Open access • Proceedings Article • DOI:10.1109/FG.2013.6553727

Face recognition based on regularized nearest points between image sets

Meng Yang, Pengfei Zhu, Luc Van Gool, Lei Zhang

Institutions: ETH Zurich

Published on: 22 Apr 2013 - IEEE International Conference on Automatic Face & Gesture Recognition

Topics: Affine hull

Face Recognition based on Regularized Nearest Points between Image Sets

Meng Yang[1], Pengfei Zhu[2], Luc Van Gool[1,3], Lei Zhang[2]
[1]Department of Information Technology and Electrical Engineering, ETH Zurich, Switzerland
[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[3]Department of Electrical Engineering/IBBT, K.U. Leuven, Belgium
{yangme@ee.ethz.ch}

***Abstract*—In this paper, a novel regularized nearest points (RNP) method is proposed for image sets based face recognition. By modeling an image set as a regularized affine hull (RAH), two regularized nearest points (RNP), one on each image set's RAH, are automatically determined by an efficient iterative solver. The between-set distance of RNP is then defined by considering both the distance between the RNPs and the structure of image sets. Compared with the recently developed sparse approximated nearest points (SANP) method, RNP has a more concise formulation, less variables and lower time complexity. Extensive experiments on benchmark databases (e.g., Honda/UCSD, CMU Mobo and YouTube databases) clearly show that our proposed RNP consistently outperforms state-of-the-art methods in both accuracy and efficiency.**

Keywords: regularized nearest points; regularized affine hull; image set; face recognition

I. INTRODUCTION

The recognition of objects of interest (e.g., human faces) is one of the most important problems in the communities of computer vision and pattern recognition. The traditional face recognition is usually formulated as a problem of identifying a human face from a single probe image, although the gallery set per subject could be a single image or multiple images. However, it is a big challenge to correctly identify a person from only a single face image in less-controlled/uncontrolled environments since the facial appearance changes dramatically due to various variations in pose, illumination, expression, disguise, etc. With the developments of video cameras and large-capacity-storage media, it becomes very convenient to collect gallery and probe image sets from video sequences or photo album for a known subject. The probe/gallery set for each subject incorporates more within-class appearance variation (e.g., the image sets shown in the bottom of Fig. 1), making the image sets based face recognition be able to achieve more satisfactory performance in practical applications.

Over the past decade, there has been growing interest in face recognition by sets of images. One special case of image-set-based recognition is video-based face recognition [1][20-23], where the images are collected from consecutive video sequences. In this paper, we focus on a more general image-set-based recognition problem, where there is no temporal information existed/provided in the image set (e.g., unordered photo album images). To solve this image-set based recognition problem researchers have proposed numerous approaches, which mainly fall into two categories [2][19]: parametric model based methods and non-parametric model-free methods. Parametric model based approaches [1][24-25] firstly represent each image set by some parametric distribution with the parameters estimated from the set data itself, and then calculate the between-set distance by measuring the similarity between two distributions (e.g., in terms of Kullback-Leibler divergence). However, the parametric methods need to solve the difficult parameter estimation problem and heavily require that the gallery and probe sets should have strong statistical correlations, which may not be true in practice.

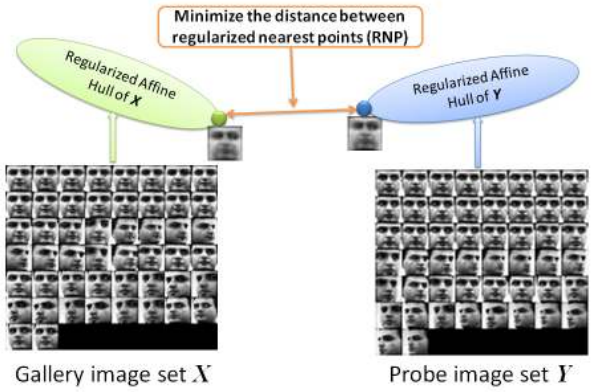

Fig. 1. Regularized Nearest Points (RNPs) of two image sets. Given two image sets (from Honda/UCSD dataset [1]), the RNPs are the two points, one on each set's regularized affine hull, between which the distance is the smallest.

In order to avoid the drawbacks of parametric methods, non-parametric model-free methods were proposed by representing an image set as a linear/affine subspace [3][19][26-28], mixture of subspaces [29-31], or nonlinear manifolds [4][17][32-33]. Usually the nonlinear-manifold methods express an image set as a combination of local linear subspaces [4][33]. Based on the representation of image sets, the between-set distance could be defined as the distance between two "exemplars" (e.g., the sample means) from these two image sets. 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 type of between-set distance for non-parametric methods is by comparing the structure of the non-

parametric model. For instance, canonical correlation analysis [9], which analyzes the principal angles and canonical correlations between linear subspaces, is widely used in [4][19][26][27][28][30][31]. Besides, Wang *et al.* [6] represented each image set with its natural second-order statistical covariance matrix, and formulated the image-set based classification as classifying points lying on a Riemannian manifold.

Very recently, Hu *et al.* [2] proposed an interesting image-set-based face recognition method, namely sparse approximated nearest points (SANP). By modeling each image set as an affine hull, Hu *et al.* selected two points (one point in each image set) with the closest distance as the sparse approximated nearest points (SANP), where SANPs are required to be sparsely represented by the original samples. The final between-set distance is the distance between the SANPs multiplied by the dimension of the affine hull. SANP achieves state-of-the-art performance compared to previous methods. However, SANP does not model the image set well although it utilizes both affine hull representation and sparse regularization in a brute-force way. The complex model (e.g., three representation terms and four unknown variables) makes SANP somewhat confusing, and the sparse constraint and many unknown variables also increase the difficulty and complexity to solve SANP.

This paper presents an efficient and effective regularized nearest points (RNP) method for image-set based face recognition. We will show that the complex formulation and the sparse constraint on the representation coefficients in SANP are not necessary. By modeling an image set as a regularized affine hull (RAH), two regularized nearest points (RNP), one on each RAH, are automatically computed, as shown in Fig. 1. Then the between-set distance is defined as the modulated distance between RNPs by the structure of image sets. Compared to SANP, RNP models the image set better and has a concise formulation with less number of parameters and unknown variables. An efficient algorithm is proposed to solve the proposed RNP with very low time complexity. Our experiments on benchmark image set databases clearly show that RNP leads to higher recognition accuracy than the previous methods, including SANP. And more importantly, the proposed RNP has a very fast speed, e.g., it is over 20 times faster than SANP on the CMU Mobo database [15].

The rest of this paper is organized as follows. Section II briefly reviews the SANP method in [2]. Section III presents the proposed RNP. Section IV conducts experiments and Section V concludes the paper.

II. Sparse Approximated Nearest Points (SANP)

Based on the work in [3] where each image set is modeled as an affine/convex hull, recently Hu *et al.* [2] proposed the sparse approximated nearest points (SANP) to combine the affine hull representation [3] and sparse representation [5]. SANP has two objectives. One is that the affine-hull regularized distance between two point sets should be small by minimizing

$$F_{\boldsymbol{\nu}_i,\boldsymbol{\nu}_j} = \left\| \left(\boldsymbol{\mu}_i + \boldsymbol{U}_i\boldsymbol{\nu}_i\right) - \left(\boldsymbol{\mu}_j + \boldsymbol{U}_j\boldsymbol{\nu}_j\right) \right\|_F^2 \tag{1}$$

where $\boldsymbol{\mu}_k$ is the sample mean of the k^{th} class data matrix $\boldsymbol{X}_k$, the columns of $\boldsymbol{U}_k$ are the orthonormal bases obtained from the singular value decomposition (SVD) of the centered data matrix of class k. It can be seen that this part is similar to the affine hull method in [3]. After minimizing Eq. (1), $\boldsymbol{\mu}_i+\boldsymbol{U}_i\boldsymbol{\nu}_i$ and $\boldsymbol{\mu}_j+\boldsymbol{U}_j\boldsymbol{\nu}_j$ are called the nearest points between the i^{th} and j^{th} classes, where $\boldsymbol{\nu}_i$ and $\boldsymbol{\nu}_j$ are the coding coefficients.

The other objective of SANP is that each of the two nearest points should be sparsely represented by the original data matrix, i.e.,

$$G_{\boldsymbol{\nu}_i,\boldsymbol{\alpha}} + \lambda_1\|\boldsymbol{\alpha}\|_1 = \left\|\left(\boldsymbol{\mu}_i + \boldsymbol{U}_i\boldsymbol{\nu}_i\right) - \boldsymbol{X}_i\boldsymbol{\alpha}\right\|_F^2 + \lambda_1\|\boldsymbol{\alpha}\|_1,$$
$$Q_{\boldsymbol{\nu}_j,\boldsymbol{\beta}} + \lambda_1\|\boldsymbol{\beta}\|_1 = \left\|\left(\boldsymbol{\mu}_j + \boldsymbol{U}_j\boldsymbol{\nu}_j\right) - \boldsymbol{X}_j\boldsymbol{\beta}\right\|_F^2 + \lambda_2\|\boldsymbol{\beta}\|_1,$$

where λ_1 and λ_2 are the parameters to tune the effect of sparse constraint.

The final model of SANP is

$$\left(\hat{\boldsymbol{\nu}}_i,\hat{\boldsymbol{\nu}}_j,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}\right) = \min_{\boldsymbol{\nu}_i,\boldsymbol{\nu}_j,\boldsymbol{\alpha},\boldsymbol{\beta}}\left(F_{\boldsymbol{\nu}_i,\boldsymbol{\nu}_j} + \gamma\left(G_{\boldsymbol{\nu}_i,\boldsymbol{\alpha}} + Q_{\boldsymbol{\nu}_j,\boldsymbol{\beta}}\right) + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_1\right) \tag{2}$$

and the final classification of a testing image set is conducted to find which class has the minimal between-set distance, which is defined as

$$D\left(c_i,c_j\right) = \left(d_i + d_j\right)\left[F_{\hat{\boldsymbol{\nu}}_i,\hat{\boldsymbol{\nu}}_j} + \gamma\left(G_{\hat{\boldsymbol{\nu}}_i,\hat{\boldsymbol{\alpha}}} + Q_{\hat{\boldsymbol{\nu}}_j,\hat{\boldsymbol{\beta}}}\right)\right] \tag{3}$$

where d_i and d_j are the dimension of the affine hulls (i.e., $\boldsymbol{U}_i$ and $\boldsymbol{U}_j$) of i^{th} class (i.e, c_i) and j^{th} class (i.e., c_j), respectively. For $\boldsymbol{X}_k$, there is another parameter, φ, as a threshold of preserving energy (e.g., φ=85%) in determining $\boldsymbol{U}_k$ and d_k.

Although SANP has achieved very interesting results on image sets based face recognition, there are several issues needing to be further considered.

a. The brute-force way to combine the affine hull representation and sparse representation makes the model of SANP rather complex (e.g., three representation terms in Eqs. (2) and (3), four parameters and four unknown variables), which increase the difficulty and complexity of solving SANP.
b. The l_1-norm sparse regularization on the representation coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ makes the solving of SANP time-consuming, although the fast solver of Accelerated Proximal Gradient (APG) method was adopted in SANP.

III. Regularized Nearest Points

In this section, we first present the model of the proposed regularized nearest points (RNP). Then we describe the solving algorithm and classification of RNP. Finally the time complexity of the proposed RNP is discussed.

A. Model of RNP

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, and $\boldsymbol{x}_{i,k}$ is the k^{th} image feature vector (e.g., raw pixel value) of the i^{th} class. Based on these sample data, in [2] and [3] the image set was approximated as an affine/convex and affine hull, respectively. In this paper, we propose a novel regularized affine hull (RAH) to model an image set:

$$RAH = \left\{ \boldsymbol{x} = \boldsymbol{X}_i \boldsymbol{\alpha} \middle| \sum\nolimits_k \alpha_k = 1, \left\|\boldsymbol{\alpha}\right\|_{l_p} \le \sigma \right\} \tag{4}$$

where $\alpha_k \in \Re$ for k=1, 2, … , n_i, and $\left\|\boldsymbol{\alpha}\right\|_{l_p}$ is the l_p-norm of the representation coefficients $\boldsymbol{\alpha}$. In order to give an intuitive illustration of RAH, we plot the solution space (i.e., the constraint) of RAH with n_i=3 and p=2 in Fig. 2. One could see that the solution space of RAH is not a hyperplane but a regularized partial region with the point of $\{\alpha_k = 1/n_i$ for k=1, 2,…, $n_i\}$ as its center. Compared to the conventional affine hull to model an image set, RAH can avoid containing the meaningless points which are too far from the sample mean.

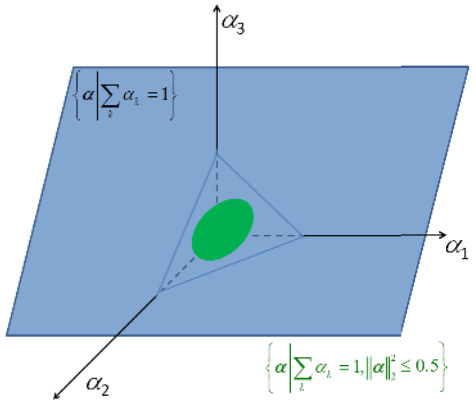

Fig. 2. An example of the solution space of RAH. The blue parallelogram represents the affine plane (i.e., the solution space of affine hull), while the green ellipse represents the solution space of the l_p-norm (p=2) RAH.

For a gallery image set and a probe image set, our basic idea is to find two nearest points, each point in the RAH of an image set, as the regularized nearest points (RNP). Let $\boldsymbol{X}_i$ be the i^{th} class data matrix in the gallery set and $\boldsymbol{Y} = \left[\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{n_y}\right]$ be the probe image set where $\boldsymbol{y}_k$ is the k^{th} image of $\boldsymbol{Y}$. Then we find the RNPs of $\boldsymbol{X}$ and $\boldsymbol{Y}$ by the following minimization:

$$\begin{aligned} &\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\|\boldsymbol{X}_i\boldsymbol{\alpha} - \boldsymbol{Y}\boldsymbol{\beta}\right\|_2^2 \\ &\text{s.t. } \sum_k \alpha_k = 1, \sum_k \beta_k = 1, \left\|\boldsymbol{\alpha}\right\|_{l_p} \le \sigma_1, \left\|\boldsymbol{\beta}\right\|_{l_p} \le \sigma_2 \end{aligned} \tag{4}$$

where the l_p-norm terms (e.g., $\left\|\boldsymbol{\alpha}\right\|_{l_p}$ and $\left\|\boldsymbol{\beta}\right\|_{l_p}$) could make the representation more stable by suppressing unnecessary samples' contribution to the representation, and the affine hull constraint (e.g., $\sum_k \alpha_k = 1, \sum_k \beta_k = 1$) could avoid the trivial solution (i.e., $\boldsymbol{\alpha}$=$\boldsymbol{\beta}$=$\boldsymbol{0}$). Using the Lagrangian formulation, the problem of RNP could be rewritten as

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left(\left\|\boldsymbol{X}_i\boldsymbol{\alpha} - \boldsymbol{Y}\boldsymbol{\beta}\right\|_2^2 + \lambda_1 \left\|\boldsymbol{\alpha}\right\|_{l_p} + \lambda_2 \left\|\boldsymbol{\beta}\right\|_{l_p} \right) \text{ s.t. } \sum_k \alpha_k = 1, \sum_k \beta_k = 1 \tag{5}$$

where λ_1 and λ_2 are the two Lagrangian multipliers.

According to the number of samples in the image set, the proposed RNP can be divided into two special cases: regularized nearest subspace classifier [7] and nearest neighbor.

In the first case, one of $\boldsymbol{X}_i$ and $\boldsymbol{Y}$ has only one sample. Taking the probe image set as an example (e.g., $\boldsymbol{y}$ for the probe image set), we could get β=1 and the proposed RNP degenerates to

$$\min_{\boldsymbol{\alpha}} \left(\left\|\boldsymbol{X}_i\boldsymbol{\alpha} - \boldsymbol{y}\right\|_2^2 + \lambda_1 \left\|\boldsymbol{\alpha}\right\|_{l_p} \right) \text{ s.t. } \sum_k \alpha_k = 1 \tag{6}$$

which is actually the regularized nearest subspace classifier [7][8][34] with an additional affine constraint.

In the second case, each image set will have only one sample (e.g., $\boldsymbol{x}_i$ for i^{th} gallery set or $\boldsymbol{y}$ for the probe set). We have α=β=1 and the proposed RNP degenerates to

$$\min \left(\left\|\boldsymbol{x}_i - \boldsymbol{y}\right\|_2^2 \right) \tag{7}$$

which is the model of nearest neighbor classifier.

B. Algorithm of RNP

The proposed RNP model has various instantiations by applying different norms to the representation coefficients. More specifically, when p=0 or 1, RNP 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 RNP with p=2 since high recognition accuracy and fast speed could be both achieved.

Since $\sum_k \alpha_k = 1, \sum_k \beta_k = 1$ are two linear equations, by relaxing them as $\sum_k \alpha_k \approx 1, \sum_k \beta_k \approx 1$ it is easy to integrate them with the first term of Eq. (5). Thus Eq. (5) with p=2 could be rewritten as

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left(\left\|z - \bar{\boldsymbol{X}}_i\boldsymbol{\alpha} - \bar{\boldsymbol{Y}}\boldsymbol{\beta}\right\|_2^2 + \lambda_1 \left\|\boldsymbol{\alpha}\right\|_2^2 + \lambda_2 \left\|\boldsymbol{\beta}\right\|_2^2 \right) \tag{8}$$

where z=[$\boldsymbol{0}$;1;1], $\bar{\boldsymbol{Y}} = \left[-\boldsymbol{Y}; \boldsymbol{0}^T; \boldsymbol{1}^T\right]$, $\bar{\boldsymbol{X}}_i = \left[\boldsymbol{X}_i; \boldsymbol{1}^T; \boldsymbol{0}^T\right]$, 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]. Eq. (8) has a closed-form solution, which, however, is not the fastest solver since a calculation of matrix inverse is needed for each pair of $\boldsymbol{X}_i$ and $\boldsymbol{Y}$.

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

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

where $\boldsymbol{P} = \left(\bar{\boldsymbol{Y}}^T\bar{\boldsymbol{Y}} + \lambda_2\boldsymbol{I}\right)^{-1}\bar{\boldsymbol{Y}}^T$. When $\boldsymbol{\beta}$ is fixed, we compute $\boldsymbol{\alpha}$ by

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

where $\boldsymbol{P}_i = \left(\bar{\boldsymbol{X}}_i^T\bar{\boldsymbol{X}}_i + \lambda_1\boldsymbol{I}\right)^{-1}\bar{\boldsymbol{X}}_i^T$.

The algorithm of RNP with p=2 is summarized in Algorithm 1. Here we initialize $\boldsymbol{\alpha}_0$=1/n_i, where n_i is the number of samples in the i^{th} class. It is easy to see that the cost function of Eq. (8) is lower bounded (≥0) and jointly convex to the variables $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Because in each step of Algorithm 1 the cost function value will decrease, the proposed Algorithm will converge to the global optimal solution.

Algorithm 1: Algorithm of Regularized Nearest Points (RNP) with p=2

Input: Projection matrices $\boldsymbol{P}_i$ and $\boldsymbol{P}$, data matrices $\bar{\boldsymbol{X}}_i$ and $\bar{\boldsymbol{Y}}$, a column vector z, and an initialization of $\boldsymbol{\alpha}_0$.

While not converged **do**

Compute the representation coefficients:

$\boldsymbol{\beta}_{t+1} = \boldsymbol{P}\left(z - \bar{\boldsymbol{X}}_i\boldsymbol{\alpha}_t\right)$;

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

End while

Output: representation coefficients $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$

C. Classifier of RNP

With the solved coefficients $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, the between-set distance of RNP is computed as follows

$$e_i = \left(\left\|\boldsymbol{X}_i\right\|_* + \left\|\boldsymbol{Y}\right\|_*\right)\cdot\left\|\boldsymbol{X}_i\hat{\boldsymbol{\alpha}} - \boldsymbol{Y}\hat{\boldsymbol{\beta}}\right\|_2^2 \tag{11}$$

where $\left\|\boldsymbol{X}\right\|_*$ (i.e., nuclear norm of $\boldsymbol{X}$) is the sum of the singular values: $\left\|\boldsymbol{X}\right\|_* = \sum_k \sigma_k(\boldsymbol{X})$, and $\left\|\boldsymbol{X}_i\hat{\boldsymbol{\alpha}} - \boldsymbol{Y}\hat{\boldsymbol{\beta}}\right\|_2^2$ represents the Euclidean distance between the two regularized nearest points.

The term $\|\boldsymbol{X}_i\|_*+\|\boldsymbol{Y}\|_*$ in Eq. (11) aims to remove the disturbance unrelated to the class information. For example, a wrong class which has much more samples than the correct class will have a lower value of $\left\|\boldsymbol{X}_i\hat{\boldsymbol{\alpha}} - \boldsymbol{Y}\hat{\boldsymbol{\beta}}\right\|_2^2$. Term $\|\boldsymbol{X}\|_*$ is the convex relaxation of the rank of matrix $\boldsymbol{X}$, which could reflect the representation ability of image set $\boldsymbol{X}$ (in our paper each column vector of $\boldsymbol{X}$ is normalized to have unit l_2-norm energy). The proposed e_i considers both the distance of RNPs and the structure of image sets, and it could well reflect the class information of $\boldsymbol{X}_i$ and $\boldsymbol{Y}$. The term of d_i+d_j in Eq. (3) of SANP also considers the structure of each image set, however, d_i+d_j is sensitive to the threshold φ (i.e., energy preserving percent).

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

$$\text{identity}\left(\boldsymbol{Y}\right) = \arg\min_i\left\{e_i\right\} \tag{12}$$

D. Complexity analysis

In this section, we compare the time complexity of the proposed RNP and the state-of-the-art sparse approximated nearest points (SANP) [2]. Some empirical analysis of sparse coding is firstly presented since SANP involves the step of sparse representation. Some fast l_1-norm minimization solvers have been recently reviewed in [11]. However, it is known that sparse coding with an $m\times n$ -sized dictionary has a computational complexity of $O(m^2n^{\varepsilon})$, where $\varepsilon\geq$1.2[12][13], m is the dimensionality of signal feature, and n is the number of dictionary atoms.

The sparse coding step of SANP has empirical complexity $O(m^2(n_i+n_y)^{\varepsilon})$ for computing the between-class distance of $\boldsymbol{X}_i$ and $\boldsymbol{Y}$, where n_i and n_y are the numbers of samples belonging to i^{th} gallery class and the probe image set, respectively. Besides, SANP needs additional calculations of SVD (e.g., $\boldsymbol{U}_i$ and $\boldsymbol{U}_y$) and variables (e.g., $\boldsymbol{v}_i$, and $\boldsymbol{v}_y$), where $\boldsymbol{U}_y$ and $\boldsymbol{v}_y$ are associated to $\boldsymbol{Y}$. Considering $\boldsymbol{U}_i$ for the gallery image set could be offline computed, the total time complexity of SANP for classifying the probe image set $\boldsymbol{Y}$ is about O_{svd}+ $\Sigma_i O(m^2(n_i+ n_y)^{\varepsilon})$. Here the summarization $\Sigma_i(.)$ means all the between-set distance of $\boldsymbol{Y}$ and $\boldsymbol{X}_i$, i=1,2,…, should be calculated, and O_{svd} denotes the time complexity of $\boldsymbol{Y}$'s SVD.

Let's analyze the complexity of the proposed RNP. For the query image set $\boldsymbol{Y}$, all the projection matrices of $\boldsymbol{P}_i$ and $\|\boldsymbol{X}_i\|_*$ for all gallery sets could be computed offline. The computing of $\boldsymbol{P}$ involves a matrix inverse, whose time complexity is roughly equal to the calculation of SVD of $\boldsymbol{Y}$ in SANP. Thus the step 1 to calculate $\boldsymbol{P}$ in RNP has a complexity of O_{svd}. The next step of RNP, i.e., the online iteration coding for $\boldsymbol{X}_i$ has a time complexity of $O(lm(n_i+n_y))$, where l is the iteration number. Usually a small value of l (e.g., l=5) could already get a good solution. In classification, $\|\boldsymbol{Y}\|_*$ could be fast computed due to it only involves the singular values. Therefore, the total complexity of RNP for a probe image set has a complexity of O_{svd}+ $\Sigma_i O(lm(n_i+ n_y))$ with l=5 in this paper.

The overall time complexity of RNP and SANP are listed in Table 1. Because $\varepsilon\geq$1.2 and the iteration number of RNP is much less than the feature dimension (e.g., l=5<<m=900 in YouTube), RNP has much lower time complexity than SANP.

TABLE 1. Time complexities of RNP and SANP for classify one probe image set.

Method	Step1	Step2
SANP	O_{svd} for SVD	$\Sigma_i O(m^2(n_i+n_y)^{\varepsilon})$ for sparse coding
RNP	O_{svd} for $\boldsymbol{P}$	$\Sigma_i O(lm(n_i+n_y))$ for iterative coding

IV. EXPERIMENTAL RESULTS

We perform experiments on benchmark image-set face databases to demonstrate the effectiveness of RNP. We first discuss the experimental setup in Section A. In Section B, we evaluate RNP on three benchmark datasets, followed by the running time comparison in Section C. In this paper, the parameters of RNP is fixed as λ_1=1e-3, and λ_2=1e-1 for all the experiments.

A. Experimenal setup

Three benchmark image set databases, including Honda/UCSD [1], CMU Mobo [15], and YouTube Celebrities [16] datasets, are used to evaluate the proposed RNP. All the face images in the three datasets were detected by using the Viola and Jone's face detector [14]. For Honda/UCSD and YouTube datasets, after histogram equalization the face images are resized to 20×20 and 30×30, respectively; and the raw pixel values of each image were directly used as the feature in 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 are conducted with the frame number 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 frame number.

The proposed RNP is compared with several state-of-the-art and representative image set classification methods, among 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], and Sparse Approximated Nearest Point (SANP) [2] are affine subspace based methods. All the competing methods are implemented by using the source codes provided by the authors, with the parameters tuned for their best results according to the recommendations in the original papers. For AHISD, CHISA and SANP, we used their linear versions since we didn't consider the kernel version of RNP in this paper. In Honda/UCSD and CMU Mobo datasets, there is a single training image set for each class. Thus following the seting of [19], each single training image set for DCC was randomly divided into two subsets to construct the within-class sets.

B. Experimental results and analysis

Honda/UCSD Dataset

The Honda/UCSD dataset contains 59 video sequences of 20 different subjects [1]. For each subject, different poses and expressions appear across different sequences, as shown in the face images in Figure 1. As the experimental setting of [1][2], 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. We can clearly see that the proposed RNP achieves the best performance in all cases, especially when the frame number is 200 all the testing sets are correctly recognized. The linear RNP outperforms SANP and even has the same performance to the kernel version of SANP [2]. When there are enough image samples in each image set, 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. Recognition rates on the Honda/UCSD Dataset

Methods/Set Length	50 Frames	100 Frames	200 Frames
DCC	76.92%	84.62%	94.87%
MMD	69.23%	87.18%	94.87%
MDA	82.05%	**94.87%**	97.44%
AHISD	**87.18%**	84.62%	89.74%
CHISD	82.05%	84.62%	92.31%
MSM	74.36%	79.49%	76.92%
SANP	84.62%	92.31%	94.87%
RNP	**87.18%**	**94.87%**	**100%**

CMU Mobo Dataset

The 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], the employed sample features are the uniform LBP histograms using circular (8, 1) neighborhoods extracted from the 8×8 squares of the gray-scale images. One image set per subject is randomly selected as the training data, with the remaining image sets as the testing data.

TABLE 3. Recognition rates on the CMU Mobo Dataset

Methods/Set Length	50 Frames	100 Frames	200 Frames
DCC	82.1%±2.7%	85.5%±2.8%	91.6%±2.5%
MMD	90.1%±2.3%	94.6±1.9%	96.4%±0.7%
MDA	86.2%±2.9%	93.2%±2.8%	95.8%±2.3%
AHISD	91.6%±2.8%	94.1%±2.0%	91.9%±2.6%
CHISD	91.2%±3.1%	93.8%±2.5%	**97.4%±1.9%**
MSM	84.3%±2.6%	86.6%±2.2%	89.9%±2.4%
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%**

Ten experiments are conducted, with the average recognition rates and the standard deviations are summarized in Table 3. In all cases, RNP has the highest identification rates. Although SANP and CHISD have close recognition accuracy to RNP, we will see that the running time of RNP is much less than that of SANP and CHISD in the following Section of running time comparison. When there are 50 frames, DCC, MSM and MDA have recognition rates lower than 90%, which

may result from the fact that extraction of discriminative information and manifold analysis depend on enough samples per image set. Compared to AHISD, the advantage of RNP is significant, which validates that the regularization of RAH indeed brings benefits to the final classification.

YouTube Celebrities Dataset

The YouTube Celebrities dataset [16] is a large-scale video dataset. 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 shown in Fig. 3. In this part, the video sequences of the first 29 celebrities are used to do the experiments. For each subject, three video sequences are randomly selected as the training data, with the 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 recognition rate and the standard deviation, are summarized in Table 4. Similar conclusions to those on the previous two datasets could be made. RNP has better performance than all the competing methods. Compared to the second best method, SANP, over 1% improvement is achieved when the frame number is 50 and 100. In this challenging test, MSM has the worst result, with average identification rates less than 70%. It is also interesting to see that AHISD's recognition rate fluctuates with the increase of the frame number, similar to what have found in the previous two datasets.

Fig.3.Some examples of the YouTube dataset.

TABLE 4. Recognition rates on the YouTube Dataset

Methods/Set Length	50 Fames	100 Frames	200 Frames
DCC	68.7%±3.2%	73.8%±4.7%	76.1%±2.5%
MMD	69.0%±3.5%	72.0%±4.6%	76.3%±4.3%
MDA	63.9%±3.9%	74.2%±5.9%	74.5%±5.0%
AHISD	73.3%±5.4%	72.6%±7.6%	66.9%±4.8%
CHISD	72.4%±5.5%	73.6%±5.2%	75.2%±5.2%
MSM	66.2%±4.6%	66.0%±8.6%	65.3%±6.5%
SANP	73.3%±3.9%	74.9%±5.9%	78.3%±4.2%
RNP	**74.9%±5.4%**	**76.1%±5.5%**	**78.9%±6.4%**

C. *Runing time comparison*

From Section B, we can see that RNP achieves higher recognition rates than all the competing methods, including the recently developed SANP [2]. Next let's compare their running time, which is one the most important concerns in practical applications.

We do face recognition on CMU Mobo dataset [15] with the same experimental setting as that in Section B. The programming environment is Matlab version 2001a. The desktop used is of i7 2.8 GHz CPU and with 4GB RAM. In order to make the running time comparison fairer, we also list the offline training time of some methods. Apart from these discriminant methods (e.g., DCC, MDA) which need a training phase, the construction of local linear subspace in MMD, the SVD of training sets in SANP, and the projection matrix learning of the training sets in RNP are also regarded as the offline training.

The offline training time and online testing time for classifying one image set with frame number as 100 is listed in Table 5. RNP has very short offline training time since only several matrix inverse computations are needed. The online testing time is more important for a classifier. From Table 5, we can see that the running time (i.e., for classifying a testing image set) of RNP is much less than all the other methods. Compared to SANP, the speedup of RNP is over 20 times. RNP is about 5 times faster than the second fastest method, MDA, with having much higher recognition accuracy.

In order to more comprehensively evaluate the running time, in Fig. 4 we plot all the methods' testing time versus different frame numbers. It can be seen that the proposed RNP is consistently faster than all the competing methods. The running time of all the methods will increase as the frame number except some special cases (e.g., DCC and MDA when the frame number is 200). Especially, AHISD's running time will dramatically rise as the frame number increases.

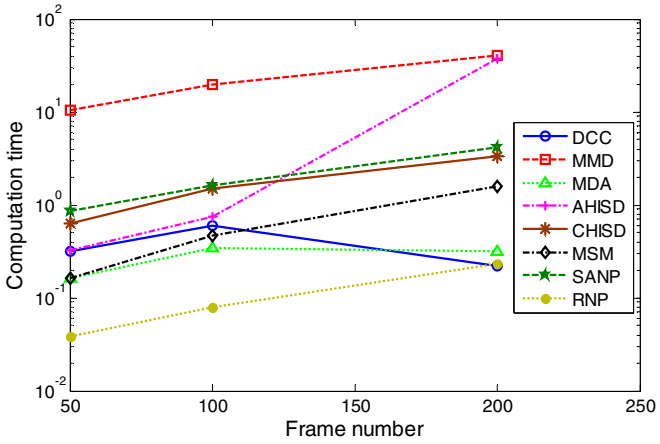

Fig. 4. The testing time for one image set versus the frame number for all the competing methods on the CMU Mobo dataset.

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

	DCC	MMD	MDA	AHISD.	CHISD	MSM	SANP	**RNP**
#1	16.4	19.8	5.87	N/A	N/A	N/A	7.71	**0.21**
#2	0.603	20.0	0.348	0.739	1.48	0.468	1.61	**0.078**

V. CONCLUSION

In this paper, we proposed a regularized nearest points (RNP) method for robust and fast face recognition based on image sets. We developed a novel regularized affine hull (RAH) to represent an image set, and defined the between-set distance as the distance between RNPs with consideration of the structure of image set. An efficient algorithm was also developed to implement RNP for image set based face recognition. We evaluated the proposed RNP on several benchmark image set databases. The extensive experimental results clearly demonstrated that RNP could achieve higher identification accuracy than the state-of-the-art methods (e.g., sparse approximated nearest points) but with much faster speed, making image sets based face recognition more applicable in practical applications. In this paper, we only discussed RNP with l_2-norm regularization. Nevertheless, RNP is a general classification scheme, and different regularizations (e.g., sparse, non-negative) and the kernel tricks (e.g., Gaussian kernel) could be employed for different applications.

ACKNOWEDGEMENT

This work was supported by the Hong Kong Polytechnic University internal grant G-YK25.

REFERENCES

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

[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] L. Zhang, M. Yang, and X. Feng, "Sparse Representation or Collaborative Representation: Which Helps Face Recognition?" in Proc. ICCV, 2011.

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