}$. It should be noted that J depends on $\mathcal{P}$ and $\mathcal{W}$ in two different lines: 1) the weighted Euclidean distance $d(\cdot, \cdot)$, that depends directly on $\mathcal{P}$ and $\mathcal{W}$; 2) the nearest prototypes $y_x^{=}$ and $y_x^{\neq}$, for any $x \in X_c$, $y_x^{=}$ and $y_x^{\neq}$ always change with the prototype positions and their associated weights. To solve this optimization problem, we will assume that, for sufficiently small variations of the positions and weights, the prototype corresponding neighborhoods stay the same.

Generally, the derivative of w_{ij}^k can be computed by the chain rule, leading to the following derivative equation:

$$\begin{aligned}\frac{\partial J}{\partial w_{ij}^k} &= \sum_{c=1}^{C}\sum_{\forall x\in X_c}\frac{1}{n_c}\frac{\partial S_\beta(Q(x))}{\partial Q(x)}\frac{\partial Q(x)}{\partial w_{ij}^k}\\ &= \sum_{c=1}^{C}\sum_{\forall x\in X_c}\frac{1}{n_c}\frac{S'_\beta(Q(x))}{d^2(x,y_x^{=})}\frac{\partial d(x,y_x^{=})}{\partial w_{ij}^k}d(x,y_x^{\neq})\\ &\quad-\sum_{c=1}^{C}\sum_{\forall x\in X_c}\frac{1}{n_c}\frac{S'_\beta(Q(x))}{d^2(x,y_x^{\neq})}d(x,y_x^{=})\frac{\partial d(x,y_x^{\neq})}{\partial w_{ij}^k}.\end{aligned}\tag{11}$$

Moreover, we can easily get

$$\frac{\partial d(x,y_x^{=})}{\partial w_{ij}^k}=\frac{w_{ij}^k(x_j-y_{xj}^{=})^2}{d(x,y_x^{=})}\tag{12}$$

$$\frac{\partial d(x,y_x^{\neq})}{\partial w_{ij}}=\frac{w_{ij}^k(x_j-y_{xj}^{\neq})^2}{d(x,y_x^{\neq})},\tag{13}$$

where $k=class(y_x^{=}), i=index(y_x^{=})$. By substituting Eqs. (12) and (13) into Eq. (11), we have

$$\begin{aligned}\frac{\partial J}{\partial w_{ij}^k} &\approx \sum_{c=1}^{C}\sum_{\substack{\forall x\in X_c\\ class(y_x^{=})=k\\ index(y_x^{=})=i}}\frac{1}{n_c}S'_\beta(Q(x))Q(x)\frac{(x_j-y_{xj}^{=})^2}{d^2(x,y_x^{=})}w_{ij}^k\\ &\quad-\sum_{c=1}^{C}\sum_{\substack{\forall x\in X_c\\ class(y_x^{\neq})=k\\ index(y_x^{\neq})=i}}\frac{1}{n_c}S'_\beta(Q(x))Q(x)\frac{(x_j-y_{xj}^{\neq})^2}{d^2(x,y_x^{\neq})}w_{ij}^k.\end{aligned}\tag{14}$$

Similarly, the gradient of J with respect to each affine coefficient v_{ij}^k (local feature-dependent weight) is:

$$\begin{aligned}\frac{\partial J}{\partial v_{ij}^k} &= \sum_{c=1}^{C}\sum_{\forall x\in X_c}\frac{1}{n_c}\frac{S'_\beta(Q(x))}{d^2(x,y_x^{\neq})}\frac{\partial d(x,y_x^{=})}{\partial v_{ij}^k}d(x,y_x^{\neq})\\ &\quad-\sum_{c=1}^{C}\sum_{\forall x\in X_c}\frac{1}{n_c}\frac{S'_\beta(Q(x))}{d^2(x,y_x^{\neq})}d(x,y_x^{=})\frac{\partial d(x,y_x^{\neq})}{\partial v_{ij}^k}.\end{aligned}\tag{15}$$

We can also get

$$\frac{\partial d(x,y_x^{=})}{\partial v_{ij}^k}=\frac{\sum_{t=1}^{d}(w_{it}^k)^2(y_{xt}^{=}-x_t)U_{k,tj}^T}{d(x,y_x^{=})}\tag{16}$$

$$\frac{\partial d(x,y_x^{\neq})}{\partial v_{ij}^k}=\frac{\sum_{t=1}^{d}(w_{it}^k)^2(y_{xt}^{\neq}-x_t)U_{k,tj}^T}{d(x,y_x^{\neq})},\tag{17}$$

where $U_{k,t}^T$ is the tth column vector of the transpose of orthonormal basis matrix U_k. By substituting Eqs. (16) and (17) into Eq. (15), we have

$$\begin{aligned}\frac{\partial J}{\partial v_{ij}^k} &\approx \sum_{c=1}^{C}\sum_{\substack{\forall x\in X_c\\ class(y_x^{=})=k\\ index(y_x^{=})=i}}\frac{1}{n_c}S'_\beta(Q(x))Q(x)\frac{\sum_{t=1}^{d}(w_{it}^k)^2(y_{xt}^{=}-x_t)U_{k,tj}^T}{d^2(x,y_x^{=})}\\ &\quad-\sum_{c=1}^{C}\sum_{\substack{\forall x\in X_c\\ class(y_x^{\neq})=k\\ index(y_x^{\neq})=i}}\frac{1}{n_c}S'_\beta(Q(x))Q(x)\frac{\sum_{t=1}^{d}(w_{it}^k)^2(y_{xt}^{=}-x_t)U_{k,tj}^T}{d^2(x,y_x^{\neq})}.\end{aligned}\tag{18}$$

In Eqs. (14) and (18), the sum of key part is calculated by visiting each image $x\in X_c$ and updating the position of prototypes and their corresponding local weights, respectively. After getting these partial derivatives, the optimizations are performed using the corresponding gradient descent update equations from iteration t to $t+1$ respectively.

For the local metric $\mathcal{W}$, the update equation is

$$(w_{ij}^k)^{(t+1)}=(w_{ij}^k)^{(t)}-\mu_{ij}^k\frac{\partial J}{\partial w_{ij}^k}.\tag{19}$$

For the position of prototype in $\mathcal{P}$, the update equation is

$$(v_{ij}^k)^{(t+1)}=(v_{ij}^k)^{(t)}-\nu_{ij}^k\frac{\partial J}{\partial v_{ij}^k},\tag{20}$$

where μ_{ij}^k and ν_{ij}^k are the learning factors of local metric learning and prototype learning, respectively. These factors can take just fixed value for all i, j and k, or we can use some simple update rules to determine them. To simplify the calculation procedure, μ_{ij}^k and ν_{ij}^k are set fixed value in this paper. Note that larger value of ν emphasizes the learning of the prototypes themselves while larger value of μ gives more importance to the learning of the local metric for these prototypes. Specifically, we think that the significance of local metric is bigger than the influence of the prototype position. Thus, in this paper, μ_{ij}^k and ν_{ij}^k are immutable across iterations, s.t. $\nu<\mu$.

For further reducing memory consumption and improving iterative efficiency, we can iteratively adjust the position of each prototype and local metric by using L-BFGS[1] algorithm (Le et al., 2011) which is a limited-memory quasi-Newton code for unconstrained optimization.

3.5. Image set classification

To find the class label of a given probe image set X_Q, we first learn the optimal $\mathcal{P}^*=\{P_1^*,P_2^*,\cdots,P_C^*\}$ and $\mathcal{W}^*=\{W_1^*,W_2^*,\cdots,W_C^*\}$ from a collection of gallery image sets $\mathcal{X}=\{X_1,X_2,\cdots,X_C\}$ by using the proposed method. After the learning process, the similarity between the probe set and any one of gallery prototype sets can be defined as the minimal feature-wise weighted distance between a pair of points belonging to either set. Thus, the probe image set can be classified into the same class with its nearest gallery prototype set.

4. Experimental evaluation

In this section, we perform extensive experiments to evaluate the performance of the proposed method for the tasks of image set-based face recognition.

4.1. Dataset settings

Three widely used image set datasets are adopted, including Honda/UCSD (Lee, Ho, Yang, & Kriegman, 2003), CMU MoBo (Gross & Shi, 2001), YouTube Celebrities (YTC) (Ren, Wu, Zhang, & Sun, 2019). To be identical to the experimental settings in Yang et al. (2017, 2013), for each dataset, we evaluate all the compared methods by using three kinds of experiments with set length 50, 100, and 200, respectively. It should be noted that all images are used for classification if the number of frames in a set is fewer than the given set length.

4.2. Comparison methods

We compare the proposed method with a number of recently proposed image set classification methods. These state-of-art methods can be roughly divided into two categories, including affine hull-based methods and non-affine hull-based methods.

[1] http://users.iems.northwestern.edu/~nocedal/lbfgs.html.

(1) Affine hull based methods, e.g., the Linear version of the Affine Hull-based Image Set Distance (AHISD) (Cevikalp & Triggs, 2010), the Convex Hull-based Image Set Distance (CHISD) (Cevikalp & Triggs, 2010), Sparse Approximated Nearest Points (SANP) (Hu et al., 2012), Regularized Nearest Points (RNP) (Yang et al., 2013), Set to Set Distance Metric Learning (SSDML) (Zhu et al., 2013), Prototype Discriminative Learning(PDL) (Wang et al., 2017). Among these methods, AHISD, SANP, PDL and SSDML are affine hull-based methods, of which AHISD can be regarded as a baseline method to find the pair of nearest neighbor points between two affine hulls. Moreover, CHISD and RNP are convex hull and regularized affine hull based methods, respectively. For convenience, all of these methods can be considered as affine hull-based ones.

(2) Non-affine hull based methods, e.g., Discriminant Canonical Correlation Analysis (DCC) (Kim et al., 2007), Manifoldto-Manifold Distance (MMD) (Wang, Shan, Chen, & Gao, 2008), Manifold Discriminant Analysis (MDA) (Hamm & Lee, 2008), Covariance Discriminative Learning (CDL) (Wang et al., 2012a), Graph Embedding Discriminant Analysis (GEDA) (Harandi et al., 2011), Deep Reconstruction Models (DRM) (Hayat et al., 2015).

For all the compared methods, we adopt the implementations provided by the respective authors on their homepages. The important parameters for all the methods are carefully optimized following the same recommendations in the original references for best performance. Specifically, for all affine hull-based methods (i.e., AHISD, CHISD, SANP, RNP, SSDML, PDL and LRPLM), we preserve 90% energy when applying the thin SVD at the time of hull modeling. For LRPLM, the number of prototypes m_c per image set is set to the same value as the number of leading singular vectors, which are obtained by performing SVD with the centered data matrix. The initial prototype set P_c are initialized by using random unit vector to initialize each affine coefficient vector v, i.e., the ith prototype in P_C is $y_{ci} = U_c v_i$. The local metric W_c is initialized by using an all-ones matrix with the corresponding dimensions as P_c. The slope parameter β of sigmoid function is fixed to 10, and the two learning factors are set to $\mu = 0.01$ and $\nu = 0.001$, respectively.

4.3. Experiment analysis

4.3.1. Performance of the proposed LRPLM method on the Honda/UCSD Dataset

The Honda/UCSD (Lee et al., 2003) dataset contains 59 video sequences of 20 different subjects. Each sequence contains approximately 12–645 frames that contain expression variations, large poses, and different illuminations. The same protocol mentioned is adopted as that in Yang et al. (2017), and the face region of each frame is cropped and resized to 30 × 30 pixels. Fig. 3 shows some cropped images from this dataset. In our experiments, each video is considered as an image set, and each subject has 20 image sets as the gallery sets for training, leaving the remaining 39 sets as the probe sets for testing. To achieve a consistency in the results, the experiments are conducted for 10 times with different randomly picked dices of training and query image sets. The average recognition rates with standard deviations of all the compared methods are summarized in Table 1. It can be seen that the proposed LRPLM method consistently achieves the best performance under different number of training frames, i.e., has the highest classification accuracy and the smallest standard deviation. Especially when the set length is 200, the classification rate of LRPLM is up to 100%. Moreover, when the set length are 100 and 200, most of the recently proposed methods (i.e., SANP, RNP, DRM, PDL, and LRPLM) can achieve a high recognition rate larger than 90%. Compared to the similar method, PDL, the improvements of LRPLM are 2.4%, 0.9% and 1.7% when the set length are 50, 100 and 200, respectively. Compared with other methods, they also achieve good performance when there are enough samples in each image set. However, the nonlinear manifold based MMD method gets the lowest recognition rate when the number of frames is 50.

Table 1
Average recognition rates (%) and standard deviations (%) of the compared methods on the Honda/UCSD dataset.

Methods	Honda/UCSD		
	50 frames	100 frames	200 frames
DCC	77.2 ± 3.4	84.9 ± 2.6	93.9 ± 2.4
MMD	69.3 ± 4.5	86.8 ± 2.0	93.9 ± 2.2
MDA	81.9 ± 5.7	94.6 ± 1.5	97.2 ± 3.7
AHISD	87.2 ± 2.3	84.7 ± 3.5	88.9 ± 1.8
CHISD	82.1 ± 2.2	84.4 ± 1.8	92.8 ± 1.5
GEDA	83.1 ± 6.2	88.7 ± 8.7	91.8 ± 5.5
SANP	84.5 ± 2.8	91.9 ± 3.4	94.9 ± 3.1
RNP	87.2 ± 3.3	95.0 ± 3.1	100.0 ± 0.0
DRM	91.5 ± 2.7	95.8 ± 1.3	100.0 ± 0.0
SSDML	83.5 ± 1.6	84.4 ± 2.1	81.3 ± 3.3
PDL	89.9 ± 2.2	96.4 ± 2.3	98.3 ± 3.1
LRPLM	**92.3** ± **1.8**	**97.3** ± **1.1**	**100.0** ± **0.0**

4.3.2. Performance of the proposed LRPLM method on the CMU Mobo Dataset

The CMU MoBo (Gross & Shi, 2001) (Motion of Body) dataset contains 96 video sequences of 24 individuals on a treadmill in total. For each video, about 300 frames covering variations of poses and expressions are captured from different walking situations inclined slow, fast, inclined, and carrying a ball. As in Wang et al. (2008), face images from each frame are cropped and resized to 20 × 20 pixels. Some face images are shown in Fig. 4. For each individual, four sequences are captured, one sequence is randomly selected for training and the remaining three ones for testing. To achieve a consistency in the results, the experiments are also conducted for 10 times with different randomly selected

Fig. 3. Some typical images of the Honda/UCSD dataset.

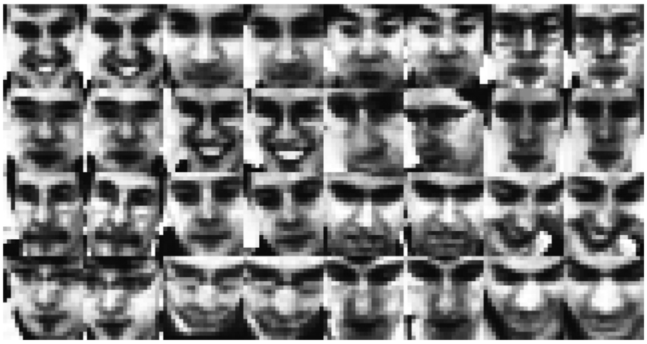


Fig. 4. Some typical images of the CMU Mobo dataset.

Table 2
Average recognition rates (%) with standard deviations (%) of the compared methods on the CMU Mobo dataset.

Methods	CMU Mobo		
	50 frames	100 frames	200 frames
DCC	81.9 ± 3.0	84.7 ± 2.7	90.9 ± 2.3
MMD	91.0 ± 2.5	93.7 ± 2.1	96.5 ± 0.8
MDA	85.9 ± 3.2	93.3 ± 2.6	95.4 ± 2.4
AHISD	90.9 ± 2.6	93.9 ± 2.1	91.8 ± 2.5
CHISD	91.3 ± 3.3	93.1 ± 2.6	97.6 ± 2.1
GEDA	86.7 ± 2.2	87.8 ± 3.2	93.3 ± 3.4
SANP	91.8 ± 3.1	94.7 ± 1.7	97.3 ± 1.3
RNP	91.9 ± 2.5	94.7 ± 1.2	97.4 ± 1.5
DRM	92.3 ± 2.1	96.2 ± 0.7	97.8 ± 1.6
SSDML	91.3 ± 3.3	95.1 ± 2.2	97.4 ± 2.5
PDL	94.2 ± 2.0	95.5 ± 1.7	96.5 ± 1.5
LRPLM	**94.3** ± **1.7**	**97.3** ± **0.7**	**98.4** ± **0.9**

Table 3
Average recognition rates (%) with standard deviations (%) of the compared methods on the YouTube Celebrities dataset.

Methods	YouTube Celebrities		
	50 frames	100 frames	200 frames
DCC	68.5 ± 2.9	73.7 ± 4.5	75.8 ± 2.2
MMD	68.9 ± 3.7	71.8 ± 4.4	76.1 ± 4.5
MDA	64.0 ± 4.1	74.3 ± 6.0	74.7 ± 4.6
AHISD	73.2 ± 5.5	72.8 ± 7.4	67.0 ± 4.5
CHISD	72.3 ± 5.6	73.3 ± 5.3	74.9 ± 5.0
GEDA	69.3 ± 3.1	73.8 ± 4.4	75.8 ± 3.3
SANP	73.3 ± 3.9	74.9 ± 5.9	78.3 ± 4.2
RNP	75.2 ± 5.4	75.4 ± 5.1	77.9 ± 5.5
DRM	70.1 ± 4.3	72.5 ± 4.7	75.4 ± 3.8
SSDML	68.3 ± 5.5	68.2 ± 5.2	72.5 ± 4.1
PDL	74.2 ± 3.3	75.3 ± 3.8	77.4 ± 3.1
LRPLM	**75.9** ± **2.5**	**76.2** ± **3.9**	**78.7** ± **2.6**

training and query image sets. The experimental results are shown in Table 2. It can be seen that the proposed LRPLM method achieves the highest average recognition rates and the lowest standard deviations under different number of training frames. Relative to PDL, the improvements of LRPLM are 1.8% and 1.9% when the set length are 100 and 200, respectively. Compared with the mostly recent deep learning method (i.e., DRM), the improvements of LRPLM are 0.8% and 1.5% when the set length are 50 and 100, respectively. Compared with other affine hull-based methods, LRPLM also shows relatively good performance.

4.3.3. Performance of the proposed LRPLM method on the YTC dataset

The YouTube-Celebrities (YTC) (Ren et al., 2019) dataset contains 1910 video clips of 47 celebrities (actors, politicians, and actresses) from YouTube. Most of the videos are low resolution, highly compression and large variations in the form of illuminations, poses, and expressions, which lead to noisy, low-quality image frames. The face images are cropped and resized to 30 × 30 pixels. Some face images are shown in Fig. 5. For performance evaluation, we used five-fold cross-validation experimental settings as followed in Wang et al. (2008) and Wang and Chen (2009). In each fold, we randomly selected nine image sets per subject, three of which are selected as the gallery sets for training and the remaining six are used as the query sets for testing. The experimental results are exhibited in Table 3. As we know, this dataset contains large appearance variations in illuminations, poses, and expressions. Moreover, faces images are not accurately cropped due to tracking errors in low quality videos. Therefore, compared with the Honda/UCSD and CMU Mobo datasets, the performances of all the comparison methods are relatively low than that on the YTC dataset. Compared with PDL, 1.7%, 0.9% and 1.3% improvements are achieved by the proposed LRPLM method when the set length are 50, 100 and 200, respectively. Similarly, LRPLM also outperforms RNP (which is the first regularized affine hull-based method) by 0.7%, 0.8% and 0.8% when the set length are 50, 100, and 200, respectively. Moreover, LRPLM is also visibly better than the deep learning method DRM.

Fig. 5. Some typical images of the YTC dataset.

In order to better demonstrate the output of LRPLM, some prototypes and their corresponding local metric of five different image sets from the YTC dataset are shown in Fig. 6a, b and c, respectively.

4.3.4. Discussion

From the results in Tables 1–3, we have the following observations.

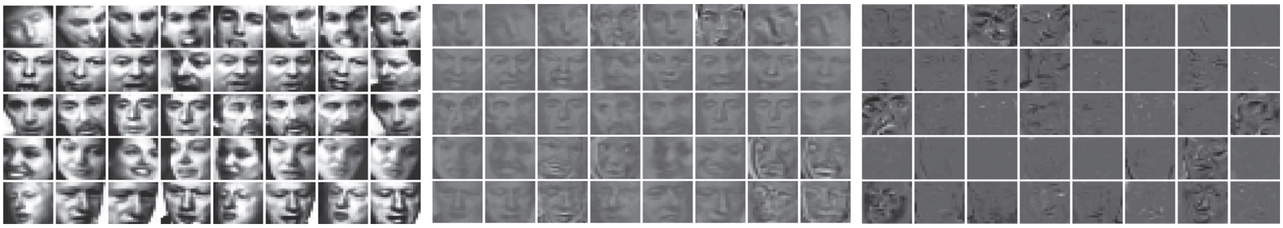
(1) The experimental results on three widely used datasets show that the performances of all the compared methods are influenced by set length. In this paper, the set length is varied from {50, 100, 200}.

(2) In general, the experimental results demonstrate that affine hull-based methods (e.g., CHISD, SANP, SSDML, RNP, PDL, and LRPLM) have higher accuracy and stability than non-affine hull-based methods under different set length settings. This suggests that affine hull model and its extended models (e.g., convex hull and regularized affine hull) are very effective modeling techniques to model image set. Notably, these non-affine hull-based methods, MSM, DCC, MMD, MDA, construct an image set by using either a linear subspace or a combination of multiple linear subspaces. With the reduction on set length, the performance of them degrade more heavily due to the reduction of discrimination information.

(3) Compared with the results of AHISD, CHISD and SANP, which all model image set similarly with PDL and our LRPLM, but do not consider learning a few prototypes to represent the whole set. From the results, it can be seen that PDL and our LRPLM achieve better performance than them. This indicates that using a reduced set of prototypes instead of the whole set can help promote the discriminative ability of image set.

(4) PDL learns a reduced set of prototypes and a global metric simultaneously for image set classification. Relative to PDL, the proposed LRPLM method learns a reduced set of prototypes and a local feature-wise metric for each gallery set simultaneously. This demonstrates that local feature-wise metric has the potential for obtaining better performance than global metric.

(5) Compared with the results of deep learning method DRM, LRPLM has gained better performance in terms of recognition rate and standard deviation. Moreover, the computational complexity of LRPLM is far less than the DRM.



(a) Some face images of five different individuals.

(b) Some prototypes of five different image sets.

(c) Some local metric matrices for these prototypes.

Fig. 6. Visualization of some prototypes and their corresponding local metric on the YTC dataset. Each row corresponds to an individual.

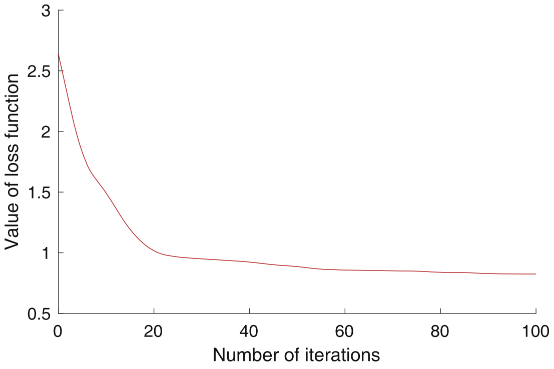


Fig. 7. Convergence of the proposed method on the YouTube Celebrities dataset.

Table 4
Execution times in seconds required for offline training and online testing of one probe set on the YTC dataset.

Methods	Training	Testing	Practicality
DCC	33.92	0.62	☆☆
MMD	N/A	38.11	☆
MDA	9.23	0.65	☆☆
AHISD	N/A	3.24	☆☆
CHISD	N/A	5.45	☆☆
GEDA	12.73	1.26	☆☆
SANP	N/A	42.40	☆
RNP	N/A	2.36	☆☆☆☆
DRM	522.34	1.52	☆☆☆
SSDML	356.58	8.35	☆☆
PDL	89.33	1.23	☆☆☆☆
LRPLM	113.7	0.67	☆☆☆☆☆

This illustrates the disadvantage of deep learning method on small-scale data sets.

4.4. Convergence analysis

To be identical to the experimental settings in Section 4.3.3, and the set length is set to 200, we conduct convergence analysis on the YTC dataset. Fig. 7 plots the convergence curve of LRPLM, which demonstrates that LRPLM achieves both faster and smoother convergence performance, usually within about 40 iterations. For space limitation, the converges of LRPLM on other datasets are not shown in this paper.

4.5. Computational time analysis

In this section, we report the average computational time of all the compared methods on the YTC dataset. Our experimental platform is MATLAB R2016b/G++ 4.2.1 under a MacBook Pro-2017 with an Intel Core i5 (2.3 GHz) CPU and 8G RAM. The implementation of LRPLM is written in c-language, other methods use MATLAB. Time in seconds required for offline training and online testing for 47 subjects (three sets per class, randomly selected 100 images per set) respectively are listed in Table 4, where ‘N/A’ means that the method does not require any offline training time. Form Table 4, the results of training time suggest that LRPLM requires more time than other affine hull-based methods (i.e., AHISD, CHISD, GEDA, SANP, RNP, SSDML and PDL). However, the training of our LRPLM is performed offline, so it can not be seen as a disadvantage. The testing of LRPLM is very efficient since it does not require matrix multiplication. More precisely, the training time and testing time of LRPLM are 113.7 s and 0.67 s, respectively. We can see that LRPLM is the fastest among all competing methods except for DCC and MDA in testing cost. Compared with PDL, LRPLM doesn’t need to compute the projections of all data, so the testing time is lower than PDL’s. It is worth noting that the computational time of the deep learning method DRM far exceeds other methods up to 522.34 s in training stage. Comprehensive analysis of the computational times and classification performance of all the compared methods, LRPLM gets five stars on practicality for the applications designed to real-time classification tasks.

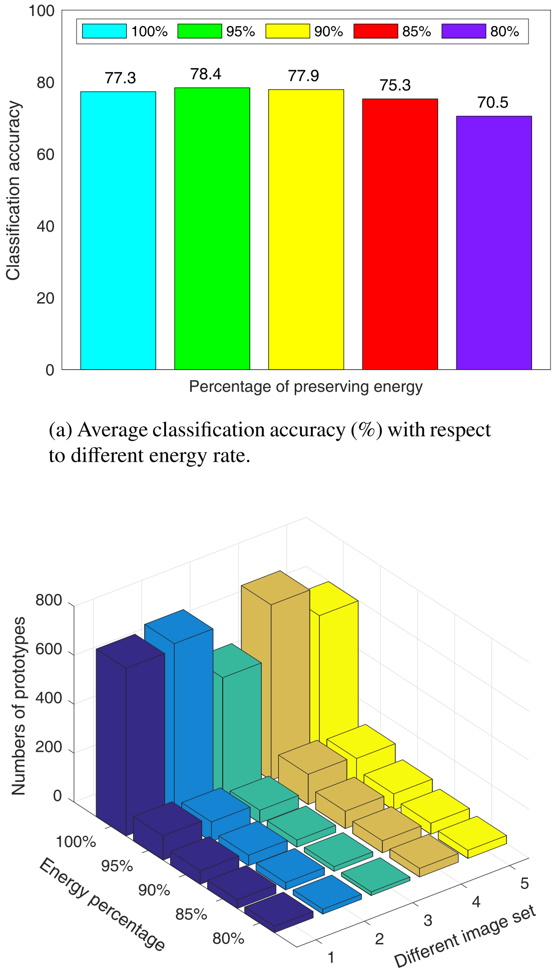
4.6. Parameters sensitivity

In this section, we study the influence of the percentage of preserving energy (*i.e.*, p) on classification accuracy and number of prototypes by tuning it from $p = \{100\%, 95\%, 90\%, 85\%, 80\%\}$. For our method, the number of prototypes m_c of an image set is the same as the number of leading singular values, which is obtained by performing SVD on the centered data matrix. Take the widely used YTC database for example, we randomly select 141 image sets (47 subjects from the YTC dataset, three sets per subject) for training and 200 image sets for testing.

Fig. 8a and b show the classification performance and the number of prototypes, respectively, of LRPLM with respect to different values of the parameter p. It can be seen that when the parameter p varies from 100% to 95%, the number of prototypes can be significantly reduced, but the accuracy is nearly unchanged. In addition, when the energy rate varies from {95%, 90%, 80%}, the accuracy is slightly decreased. It demonstrates that using fewer prototypes instead of the original whole image set will not affect the classification performance, but can reduce the storage and time costs while preventing over-fitting. All things considered, $p = 90\%$ is hold in our work.

5. Conclusion and future work

In this paper, we have proposed a novel method for image set classification, which is called Learning of Reduced Prototypes and Local Metric (LRPLM). It aims to significantly overcome the flaw of affine hull caused by too loose affine approximation. For each image set, a joint discriminative information, a reduced set of virtual



(a) Average classification accuracy (%) with respect to different energy rate.

(b) Number of prototypes per image set with respect to different energy rate.

Fig. 8. Parameter sensitivity analysis of LRPLM on the YouTube-Celebrities dataset, where the energy percentage parameter p is tuned from {100%, 95%, 90%, 85%, 80%}. Note: for convenience, we only shown the number of prototypes of the first five image sets for comparison.

prototypes and an optimal local feature-wise metric, are simultaneously learned. Therefore, LRPLM not only inherits the advantage of affine hull with better representation to account for the unseen appearances, but also significantly improves the discrimination ability of gallery sets. The promising results obtained on three benchmark datasets clearly demonstrate the superiority of LRMLM in terms of both accuracy and computation complexity.

Besides the contents presented in this paper, future work consists of: 1) introduce some regularization terms to avoid overfitting, e.g., ℓ_1-norm, ℓ_2-norm, and elastic net; 2) plan to extend LRPLM to the deep learning framework to boost its performance; 3) explore more classification discriminant functions to measure the classification loss; 4) apply LRPLM to other expert system applications, e.g., gesture recognition, object recognition, and scene classification.

Declaration of Competing Interest

All the authors declare that there are no conflicts of interest regarding the publication of this article.

Credit authorship contribution statement

Zhenwen Ren: Conceptualization, Methodology, Software, Visualization, Writing - original draft. **Bin Wu:** Supervision, Validation, Investigation. **Quansen Sun:** Writing - review & editing, Supervision. **Mingna Wu:** Data curation, Writing - original draft.

Acknowledgments

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

The authors would like to Wen Wang, who shared their PDL code on their homepage. The authors would also like to thank the anonymous reviewers who provided substantive suggestions for improving our work.

References

Affonso, C., Rossi, A. L. D., Vieira, F. H. A., & de Leon Ferreira, A. C. P. (2017). Deep learning for biological image classification. *Expert Systems with Applications, 85*, 114–122.

Behera, S. K., Dogra, D. P., & Roy, P. P. (2018). Fast recognition and verification of 3d air signatures using convex hulls. *Expert Systems with Applications, 100*, 106–119.

Cevikalp, H., & Triggs, B. (2010). Face recognition based on image sets. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on* (pp. 2567–2573). IEEE.

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.

Glowacz, A. (2018a). Acoustic-based fault diagnosis of commutator motor. *Electronics, 7*(11), 299.

Glowacz, A. (2018b). Recognition of acoustic signals of commutator motors. *Applied Sciences, 8*(12), 2630.

Gross, R., & Shi, J. (2001). The cmu motion of body (mobo) database,.

Hamm, J., & Lee, D. D. (2008). Grassmann discriminant analysis: A unifying view on subspace-based learning. In *Proceedings of the 25th international conference on machine learning* (pp. 376–383). ACM.

Hanmandlu, M., & Mamta. (2014). Robust authentication using the unconstrained infrared face images. *Expert Systems with Applications, 41*(14), 6494–6511.

Harandi, M. T., Sanderson, C., Shirazi, S., & Lovell, B. C. (2011). Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on* (pp. 2705–2712). IEEE.

Hayat, M., Bennamoun, M., & An, S. (2015). Deep reconstruction models for image set classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(4), 713–727.

Hu, Y., Mian, A. S., & Owens, R. (2012). Face recognition using sparse approximated nearest points between image sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(10), 1992–2004.

Karanam, S., Wu, Z., & Radke, R. J. (2018). Learning affine hull representations for multi-shot person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology, 28*(10), 2500–2512.

Kim, T.-K., Kittler, J., & Cipolla, R. (2007). Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(6), 1005–1018.

Köstinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2013). Joint learning of discriminative prototypes and large margin nearest neighbor classifiers. In *Computer vision (ICCV), 2013 IEEE international conference on* (pp. 3112–3119). IEEE.

Le, Q. V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., & Ng, A. Y. (2011). On optimization methods for deep learning. In *Proceedings of the 28th international conference on international conference on machine learning* (pp. 265–272). Omnipress.

Lee, K.-C., Ho, J., Yang, M.-H., & Kriegman, D. (2003). Video-based face recognition using probabilistic appearance manifolds. *Computer vision and pattern recognition, 2003. proceedings. 2003 IEEE computer society conference on*: 1. IEEE. I–I

Leng, M., Moutafis, P., & Kakadiaris, I. A. (2015). Joint prototype and metric learning for set-to-set matching: Application to biometrics. In *Biometrics theory, applications and systems (BTAS), 2015 IEEE 7th international conference on* (pp. 1–8). IEEE.

Liu, W., Xu, D., Tsang, I. W., & Zhang, W. (2019). Metric learning for multi-output tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 408–422.

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 Transactions on Image Processing, 22*(12), 5252–5262.

Paredes, R., & Vidal, E. (2006). Learning weighted metrics to minimize nearest-neighbor classification error. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(7), 1100–1110.

Ren, Z., Wu, B., Zhang, X., & Sun, Q. (2019). Image set classification using candidate sets selection and improved reverse training. *Neurocomputing, 341*, 60–69.

Shah, S. A. A., Bennamoun, M., & Boussaid, F. (2016). Iterative deep learning for image set based face and object recognition. *Neurocomputing, 174*, 866–874.

Shao, J., Huang, F., Yang, Q., & Luo, G. (2018). Robust prototype-based learning on data streams. *IEEE Transactions on Knowledge and Data Engineering, 30*(5), 978–991.

Soleymani, R., Granger, E., & Fumera, G. (2018). Progressive boosting for class imbalance and its application to face re-identification. *Expert Systems with Applications, 101*, 271–291.

Tan, C., Wahidin, L., Khalil, S., Tamaldin, N., Hu, J., & Rauterberg, G. (2016). The application of expert system: A review of research and applications. *ARPN Journal of Engineering and Applied Sciences, 11*(4), 2448–2453.

Wang, R., & Chen, X. (2009). Manifold discriminant analysis. In *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on* (pp. 429–436). IEEE.

Wang, R., Guo, H., Davis, L. S., & Dai, Q. (2012a). Covariance discriminative learning: A natural and efficient approach to image set classification. In *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on* (pp. 2496–2503). IEEE.

Wang, R., Shan, S., Chen, X., Dai, Q., & Gao, W. (2012b). Manifold–manifold distance and its application to face recognition with image sets. *IEEE Transactions on Image Processing, 21*(10), 4466–4479.

Wang, R., Shan, S., Chen, X., & Gao, W. (2008). Manifold-manifold distance with application to face recognition based on image set. In *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE conference on* (pp. 1–8). IEEE.

Wang, W., Wang, R., Shan, S., & Chen, X. (2015). Probabilistic nearest neighbor search for robust classification of face image sets. In *Automatic face and gesture recognition (fg), 2015 11th IEEE international conference and workshops on: 1* (pp. 1–7). IEEE.

Wang, W., Wang, R., Shan, S., & Chen, X. (2017). Prototype discriminative learning for image set classification. *IEEE Signal Processing Letters, 24*(9), 1318–1322.

Wu, Y., Minoh, M., & Mukunoki, M. (2013). Collaboratively regularized nearest points for set based recognition.. In *BMVC: 2* (p. 5).

Xuan, L., Wang, Z., Zhao, W., & Liu, Y. (2013). Image set classification based on low-rank representation. *Journal of Tongji University (Natural Science), 2*, 022.

Yang, M., Wang, X., Liu, W., & Shen, L. (2017). Joint regularized nearest points for image set based face recognition. *Image and Vision Computing, 58*, 47–60.

Yang, M., Zhu, P., Van Gool, L., & Zhang, L. (2013). Face recognition based on regularized nearest points between image sets. In *Automatic face and gesture recognition (fg), 2013 10th IEEE international conference and workshops on* (pp. 1–7). IEEE.

Zeng, S., Gou, J., & Deng, L. (2017). An antinoise sparse representation method for robust face recognition via joint l1 and l2 regularization. *Expert Systems with Applications, 82*, 1–9.

Zhang, H., Wang, F., Chen, Y., Zhang, W., Wang, K., & Liu, J. (2016). Sample pair based sparse representation classification for face recognition. *Expert Systems with Applications, 45*, 352–358.

Zhang, S., Qi, Y., Jiang, F., Lan, X., Yuen, P. C., & Zhou, H. (2018). Point-to-set distance metric learning on deep representations for visual tracking. *IEEE Transactions on Intelligent Transportation Systems, 19*(1), 187–198.

Zhu, P., Zhang, L., Zuo, W., & Zhang, D. (2013). From point to set: Extend the learning of distance metrics. In *Computer vision (iccv), 2013 IEEE international conference on* (pp. 2664–2671). IEEE.

Zhu, P., Zuo, W., Zhang, L., Shiu, S. C.-K., & Zhang, D. (2014). Image set-based collaborative representation for face recognition. *IEEE Transactions on Information Forensics and Security, 9*(7), 1120–1132.