Joint Regularized Nearest Points for Image Set based Face Recognition

Meng Yang[1], Weiyang Liu[2], and Linlin Shen[1]
[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[2] School of Electronic & Computer Engineering, Peking University

***Abstract*—Face recognition based on image set has attracted much attention due to its promising per-formance to overcome various variations. Recently, (collaborative) regularized nearest points (C)RNP has achieved the state-of-art performance by measuring the between-set distance as the distance between nearest points generated in each image set. However, the nearest point of the query set in RNP changes in computing its distance to nearest points of different gallery image sets, which may result in that a wrong gallery image set can also has a small between-set distance; CRNP used collaborative representation to overcome this issue but it doesn't explicitly minimize the between-set distance. In order to solve these issues and fully exploit the advantages of nearest point based approaches, in this paper a novel joint regularized nearest points (JRNP) is proposed for face recognition based on image sets. In JRNP, the nearest point in the query set keeps the same when computing its distance to the image sets of different classes; at the same time, it explicitly minimize the between-set distance of facial images. An efficient algorithm was proposed to solve this problem, and the classification is then based on the joint distance between the regularized nearest points in image sets. Extensive experiments on benchmark databases were conducted on benchmark databases (e.g., Honda/UCSD, CMU Mobo, and YouTube databases). The experimental results clearly show that our JRNP leads the performance in face recognition based on image sets.**

I. INTRODUCTION

Recognizing the objects of interest (e.g., human faces) is one of the most important and fundamental problems in the communities of computer vision and pattern recognition. Although face recognition (FR) has been extensively studied in the past several decades, the traditional face recognition usually assumes there is only a single query face image, from which a human identity is recognized. Although there are multiple images in the gallery set per subject, it is still a big challenge to correctly recognize a person's identity based on only a single query face image captured in less-controlled/uncontrolled environments due to different variations (e.g., lighting, expression, pose, disguise changes) existing in facial appearance images.

With the wide installation of video cameras and the developments of large-capacity-storage media, it becomes very convenient to collect multiple images from video sequences or photo albums for a known subject and store these images as the gallery and query image sets. Multiple face images in the query and gallery set for each subject incorporates more within-class appearance variations and provides richer information for face recognition. Compared to the traditional face recognition with a single query face image, face recognition based on image sets could achieve more satisfactory performance in practical face recognition applications and is more promising framework of face recognition.

Face recognition based on image sets has been attracting much attention from researchers over the past decades. The image sets could either be the consecutive video sequences with temporal information, or unordered photo album images collected from web at different times. Compared to video-based face recognition [1][20-23][38-39], face recognition based on general image sets, in which the temporal information is not available, has wider applications. In this paper we mainly focus on the face recognition problem based on general image sets. Numerous approaches have been proposed for this kind of image-set based recognition problem.

One major category of face recognition based on image set is the parameter model based approaches [39][24-25]. These parametric model based approaches [39][24-25] firstly represent each image set by some parametric distribution with the parameters estimated from the data itself, and then calculate the between-set distance by measuring the similarity between these two distributions (e.g., in terms of Kullback-Leibler divergence [37]). However, the parametric methods need to solve the difficult parameter estimation problem and require strong statistical correlations between the gallery and query sets, which may not exist in practice. To overcome the shortcomings of parameter model based approaches, recently Lu *et al.* [36] directly extracted the multiple order statistics features from the image set and developed a multi-kernel metric learning method to combine different order information.

In order to avoid the drawbacks of model-based methods, non-parametric model-free based approaches were proposed based on representing an image set as a convex/affine subspace [3][19][26-28], mixture of subspaces [29-31], or nonlinear manifolds [4][17][32-33]. In nonlinear-manifold methods, the manifold of an image set is usually represented as a combination of local linear subspaces [4][33]. In this model-free face recognition based on image sets, how to measure between-set distance is the key problem. A popular way is to define the between-set distance as the distance between two "exemplars" (e.g., the mean of samples) chosen from these two image sets. For instance, Cevikalp *et al.* [3] characterized each image set by an affine/convex hull spanned by its samples, and selected two points (one point in one hull) with the closest approach as the "exemplars". Another way of measuring the between-set distance for non-parametric approach is to compare the structure of the non-parametric model. For instance, Canonical correlation analysis (CCA) [9], which analyzes the principal angles and canonical correlations between linear subspaces, is widely used in the works of

978-1-4799-6026-2/15/$31.00 ©2015 IEEE

[4][19][26][27][28][30][31]. Besides, the natural second-order statistic-covariance matrix was used to represent each image set in [6], and the image-set based classification was formulated as classifying points lying on a Riemannian manifold.

Recently inspired the success of sparse representation on face recognition [5], Hu *et al.* [2] proposed a sparse approximated nearest points (SANP) approach for image-set based face recognition. By modeling each image set as an affine hull, Hu *et al.* selected two points (one point in each hull) with the closest distance as the sparse approximated nearest points (SANP), where SANPs were required to be sparsely represented by the original samples. The final between-set distance of SANPs is the result of multiplication of the distance between the found SANPs and the dimension of the affine hull. Although SANP has achieved a good performance, its model is a little complex. In order to improve it, Yang *et al.* [34] proposed a regularized nearest points (RNP) method, which modeled each image set as a regularized affine hull and use the regularized nearest points to measure the similarity of these two image sets. Following RNP and collaborative representation based classification [7], Wu *et al.* [35] find all the regularized nearest points simultaneously in the framework of collaborative representation. Although RNP and CRNP have shown promising performance, RNP finds different regularized nearest points in the query set when computing the distance between the query set and different gallery sets, easily resulting in over-fitting. CRNP didn't explicitly minimize the between-set distance, which reduces its discrimination, because the objective function of CRNP aims to minimize the distance between the query set and the entire gallery set.

This paper presents an efficient and effective joint regularized nearest points (JRNP) method for image-set-based face recognition. JRNP minimizes the joint representation model by finding a unique nearest point in the query set and explicitly minimizing the image set based between-class distance. An efficient algorithm was proposed to solve this problem, and the classification is then based on the joint distance between the regularized nearest points in image sets. Compared to RNP, the nearest point in the query set keeps the same for different gallery image sets to avoid over-fitting. Different from CRNP, JRNP explicitly minimize the between-set distance to enhance the discrimination of the model. Our experiments on benchmark image set databases clearly show that JRNP has achieved better recognition accuracy than the previous methods, including SANP, RNP and CRNP. Meanwhile, the proposed RNP also has a very fast speed; e.g., it is over 14 times faster than SANP in the CMU Mobo database [15].

The rest of this paper is organized as follows. Section II briefly reviews the RNP method in [34] and CRNP method in [35]. Section III presents the proposed JRNP. Section IV conducts experiments and Section V concludes the paper.

II. RELATED WORK

In this section, two related work, regularized nearest points (RNP) [34] and collaboratively regularized nearest points (CRNP) [35], are reviewed. Both of them use the nearest points generated by each face images set to compute the distance between different subjects.

Denote $\boldsymbol{X}_i = \left[\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \cdots, \boldsymbol{x}_{i,n_i} \right]$ as the data matrix of i-th class, n_i is the image number of i-th class, and $\boldsymbol{x}_{i,k}$ is the k-th feature vector of i-th class. Let $\boldsymbol{Y} = \left[\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{n_y} \right]$ be the data matrix of the query class, with $\boldsymbol{y}_k$ as its k-th feature vector. RNP computed the nearest points class by class. For instance, the coding phase for i-th class and the query set is

$$\min_{\boldsymbol{\alpha}_i, \boldsymbol{\beta}} \left\| \boldsymbol{X}_i \boldsymbol{\alpha}_i - \boldsymbol{Y} \boldsymbol{\beta} \right\|_2^2 \text{ s.t. } \sum\nolimits_k \alpha_{i,k} = 1, \sum\nolimits_k \beta_k = 1, \left\| \boldsymbol{\alpha}_i \right\|_2 \le \sigma_1, \left\| \boldsymbol{\beta} \right\|_2 \le \sigma_2 \tag{1}$$

where $\alpha_{i,k}$ is the k-th entry of $\boldsymbol{\alpha}_i$, $\boldsymbol{X}_i\boldsymbol{\alpha}_i$ and $\boldsymbol{Y}\boldsymbol{\beta}$ are the generated nearest points based on $\boldsymbol{X}_i$ and $\boldsymbol{Y}$, respectively. With the solved coding coefficients $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}$, then the distance between these two sets is computed as

$$e_i = \left(\left\| \boldsymbol{X}_i \right\|_* + \left\| \boldsymbol{Y} \right\|_* \right) \cdot \left\| \boldsymbol{X}_i \boldsymbol{\alpha}_i - \boldsymbol{Y} \boldsymbol{\beta} \right\|_2^2 \tag{2}$$

where $\|\boldsymbol{X}\|_*$ is the nuclear norm of $\boldsymbol{X}$, i.e., the sum of the singular values of $\boldsymbol{X}$.

Opposite to RNP, CRNP computed the virtual samples for all classes at the same time

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \left\| \boldsymbol{Y} \boldsymbol{\beta} - \boldsymbol{X} \boldsymbol{\alpha} \right\|_2^2 + \lambda_1 \left\| \boldsymbol{\alpha} \right\|_2^2 + \lambda_2 \left\| \boldsymbol{\beta} \right\|_2^2 \text{ s.t. } \sum\nolimits_k \beta_k = 1, \sum\nolimits_i \sum\nolimits_k \alpha_{i,k} = 1 \tag{3}$$

where $\boldsymbol{X}$=[$\boldsymbol{X}_1$,$\boldsymbol{X}_2$,…,$\boldsymbol{X}_c$] and $\boldsymbol{\alpha}$=[$\boldsymbol{\alpha}_1$;$\boldsymbol{\alpha}_2$;…;$\boldsymbol{\alpha}_c$], where $\boldsymbol{\alpha}_i$ is the sub-coefficient vector associated to $\boldsymbol{X}_i$. Compared to RNP, the nearest point for each class (e.g., $\boldsymbol{X}_i\boldsymbol{\alpha}_i$ for i-th class) and the nearest point in the query set (i.e., $\boldsymbol{Y}\boldsymbol{\beta}$), are computed simultaneously. After solving the coding coefficients, $\boldsymbol{\alpha}_i$ (i=1,…,c) and $\boldsymbol{\beta}$, CRNP computed the between-class distance as

$$e_i = \left(\left\| \boldsymbol{X}_i \right\|_* + \left\| \boldsymbol{Y} \right\|_* \right) \cdot \left\| \boldsymbol{X}_i \boldsymbol{\alpha}_i - \boldsymbol{Y} \boldsymbol{\beta} \right\|_2^2 \Big/ \left\| \boldsymbol{\alpha}_i \right\|_2^2 \tag{4}$$

Although RNP and CRNP have achieved promising results on face recognition based on image sets, there are several issues to be further considered.

The nearest point of the query set in RNP changes when computing the between-set distance of query set to the gallery sets of different classes. Since face images have small between-class variance and big within-class variance, the distance measured by RNP is easily over-fitting, e.g., the between-set distance of wrong classes could also be small.

CRNP inherits some merits of collaborative representation based classification [7], e.g., the across-class collaboration in representation and competing of different classes in classification. However, compared to RNP, CRNP didn't explicitly minimize the distance between the query set and the gallery set of different classes. This would reduce the discrimination ability of CRNP.

III. JOINT REGULARIZED NEAREST POINT MODEL

In order to overcome the shortcomings of RNP and CRNP, we proposed a joint regularized nearest points (JRNP) to preserve the advantages of RNP and CRNP, and overcome their shortcomings at the same time. In this section, we first

present the model of the proposed joint regularized nearest points (JRNP). Then we describe the solving algorithm and classification of JRNP. Finally the time complexity of the proposed JRNP is presented.

A. Model of JRNP

Given a query set $\boldsymbol{Y}$ and the entire training dataset $\boldsymbol{X}=[\boldsymbol{X}_1, \boldsymbol{X}_2,\ldots, \boldsymbol{X}_c]$, we want to find a so-called regularized nearest point in $\boldsymbol{Y}$, i.e., $\boldsymbol{Y\beta}$, which is close to not only the subspace built by $\boldsymbol{X}$ but also to the subspace of the unknown correct class. Here the requirement that $\boldsymbol{Y\beta}$ is close to a point in the subspace built by $\boldsymbol{X}$ aims to introduce the competition between different gallery classes; while it could enhance the discrimination by making $\boldsymbol{Y\beta}$ close to a nearest point generated by image set of the unknown correct class. Suppose the correct class label is i, our goal could be formulated as

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}_i} \left\{ \begin{array}{l} \|\boldsymbol{Y\beta}-\boldsymbol{X\alpha}\|_2^2+\lambda_1\|\boldsymbol{\alpha}\|_{l_p}+\lambda_2\|\boldsymbol{\beta}\|_{l_p} \\ +\|\boldsymbol{Y\beta}-\boldsymbol{X}_i\boldsymbol{\gamma}_i\|_2^2+\lambda_1\|\boldsymbol{\gamma}_i\|_{l_p} \end{array} \right\} \text{s.t.} \sum\nolimits_k \beta_k=1 \qquad (5)$$

where λ_1 and λ_2 are scalar parameters, the l_p-norm regularization terms (e.g., $\|\boldsymbol{\alpha}\|_{l_p}$, $\|\boldsymbol{\gamma}_i\|_{l_p}$ and $\|\boldsymbol{\beta}\|_{l_p}$) could make the coding more stable by suppressing unnecessary samples' contribution to the representation, and the affine hull constraint (e.g., $\sum_k \beta_k=1$) could avoid the trivial solution (e.g., $\boldsymbol{\alpha}=\boldsymbol{\beta}=\boldsymbol{0}$). Here we don't require $\sum_k \alpha_k=1$ and $\sum_k \gamma_{i,k}=1$ to reduce the complexity of coding because the constraint $\sum_k \beta_k=1$ and the minimization of $\|\boldsymbol{Y\beta}-\boldsymbol{X\alpha}\|_2$ and $\|\boldsymbol{Y\beta}-\boldsymbol{X}_i\boldsymbol{\gamma}_i\|_2$ could avoid $\boldsymbol{\alpha}=\boldsymbol{0}$ and $\boldsymbol{\gamma}_i=\boldsymbol{0}$.

However, the model of Eq. (5) is too ideal to be used in testing phase because the label of the correct class is unknown. Thus we proposed a practical joint regularized nearest points (JRNP) model

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}_i} \left\{ \begin{array}{l} \|\boldsymbol{Y\beta}-\boldsymbol{X\alpha}\|_2^2+\lambda_1\|\boldsymbol{\alpha}\|_{l_p}+\lambda_2\|\boldsymbol{\beta}\|_{l_p} \\ +\sum_{i=1}^{c} w_i\left(\|\boldsymbol{Y\beta}-\boldsymbol{X}_i\boldsymbol{\gamma}_i\|_2^2+\lambda_1\|\boldsymbol{\gamma}_i\|_{l_p}\right) \end{array} \right\} \text{s.t.} \sum_k \beta_k=1 \qquad (6)$$

where we use an estimated weight value w_i to indicate the focus on the minimization of $\|\boldsymbol{Y\beta}-\boldsymbol{X}_i\boldsymbol{\gamma}_i\|_2$. An illustration of the proposed JRNP (i.e., Eq.(6)) is shown in Fig.1. We can observe that the nearest point in the query set, $\boldsymbol{Y\beta}$, keeps the same for different gallery subjects, which could avoid the possible overfitting in RNP. Meanwhile, compared to CRNP, the weight would make $\boldsymbol{Y\beta}$ closer to the nearest point of estimated correct class. Especially when the correct class weight value is close to 1, the model of JRNP will approach to the ideal model of Eq. (5), and the advantage of RNP and CRNP could be utilized.

From Eq. (6), we can observe that only $\boldsymbol{\beta}$ is directly affected by the weight value $\boldsymbol{w}_i$. The weight value w_i is used to control the contribution of each subject to update $\boldsymbol{\beta}$. Here we simply require $\sum_i w_i=1$ to make a balance between collaborative representation (i.e., $\|\boldsymbol{Y\beta}-\boldsymbol{X\alpha}\|_2^2$) and subspace representation (e.g., $\|\boldsymbol{Y\beta}-\boldsymbol{X}_i\boldsymbol{\gamma}_i\|_2^2$). In ideal case, the correct subject will have a big weight value while the wrong subjects have low or near-zero weight values, as shown in Fig.1. This weight value could be updated in the optimization of JRNP.

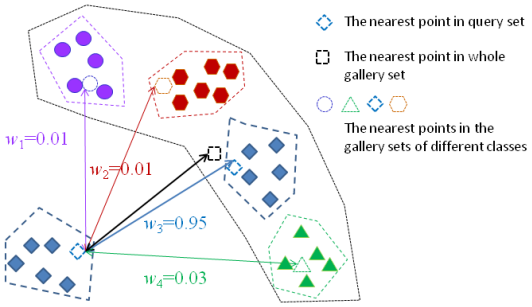

Fig. 1. An illustration of the proposed joint regularized nearest point model.

According to the number of samples in the probe image set, the proposed JRNP could change to a special case, joint collaborative representation based classifier, when the probe image set only has one sample. Denote $\boldsymbol{y}$ as the probe image, we could get β=1 and the proposed JRNP degenerates to

$$\min_{\boldsymbol{\alpha},\boldsymbol{\gamma}_i} \|\boldsymbol{y}-\boldsymbol{X\alpha}\|_2^2+\lambda\|\boldsymbol{\alpha}\|_{l_p}+\sum\nolimits_{i=1}^{c} w_i\left(\|\boldsymbol{y}-\boldsymbol{X}_i\boldsymbol{\gamma}_i\|_2^2+\lambda\|\boldsymbol{\gamma}_i\|_{l_p}\right) \qquad (7)$$

which is a joint version of collaborative representation based classifier [7] and nearest subspace classifier [8].

B. Solving algorithm of JRNP

The proposed JRNP model has various instantiations by applying different norms to the representation coefficients. More specially, when p=0 or 1, JRNP is regularized by l_0/l_1-norm sparse constraint; when p=2, l_2-norm regularization is applied to the representation coefficients. Some other constraints (e.g., non-negative constraint) could also be additively imposed on the representation coefficients. In this paper, we prefer to focus on a special instantiation of JRNP with p=2 since high recognition accuracy and fast speed could be both achieved.

Given known weight value w_i, we present how to solve the coding problem in Eq. (6) where $\sum_k \beta_k$=1is a linear equation. By relaxing it as $\sum_k \beta_k \approx 1$, it is easy to integrate the linear constraint into the objective function of Eq. (6). Thus Eq. (6) with p=2 could be rewritten as

$$\begin{array}{l} \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}_i} \|\boldsymbol{z}-\bar{\boldsymbol{Y}}\boldsymbol{\beta}-\bar{\boldsymbol{X}}\boldsymbol{\alpha}\|_2^2+\lambda_1\|\boldsymbol{\alpha}\|_2^2+\lambda_2\|\boldsymbol{\beta}\|_2^2 \\ \quad +\sum_{i=1}^{c} w_i\left(\|\boldsymbol{z}-\bar{\boldsymbol{X}}_i\boldsymbol{\gamma}_i-\bar{\boldsymbol{Y}}\boldsymbol{\beta}\|_2^2+\lambda_1\|\boldsymbol{\gamma}_i\|_2^2\right) \end{array} \qquad (8)$$

where $\boldsymbol{z}=[\boldsymbol{0};1]$, $\bar{\boldsymbol{Y}}=\left[-\boldsymbol{Y};\boldsymbol{1}^T\right]$, $\bar{\boldsymbol{X}}=\left[\boldsymbol{X};\boldsymbol{0}^T\right]$, $\bar{\boldsymbol{X}}_i=\left[\boldsymbol{X}_i;\boldsymbol{0}\right]$ and the two column vectors, $\boldsymbol{0}$ and $\boldsymbol{1}$, have appropriate sizes based on the context. In fact, the l_2-norm regularized model of Eq. (8) is the Ridge Regression, which is also a shrinkage method as the l_1-norm regularized sparse coding (i.e., Lasso) [10]. Since Eq. (8) is based on l_2-norm, it has a closed-form solution. However, this analytic solution needs several calculations of matrix inverse operation in online testing, which has a big computation burden.

For the face recognition problem based on image sets, Eq. (8) could be solved very efficiently by alternatively calculating $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}_i$, and $\boldsymbol{\beta}$.

When $\boldsymbol{\beta}$ is fixed, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}_i$ could be updated as

$$\boldsymbol{\alpha} = \boldsymbol{P}\left(z - \bar{\boldsymbol{Y}}\boldsymbol{\beta}\right) \tag{9}$$

and

$$\boldsymbol{\gamma}_i = \boldsymbol{P}_i\left(z - \bar{\boldsymbol{Y}}\boldsymbol{\beta}\right) \tag{10}$$

where the projection matrices are defined as $\boldsymbol{P} = \left(\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}} + \lambda_1\boldsymbol{I}\right)^{-1}\bar{\boldsymbol{X}}^T$ and $\boldsymbol{P}_i = \left(\bar{\boldsymbol{X}}_i^T\bar{\boldsymbol{X}}_i + \lambda_1\boldsymbol{I}\right)^{-1}\bar{\boldsymbol{X}}_i^T$.

When w_i, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}_i$ is fixed, the coding vector $\boldsymbol{\beta}$ is updated as

$$\boldsymbol{\beta} = \boldsymbol{Q}\left\{\left(z - \bar{\boldsymbol{X}}\boldsymbol{\alpha}\right) + \sum_{i=1}^{c} w_i\left(z - \bar{\boldsymbol{X}}_i\boldsymbol{\gamma}_i\right)\right\} \tag{11}$$

where the project matrix is computed as $\boldsymbol{Q} = \left(2\bar{\boldsymbol{Y}}^T\bar{\boldsymbol{Y}} + \lambda_2\boldsymbol{I}\right)^{-1}\bar{\boldsymbol{Y}}^T$.

The algorithm of JRNP with p=2 and known weight values is summarized in Table 1. Here we initialize $\boldsymbol{\beta}_0$=1/n, where n is the number of the query samples. It is easy to see that the cost function of Eq. (8) is lower bounded (≥0) and jointly convex to the variables $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}_i$, and $\boldsymbol{\beta}$. Since each step in the algorithm of JRNP will decrease the cost function value, the proposed algorithm to solve Eq.(8) will converge, as shown in Fig. 2.

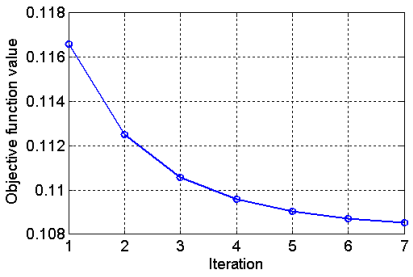

Fig. 2. An example of the convergence of JRNP with known weight values in the YouTube Celebrities Dataset.

Table 1. Algorithm of Joint Regularized Nearest Points

Algorithm 1: Algorithm of Joint Regularized Nearest Points (JRNP)

Input: Projection matrices $\boldsymbol{Q}$, $\boldsymbol{P}_i$ and $\boldsymbol{P}$, data matrices $\bar{\boldsymbol{X}}_i$ and $\bar{\boldsymbol{Y}}$, a column vector $\boldsymbol{z}$, weight value w_i, and an initialization of $\boldsymbol{\beta}^0$.

While not converged **do**

Compute the representation coefficients:

$\boldsymbol{\alpha}^{t+1} = \boldsymbol{P}\left(z - \bar{\boldsymbol{Y}}\boldsymbol{\beta}^t\right)$

$\boldsymbol{\gamma}_i^{t+1} = \boldsymbol{P}_i\left(z - \bar{\boldsymbol{Y}}\boldsymbol{\beta}^{t+1}\right)$

$\boldsymbol{\beta}^{t+1} = \boldsymbol{Q}\left\{\left(z - \bar{\boldsymbol{X}}\boldsymbol{\alpha}^{t+1}\right) + \sum_{i=1}^{c} w_i\left(z - \bar{\boldsymbol{X}}_i\boldsymbol{\gamma}_i^{t+1}\right)\right\}$

t=t+1

End while

Output: representation coefficients $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\gamma}}_i$ and $\hat{\boldsymbol{\beta}}$.

With the coding coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}_i$, and $\boldsymbol{\beta}$ known, we show how to estimate the weight value w_i. Since we want to introduce the competition between different gallery classes, we update w_i according to the distance between $\boldsymbol{Y}\boldsymbol{\beta}$ and the collaborative representation vector $\boldsymbol{\alpha}$=[$\boldsymbol{\alpha}_1$,$\boldsymbol{\alpha}_2$,…,$\boldsymbol{\alpha}_c$] for each gallery glasses. In this paper we update the weight value as

$$w_i = \exp\left(-\mu e_i/\bar{e}\right)\Big/\sum_i \exp\left(-\mu e_i/\bar{e}\right) \tag{12}$$

where $e_i = \left(\|\boldsymbol{X}_i\|_* + \|\boldsymbol{Y}\|_*\right)\cdot\|\boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{Y}\boldsymbol{\beta}\|_2^2\big/\|\boldsymbol{\alpha}_i\|_2^2$, μ is a scalar parameter, and $\bar{e}$ is the mean value of e_1, e_2, …, e_c (the detailed meaning of e_i will be illustrated in subsection D). Here $\|\boldsymbol{X}_i\|_*$ (i.e., nuclear norm of $\boldsymbol{X}_i$) is the sum of the singular values: $\|\boldsymbol{X}_i\|_* = \sum_k \sigma_k(\boldsymbol{X}_i)$, and $\|\boldsymbol{X}_i\boldsymbol{\alpha} - \boldsymbol{Y}\boldsymbol{\beta}\|_2^2$ represents the Euclidean distance between these two regularized nearest points in the query set and a gallery set. We initialize the weight values before algorithm of JRNP. In the iteration of JRNP, we also introduce several updates of weight values since some improvements could be achieved.

C. Classifier of JRNP

With the solved representation coefficients $\hat{\boldsymbol{\gamma}}_i$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, the between-set distance of JRNP is defined as

$$e_i = \left(\|\boldsymbol{X}_i\|_* + \|\boldsymbol{Y}\|_*\right)\cdot\left(\left\|\boldsymbol{X}_i\hat{\boldsymbol{\alpha}}_i - \boldsymbol{Y}\hat{\boldsymbol{\beta}}\right\|_2^2 + \left\|\boldsymbol{X}_i\hat{\boldsymbol{\gamma}}_i - \boldsymbol{Y}\hat{\boldsymbol{\beta}}\right\|_2^2\right)\Big/\|\hat{\boldsymbol{\alpha}}_i\|_2^2 \tag{13}$$

The term $||\boldsymbol{X}_i||_*$+$||\boldsymbol{Y}||_*$ in Eq. (13) aims to remove the disturbance unrelated to the class information. One example is that a wrong class, which has much more samples than the correct class, will have a lower value of $\left\|\boldsymbol{X}_i\hat{\boldsymbol{\alpha}} - \boldsymbol{Y}\hat{\boldsymbol{\beta}}\right\|_2^2 + \left\|\boldsymbol{X}_i\hat{\boldsymbol{\gamma}}_i - \boldsymbol{Y}\hat{\boldsymbol{\beta}}\right\|_2^2$. $||\boldsymbol{X}||_*$ is the convex relaxation of the rank of matrix $\boldsymbol{X}$, which could reflect the representation ability of the image set $\boldsymbol{X}$ (in our paper each column vector of $\boldsymbol{X}$ is normalized to have unit l_2-norm energy). Compared to RNP and CRNP, the proposed e_i jointly considers the distance measurements of RNP and CRNP, which could both use their advantages.

The identity of the probe image set $\boldsymbol{Y}$ is decided by

$$\text{identity}\left(\boldsymbol{Y}\right) = \arg\min_i\left\{e_i\right\} \tag{14}$$

D. Complexity analysis

In this section, we analyze the time complexity of the proposed JRNP. For the query image set $\boldsymbol{Y}$, all the projection matrices of $\boldsymbol{P}$ and $\boldsymbol{P}_i$ for all gallery sets could be computed offline. The computing of $\boldsymbol{Q}$ in Eq. (11) involves a matrix inverse, whose time complexity is roughly equal to the calculation of SVD of $\boldsymbol{Y}$ in SANP [2] (SANP needs to compute the orthonormal bases via SVD). Thus the calculation of $\boldsymbol{Q}$ in JRNP approximately has a complexity of O_{svd}.

Let m be the feature dimensionality and n_i be the number of gallery samples for the i-th class. In the algorithm of JRNP, the updating of $\boldsymbol{\alpha}$ has a time complexity of $O(lm(\sum_i n_i+n_y))$; the updating of $\boldsymbol{\gamma}_i$, i=1,2,…,c has a time complexity of $O(lm(\sum_i n_i+n_y))$; and the updating of $\boldsymbol{\beta}$ has a time complexity of $O(lm(2\sum_i n_i+n_y))$, where l is the iteration number. Consider the updating of w_i, the total time complexity is approximately $O(lm(5\sum_i n_i+4n_y))+O_{svd}$.

CRNP and RNP are the fastest methods in face recognition based on image set. We argue that the time complexity of

JRNP is less than the sum of CRNP and RNP. In algorithm, the updating of α_i, γ_i, i=1,2,…,c, and β has similar time complexity to these of CRNP and RNP. However, both RNP and CRNP need to compute the inverse of the projection matrix (e.g., Q) for β. Since CRNP and RNP need to compute this matrix inverse twice in total, the time complexity of JRNP is less than the sum of RNP's and CRNP's time complexity.

As the analysis in RNP [34], the time complexity of SNAP is roughly $O_{svd}+\Sigma_i O(m^2(n_i+n_y)^{\varepsilon})$, $\varepsilon \geq 1.2$[11][12][13]. Since the iteration number of JRNP is usually very small (e.g., l=5<<m=900 in YouTube), JRNP has less computational burden than SANP [2].

IV. Experimental results

We perform experiments on benchmark image-set face databases to demonstrate the effectiveness of JRNP. We first discuss the experimental setup in Subsection A. In Subsection B, we evaluate JRNP on three benchmark datasets, followed by the running time comparison in Subsection C. In this paper, if there is no specific instruction, the parameters of JRNP is fixed as λ_1=1e-1, λ_2=1e-3 and μ=10 for all the experiments.

A. Experimental setup

Three benchmark image set databases, including Honda/UCSD [39], CMU Mobo [15], and YouTube Celebrities [16] datasets, are used to evaluate the proposed JRNP. All the face images in the three datasets were detected using Viola and Jone's face detector [14]. For Honda/UCSD and YouTube datasets, after histogram equalization the face images were resized to 20×20 and 30×30, respectively; and the feature of raw pixel values for each image was directly used as the column vector of the data matrix. For CMU Mobo dataset, the histogram of LBP feature [18] was extracted as the facial feature. For each dataset, three kinds of experiments with the frame number 50, 100 and 200 are conducted, 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 frame number.

The proposed JRNP is compared with several recent image set classification methods, of which the Discriminant Canonical Correlations (DCC) [19] and Mutual Subspace Method (MSM) [28] are linear subapce based methods; Manifold-Manifold Distance (MMD) [4] and Manifold Discriminant Analysis (MDA) [33] are nonlinear manifold based methods; and Affine Hull based Image Set Distance (AHISD) [3], Convex Hull based Image Set Distance (CHISD) [3], Regularized Nearest Points (RNP) [34], Collaboratively Regularized Nearest Points (CRNP) [35] and Sparse Approximated Nearest Point (SANP) [2] are affine subspace based methods. All the competing methods were implemented using the source codes provided by the authors, with the parameters tuned according to the recommendations in the original references. For AHISD, CHISA and SANP, we used their linear versions since we didn't consider the kernel version of JRNP in this paper. In Honda/UCSD and CMU Mobo datasets, there is a single training image set for each class. Thus following the setting of [19] each single training image set for DCC was randomly divided into two subsets to construct the within-class sets.

B. Results and analysis

Honda/UCSD Dataset

It contains 59 video sequences of 20 different subjects in Honda/UCSD Dataset [39]. For each subject, different poses and expressions appear across different sequences. As the experimental setting of [39][2][34], we use 20 sequences for training, with the remaining sequences for testing.

The recognition results using different number of training frames are reported in Table 2. Here λ_1=0.001 and λ_2=0.1. We can clearly see that the proposed JRNP could significantly outperform all the other methods, including the state-of-the-art methods like SANP, RNP and CRNP. For instance, JRNP correctly recognize all query sets when the set length is 100 and 200; when the set length is 50, JRNP outperforms the second best, CRNP, by over 2%. JRNP even substantially outperforms the kernel version of SANP (e.g., 87.18% with set length as 50 and 94.87% with set length as 100). When image samples in each image set are enough, good performance could be achieved by all the methods, except MSM, which usually gets the worst result. When the number of image samples is not high (e.g., 50), the nonlinear manifold based methods (e.g., MMD) could not get a high recognition rate. However, the performance of the affine subspace based methods (e.g., AHISD, SANP) is still good.

Table 2. Identification rates on Honda/UCSD Dataset

Methods	50 Frames	100 Frames	200 Frames
DCC[19]	76.92%	84.62%	94.87%
MMD[4]	69.23%	87.18%	94.87%
MDA[33]	82.05%	94.87%	97.44%
AHISD[3]	87.18%	84.62%	89.74%
CHISD[3]	82.05%	84.62%	92.31%
MSM[28]	74.36%	79.49%	76.92%
SANP[2]	84.62%	92.31%	94.87%
RNP[34]	87.18%	94.87%	**100%**
CRNP[35]	89.74%	97.44%	97.44%
JRNP	**92.31%**	**100%**	**100%**

CMU Mobo Dataset

CMU Mobo (Motion Boday) dataset [15] contains 96 sequences of 24 subjects walking on a treadmill. For each subject, there are 4 video sequences (with significant pose variation) collected in four walking patterns, respectively. As [2][34], the sample features are the uniform LBP histograms using circular (8,1) neighborhoods extracted from the 8×8 squares of the gray-scale images. While one image set per subject is randomly selected for training, the remaining image sets are used for testing.

Ten experiments are conducted, and the average identification rates and the standard deviations are summarized in Table 3. In all cases, JRNP has the highest identification rates. When the set length is 50, the improvements of JRNP over the second best, RNP, are 2%. Compared to CRNP, the improvements of JRNP are 0.8%, 0.4% and 0.5% when the set length is 50, 100 and 200, respectively. Compared to another state-of-the art method, SANP, over 2% improvement is achieved when the set length

is 50. Compared to other competitors, the advantage of JRNP is also significant. When there are 50 frames, the recognition rates of DCC, MSM and MDA are lower than 90%, which suggests that the extraction of discriminative information and manifold analysis depends on large number of samples per image set.

Table 3. Identification rates on CMU Mobo Dataset

Methods	50 Frames	100 Frames	200 Frames
DCC[19]	82.1%±2.7%	85.5%±2.8%	91.6%±2.5%
MMD[4]	90.1%±2.3%	94.6±1.9%	96.4%±0.7%
MDA[33]	86.2%±2.9%	93.2%±2.8%	95.8%±2.3%
AHISD[3]	91.6%±2.8%	94.1%±2.0%	91.9%±2.6%
CHISD[3]	91.2%±3.1%	93.8%±2.5%	**97.4%±1.9%**
MSM[28]	84.3%±2.6%	86.6%±2.2%	89.9%±2.4%
SANP[2]	91.8%±3.1%	94.7%±1.7%	97.3%±1.3%
RNP[34]	91.9%±2.5%	94.7%±1.2%	**97.4%±1.5%**
CRNP[35]	93.1%±2.5%	94.9%±3.0%	96.9%±2.5%
JRNP	**93.9%±2.2%**	**95.3%±2.5%**	**97.4%±1.9%**

YouTube Celebrities Dataset

YouTube Celebrities dataset [16] is a large-scale vide dataset, which contains 1910 video sequences of 47 celebrities (actors, actress and politicians). This dataset is more challenging than the previous two datasets since the images are mostly low resolution and have large pose/expression variation, motion blur, etc. As in RNP [34], we use the first 15 video sequences of 29 celebrities to do the experiments. For each subject, three video sequences are randomly selected as the training data, with other three randomly selected sequences as the testing data. We conduct 5 experiments by repeating the random selection of training/testing data.

The experimental results, including the average identification rate and the standard deviation, are summarized in Table 4. Again we could observe that JRNP get the best performance. Compared to SANP, 3.5%, 3.3% and 1.2% improvements are achieved by JRNP when set length is 50, 100 and 200, respectively. JRNP also outperforms RNP by 1.6%, 2.8% and 1.6% when the set length is 50, 100, and 200. Similarly JRNP is visibly better than CRNP. In this challenging task, MSM doesn't have a good result, with average identify rates lower than 70%. It is also interesting to see that AHISD's identification rate fluctuates with the increase of the frame number, and the similar trend of AHISD could also be found in the previous two datasets.

Table 4. Identification rates on You Tube Dataset

Methods	50 Frames	100 Frames	200 Frames
DCC[19]	68.7%±3.2%	73.8%±4.7%	76.1%±2.5%
MMD[4]	69.0%±3.5%	72.0%±4.6%	76.3%±4.3%
MDA[33]	63.9%±3.9%	74.2%±5.9%	74.5%±5.0%
AHISD[3]	73.3%±5.4%	72.6%±7.6%	66.9%±4.8%
CHISD[3]	72.4%±5.5%	73.6%±5.2%	75.2%±5.2%
MSM[28]	66.2%±4.6%	66.0%±8.6%	65.3%±6.5%
SANP[2]	73.3%±3.9%	74.9%±5.9%	78.3%±4.2%
RNP[34]	75.2%±5.4%	75.4%±5.1%	77.9%±5.5%
CRNP[35]	73.8%±2.2%	77.5%±4.5%	77.7%±3.3%
JRNP	**76.8%±2.5%**	**78.2%±3.9%**	**79.5%±2.9%**

C. Running time comparison

From Subsection B, we can see that JRNP achieves higher identification rates than all the competing methods, including the state-of-the-art approaches, such as SANP [2], RNP[34] and CRNP[35]. In this section, we are mainly comparing their running time, which is an important concern in practical applications.

We do face recognition on CMU Mobo dataset [15] with the same experimental setting as that in Subsection B. The programming environment is Matlab version 2011a. The desktop used is of i7 2.8 GHz CPU and with 4GB RAM. In Section III, we have shown JRNP has lower time complexity than the summarization of two fastest methods, RNP and JRNP. Here we mainly compare the running time of JRNP with that of other competing methods. In order to make the timing comparison fair, we also list the offline training of some methods. Besides the training phase of these discriminant methods (e.g., DCC, MDA) , the constructing of local linear subspace in MMD, the SVD of training sets in SANP, and the projection matrix of training sets in, JRNP, are also regarded as the offline training.

The offline training time and online testing time for classifying one image set with the frame number 100 are listed in Table 5. We can observe that JRNP has a very small offline training time compared to other competitors. From Table 5, we can see that the online running time (i.e., for classifying a testing image set) of JRNP are much less than all the other methods. Compared to SANP, the speedup of JRNP is over 14 times. In that case, the testing time of CRNP and RNP is 0.041 and 0.078, respectively, of which the summarization is larger than the testing time of JRNP. Although JRNP is a little slower than CRNP and RNP, the online running time of JRNP, 0.111 second per query image set, is still very fast.

In order to comprehensively evaluate the running time, Fig. 3 plots all the methods' testing time versus different frame number. It can be seen that the proposed JRNP is consistently faster than the competing methods. The running time of all the methods increase with the frame number except some special cases (e.g., DCC and MDA when the frame number is 200). Especially, AHISD's running time dramatically rise as the frame number increases.

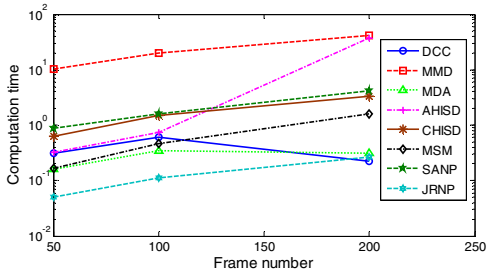

Fig.3 The testing time for one image set versus the frame number for all the competing methods in CMU Mobo dataset.

TABLE 5. Computation time (seconds) of different methods on CMU Mobo dataset with 100 frames for training and testing (classification of one image set). #1: offline training, #2: online testing.

	DCC	MMD	MDA	AHISD.	CHISD	MSM	SANP	**JRNP**
#1	16.4	19.8	5.87	N/A	N/A	N/A	7.71	2.41
#2	0.603	20.0	0.348	0.739	1.48	0.468	1.61	0.111

V. CONCLUSION

In this paper, we proposed jointly regularized nearest points (JRNP) for robust and fast face recognition based on image sets. The JRNP simultaneously minimize the distance between the query set and the whole gallery set and the between-set distances for each gallery class. Compared to RNP, the nearest point in the query set keeps the same for different gallery classes, which is beneficial to avoid over-fitting. Meanwhile, different from CRNP, the between-set distance is explicitly minimized to enhance discrimination. An efficient algorithm was proposed to solve the model of JRNP and a joint between-class distance was presented to combine the advantages of RNP and JRNP. We evaluated the proposed JRNP on several benchmark image set databases. The extensive experimental results clearly demonstrated that JRNP could achieve higher identification accuracy than the state-of-the-art but with a fast speed, making face recognition based on image sets more applicable in practical applications.

ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China under Grants no. 61402289 and 61272050, and Shenzhen Scientific Research and Development Funding Program under Grants JCYJ20140509172609171 and JCYJ20130329115750231.

REFERENCES

[1] Z. Cui, H. Chang, S.G. Shan, B.P. Ma, and X.L. Chen, Joint sparse representation for video-based face recognition, Neurocomputing 135: 306-312, 2014.

[2] Y. Q. Hu, A. S. Mian and R. Owens, "Face recognition using sparse approximated nearest points between image sets," IEEE PAMI 34(10), 1992-2004, 2012.

[3] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in Proc. CVPR 2010.

[4] R. Wang, S. Shan, X. Chen and W. Gao, "Manifold-manifold distance with application to face recogniton based on image set," in Proc. CVPR, 2008.

[5] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma, "Robust face recognition via sparse representation," IEEE PAMI 31(2): 21—227, 2009.

[6] R. Wang, H. Guo, L. S. Davis, Q. Dai, "Covariance Discriminative Learning: A Natural and Efficient Approach to Image Set Classification," in Proc. CVPR 2012.

[7] L. Zhang, M. Yang, X. Feng, Y. Ma and D. Zhang, "Collaborative Representation based Classification for Face Recognition," arXiv:1204.2358.

[8] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," IEEE PAMI 32(11): 2106-2112, 2010.

[9] H. Hotelling, "Relations between tow sets of variates," Biometrika, 28(3-4): 321-377, 1936.

[10] T. Hastie, R. Tibshirani, and J. Fridrman, The Elements of Statistical Learning, 2nd ed., Springer, 2009.

[11] A. Yang, A. Ganesh, Z. H. Zhou, S. Sastry, and Y. Ma, "Fast L1-Minimization Algorithms for Robust Face Recognition," (preprint)

[12] S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "A interior-point method for large-scale l1-regularized least squares," IEEE Journal on Selected Topics in Signal Processing, 1(4):606–617, 2007.

[13] Y. Nesterov, A. Nemirovskii, "Interior-point polynomial algorithms in convex programming," SIAM Philadelphia, PA, 1994.

[14] P. Viola and M. J. Jones, "Robust real-time face detection," International Journal of Computer Vision, 57(2): 137-154, 2004.

[15] R. Gross and J. Shi, The CMU Motion of Body (MoBo) Database. Technical Report CMU-RI-TR-01-18, Robust institute, 2001.

[16] M. Kim, S. Kumar, V. Pavlovic and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in Proc. CVPR, 2008.

[17] T. Wang and P. Shi, "Kernel grassmannian distances and discriminant analysis for face recogntion from image sets," Pattern Recognition Lettler, 30(13): 1161-1165, 2009.

[18] T. Ahonen, A. Hadid and M. Pietikainen, "Face description with local binary patterns: Application to face recogntion," IEEE PAMI, 28(12): 2037-2041, 2006.

[19] T.-K. Kim, O. Arandjelovic and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," IEEE PAMI, 29(6): 1005-1018, 2007.

[20] W. Liu, Z. Li, and X. Tang, "Spatio-temporal Embedding for Statistical Face Recognition from Video," in Proc. ECCV, 2006.

[21] X. Liu and T. Chen, "Video-Based Face Recognition Using Adaptive Hidden Markov Models," in Proc. CVPR, 2003.

[22] J. Stallkamp, H. K. Ekenel, R. Stiefelhagen, "Video-based Face Recognition on Real-World Data," in Proc. ICCV 2007.

[23] S. Zhou and R. Chellappa, "Probabilistic Human Recognition from Video," in Proc. ECCV, 2002.

[24] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla and T. Darrel, "Face recognition with image sets using manifold density divergence," in Proc. CVPR, 2005.

[25] G.Shakhnnarvovich, J. W. Fisher and T. Darrel, "Face recognition from long-term observation," in Proc. ECCV, 2002.

[26] K.-C. Lee, J. Yamaguchi, "The kernel orthogonal mutual subspace method and its application to 3D object recognition," in Proc. ACCV, 2007.

[27] M. Nishiyama, O. Yamaguchi and K. Fukui, "Face recognition with the multiple constrained constrained mutual subspace method," in Proc. AVBBPA, 2005.

[28] O. Yamaguchi, K. Fukui and K.-i. Maeda, "Face recognition using temporal image sequence," in Proc. FG, 1998.

[29] M. Nishiyama, M. Yuasa, T. Shibata, T. Wakasugi, T. Kawahara and O. Yamaguchi, "Recognizing faces of moving people by hierarchical image-set matching," in Proc. CVPR, 2007.

[30] T.-K. Kim, J. Kittler and R. Cipollar, "Incremental learning of locally orthogonal subspaces for set-based object recogntion," in Proc. BMVC, 2006.

[31] W. Fan and D.-Yeung, "Locally linear models on face appearance manifolds with application to dual-subspace based classification," in Proc. CVPR, 2006.

[32] A. W. Fitzgibbon and A. Zisserman, "Joint manifold distance: a new aproach to appearance based clustering," in Proc. CVPR, 2003.

[33] R. Wang and X. Chen, "Manifold discriminant analysis," in Proc. CVPR, 2009.

[34] M. Yang, P.F. Zhu, L. Van Gool and L. Zhang, "Face recognition based on regularized nearest points between image sets", In Proc. FG 2013.

[35] Y. Wu, M. Minoh and M. Mukunoki, "Collaboratively regularized nearest points", in Proc. BMVC, 2013.

[36] J. W. Lu, G. Wang and P. Moulin, "Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning", in Proc. ICCV, 2013.

[37] T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley, New York, 1991.

[38] Z.Cui, S.G. Shan, H.H. Zhang, S.H. Lao, and X.L. Chen, Image sets alignment for video-based face recognition, in Proc. CVPR 2012.

[39] K.-C. Lee, J. Ho, M.-H. Yang and D. Kriegman, "Video-base face recognition using probabilistic appearance manifolds," in Proc. CVPR, 2003.