Probabilistic Nearest Neighbor Search for Robust Classification of Face Image Sets

Wen Wang[1,2], Ruiping Wang[1], Shiguang Shan[1], Xilin Chen[1]
[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
wen.wang@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

***Abstract*— Classification with image sets is recently a compelling technique for video-based face recognition. Previous methods in this line mostly assume each image set is pure, i.e., containing well-aligned face images of the same subject, which however is hardly satisfied in real-world applications due to incorrect face detection, questionable tracking, or multiple faces in a single image. This paper proposes a Probabilistic Nearest Neighbor (ProNN) search method to enhance the robustness of NN search against impure image sets by leveraging the statistical distribution of the involved image sets. Specifically, we represent image sets by affine hull, a well-recognized set model, to account for the unseen appearances in each image set. We further exploit a constraint that these unseen appearances statistically follow some pre-specified distribution (Gaussian in this work). Finally, in search of a pair of nearest neighbor points (one per hull), at the same time their distance being minimized, the probability of each point belonging to the same class as that of its corresponding hull is maximized. The proposed ProNN method is evaluated on three widely-studied public databases, Honda/UCSD, YouTube Celebrities and Multiple Biometric Grand Challenge (MBGC), under two kinds of experimental settings where image sets are contaminated either with false positive faces or images of other subjects. Extensive experiments demonstrate the superiority of the proposed approach over state-of-the-art methods.**

I. Introduction

With the rapid progress of video technologies, image sets are commonly available and can be easily collected by video surveillance, multi-view cameras, photo albums or long term observations. Compared with a single face image, richer information is embedded in an image set as it can cover a lot of variations of the person's facial appearance. Therefore, face recognition with image sets has attracted increasing interest recently and demonstrated promising performance in realistic environment [1]–[18]. For classification with face image sets, both the gallery and probe samples are image sets, each of which is assumed to contain facial images or video frames belonging to one single person.

According to how to model the image sets, relevant approaches mainly fall into four categories: statistical model based methods [1]–[5] , linear subspace based methods [6]–[9], nonlinear manifold based methods [10]–[14] and affine subspace based methods [15]–[18].

Several works tend to model statistical nature of image sets. In earlier years, researchers attempt to represent the image set with some well studied probability density functions, such as single Gaussian in [1] and Gaussian Mixture model(GMM) in Manifold Density Method (MDM) [2]. The similarity between two distributions is then measured by the classical Kullback-Leibler Divergence (KLD). More recently, Wang *et al.* [3] propose a Covariance Discriminative Learning(CDL) method to model the image set by its natural second-order statistic, i.e. covariance matrix, and further conduct discriminative learning on a Riemannian manifold. While only covariance information is modeled in CDL, Lu *et al.* [4] propose to combine multiple order statistics as features of image sets, and develop a localized multi-kernel metric learning (LMKML) algorithm for classification.

Rather than modeling the statistical nature of image sets, linear subspace based methods make the assumption that each image set spans a linear subspace. Specially, the Mutual Subspace Method (MSM) [6] and Discriminant-analysis of Canonical Correlations (DCC) [7] represent each image set

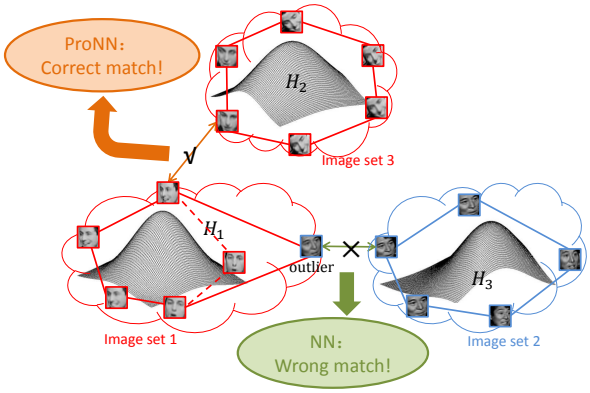


Fig. 1: Conceptual illustration of the proposed approach. H_1, H_2 and H_3 are affine hulls (the polygons with solid lines) formed by each of the three image sets (the cloud shapes) respectively. In the figure, face images superimposed with different colors denote different subjects. Since image set 1 contains an outlier sample, when matching it with other image sets 2 and 3, nearest neighbor points selected based on affine hull model (i.e. "NN") would cause a wrong match to image set 2. In our approach "ProbNN", with an additional consideration of the statistical structure of the image sets, the outlier in set 1 can be easily filtered out and a more accurate affine hull model with shrinking region (bounded by dashed red lines) can be formed which yields the correct match of image set 1 to set 3.

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

as a single linear subspace and compute the principal angles of two linear subspaces for classification. Grassmann Discriminant Analysis (GDA) [8] and Grassmann Embedding Discriminant Analysis (GEDA) [9] also model the image sets as linear subspaces but perform classification on the Grassmann manifold spanned by the subspaces according to Riemannian geometry.

In the literature, non-linear manifold has also been employed to represent an image set. In Manifold-Manifold Distance (MMD) [10] and Manifold Discriminant Analysis (MDA) [11], the similarity between manifolds is converted to integrating the distances between pair-wise subspaces. Further, Cui *et al.* [12] adopt the similar set modeling strategy, but attempt to align the image sets with a generic reference sets for more precise local model matching. Chen *et al.* [13] propose to measure the distance between two image sets with the distance between two nearest local linear subspaces searched by joint sparse approximation. In a more recent work [14], deep reconstruction models are constructed to automatically learn the underlying manifold structure.

Besides the above three trends, the affine subspace is proposed to characterize the large variations in image sets, which tends to represent the unseen appearances in each image set via linear combinations of images in the set. For instance, Affine Hull based Image Set Distance (AHISD) [15] and Convex Hull based Image Set Distance (CHISD) [15] are proposed to model each image set by an affine/convex hull model respectively, and the dissimilarity between two hulls are defined as the distance between a pair of nearest points belonging to either hull respectively. However, the affine hull model would fail when image sets of different classes have intersections. To address the issue, Sparse Approximated Nearest Points (SANP) [16] is proposed to generate a pair of virtual nearest points respectively approximated by the images in either set under sparse representation constraint. More recently, Regularized Nearest Points (RNP) [18] is proposed to approximate each image set by a regularized affine hull model, which exploits a constraint to regularize the structure of image sets. Further in [17] an Adaptive Multi Convex Hull metric is presented to use multiple local convex hulls to approximate an image set.

In real world applications, the image sets usually contain images/frames of low-quality with complicated variations in facial appearance due to changes in pose, expression and illumination. Therefore, face image sets in real-world applications might contain outliers such as non-faces (false face detection), badly-aligned faces, and even other persons' face images due to failure of tracking or the existence of multiple faces in the original image/frame. None of the methods mentioned above has ever explicitly considered the outlier problem. Therefore, most of them suffer from accuracy degradation to some extent when image sets are contaminated by outliers.

To deal with the outliers in the image sets and achieve robust classification of image sets, a Probabilistic Nearest Neighbor (ProNN) Search method is proposed with consideration of both the nearest neighbor distance and the statistical structure of the image sets. Specifically, we represent image sets by affine hull models to account for the unseen appearances via the affine combination of the images in each image set [15], [16]. Simultaneously, we characterize the statistical structure of the facial appearances in each set by making a natural assumption that they follow some distribution. Finally, given two image sets, we discover a pair of ProNNs, one per hull, satisfying two criteria: 1) they are nearest points that can be virtual; 2) each of them has large probability belonging to the class of most of the images in its hull. Fig. 1 gives an intuitive explanation about the proposed ProNN.

In sum, the contributions of our proposed method are as follows. On the one hand, our method inherits the advantages of affine hull model that can account for unseen appearances in the form of affine combinations of sample images. On the other hand, we propose a probability based approach to enhance the robustness of affine hull model to outliers that contaminate face image sets. It manifests the effect of modeling statistical structure of each image set for robust face image set classification. Finally, extensive experiments have been conducted to demonstrate the superiority of our proposed approach over state-of-the-art methods.

The rest of this paper is organized as follows. In Sec. II we introduce our Probabilistic Nearest Neighbor (ProNN) Search method in detail. Then Sec. III discusses about the main distinctions between some most related previous works and our proposed method. In Sec. IV, we compare the performance of our method with state-of-the-art methods in two groups of simulation experiments on three public face databases respectively. Finally, conclusions with possible future directions are summarized in Sec. V.

II. Probabilistic Nearest Neighbor Search

In this section, we first describe the conventional nearest neighbor search methods based on the affine hull model. Then we elaborate how to conduct Probabilistic Nearest Neighbor (ProNN) Search. Finally, we present optimization method of ProNN.

A. Nearest neighbor search based on the affine hull model

Given a total of N image sets, we denote $X_i = \{x_1^i, x_2^i, ..., x_{N_i}^i\}$ as the i-th set containing N_i samples. X_i belongs to one of the classes denoted by $Class_i$ and x_k^i is the d-dimensional feature vector of the k-th image in X_i. An image set can be characterized as an affine hull spanned by samples [15]:

$$H_i = \left\{ x = \sum_{k=1}^{N_i} \alpha_k^i x_k^i \middle| \sum_{k=1}^{N_i} \alpha_k^i = 1 \right\}, i = 1, ..., N. \quad (1)$$

Then by using the sample mean $\mu_i = \frac{1}{N_i} \sum_{k=1}^{N_i} x_k^i$, we can parameterize the affine hull as follows:

$$H_i = \left\{ x = \mu_i + U_i v_i | \, v_i \in \mathbb{R}^l \right\}. \quad (2)$$

where $U_i = [u_{i1}, ..., u_{il}]$ is an orthonormal basis which is obtained from the Singular Value Decomposition (SVD) of

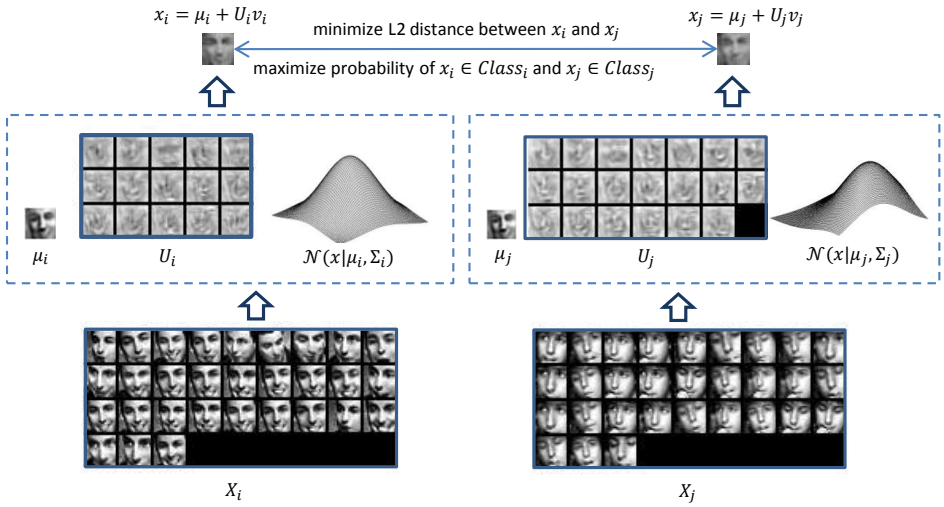


Fig. 2: An illustration of the Probabilistic Nearest Neighbors (ProNNs) taking two image sets as examples. Given two face image sets X_i and X_j, we represent the points in each of them with the affine hull model (μ_i, U_i) and (μ_j, U_j) which are linear combinations of images in each set. The points in the two hulls are assumed to follow Gaussian distribution $\mathcal{N}(x|\mu_i, \Sigma_i)$ and $\mathcal{N}(x|\mu_j, \Sigma_j)$ respectively. Thus the ProNNs are searched by minimizing distance between neighbor points and meanwhile maximizing probability of each point belonging to the same class of its hull.

$\left[x_1^i - \mu_i, ..., x_{N_i}^i - \mu_i\right]$. Note that the directions corresponding to near-zero singular values are discarded, leading to l $(l < N_i)$ singular vectors in U_i.

According to [15], nearest neighbor search based on the affine hull model tends to compute the affine-hull distance between two sets by

$$dist(H_i, H_j) = \min_{y,z} \|y - z\|_2^2, \quad y \in H_i, z \in H_j \tag{3}$$

Following (3), we have $y = \mu_i + U_i v_i$, $z = \mu_j + U_j v_j$, then the optimization problem can be rewritten as follows,

$$dist(H_i, H_j) = \min_{v_i, v_j} \|(\mu_i + U_i v_i) - (\mu_j + U_j v_j)\|_2^2. \tag{4}$$

By defining $U = (U_i, -U_j)$ and $v = \begin{pmatrix} v_i \\ v_j \end{pmatrix}$, the optimization becomes a standard least squares problem

$$\min_v \|Uv + \mu_i - \mu_j\|_2^2. \tag{5}$$

and we can compute its analytical solution as follows:

$$v = (U^T U)^{-1} U^T (\mu_j - \mu_i) \tag{6}$$

It follows that the distance between the two hulls can be rewritten as $\|(I - U(U^T U)^{-1} U^T)(\mu_j - \mu_i)\|$. Finally a simple NN classifer can be used to conduct classification.

Here we take Fig. 1 as an example. As shown in the figure, we aim at classifying image set 1 containing an outlier sample either to set 2 or set 3. According to affine hull based NN search, set 1 is wrongly matched to set 2 due to the effect of the outlier and that the affine hull model doesn't take into consideration of statistical structure of the set.

B. Probabilistic Nearest Neighbor Search

To enhance the robustness to outliers, we additionally take statistical structure of each image set into consideration while modeling the image set using affine hull.

For characterizing the statistical structure of each image set, we make an assumption that points in each affine hull are generated i.i.d from some distribution, which is assumed to be Gaussian distribution here, and estimate the distribution by samples in each set.

First we estimate the mean and covariance matrix by image features for each set.

$$\begin{aligned} x_k^i &\sim \mathcal{N}(x|\mu_i, \Sigma_i), \\ \mu_i &= \frac{1}{N_i} \sum_{k=1}^{N_i} x_k^i, \\ \Sigma_i &= \frac{1}{N_i} \sum_{k=1}^{N_i} (x_k^i - \mu_i)(x_k^i - \mu_i)^T. \end{aligned} \tag{7}$$

where $\mathcal{N}(x|\mu_i, \Sigma_i)$ denotes a Gaussian distribution with mean μ_i and covariance matrix Σ_i, and its Probability Density Function (PDF) is:

$$\mathcal{N}(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)} \tag{8}$$

Assuming that for two points $y \in H_i$ and $z \in H_j$, we have $y \sim \mathcal{N}(x|\mu_i, \Sigma_i)$, and similarly $z \sim \mathcal{N}(x|\mu_j, \Sigma_j)$. Then we can express the probability of $y \in Class_i$ and $z \in Class_j$ as:

$$\begin{aligned} P(y \in Class_i) &\propto \mathcal{N}(y|\mu_i, \Sigma_i) \\ &\propto \exp\left[-\frac{1}{2}(y - \mu_i)^T \Sigma_i^{-1} (y - \mu_i)\right]; \\ P(z \in Class_j) &\propto \mathcal{N}(z|\mu_j, \Sigma_j) \\ &\propto \exp\left[-\frac{1}{2}(z - \mu_j)^T \Sigma_j^{-1} (z - \mu_j)\right]. \end{aligned} \tag{9}$$

We tend to maximize the probability of each point belonging to the same class as the images in corresponding set as follows:

$$\begin{aligned} &\max_{y} P(y \in Class_i); \\ &\max_{z} P(z \in Class_j). \end{aligned} \tag{10}$$

The above two constraints, i.e., minimizing the distance in (3) and maximizing the probability in (10), are combined together to form a simple and direct objective function, which can be minimized to discover a pair of nearest unseen points, one per hull, with large probability belonging to the same class as the images in corresponding set. Fig. 2 demonstrates the basic scheme of our method.

Therefore, we search for a pair of ProNNs $\{y \in H_i, z \in H_j\}$ by the optimization problem below:

$$\begin{aligned} \{y^*, z^*\} &= \arg\min_{y \in H_i, z \in H_j} J(y, z) \\ &= \arg\min_{y \in H_i, z \in H_j} \frac{dist(y, z)}{P(y \in Class_i) \cdot P(z \in Class_j)} \end{aligned} \tag{11}$$

Since $y \in H_i$ and $z \in H_j$, we have $y = \mu_i + U_i v_i$, $z = \mu_j + U_j v_j$. Then by feeding this into (11) and considering (9), the objective function can be rewritten as follows:

$$\begin{aligned} J(v) &= \|(\mu_i + U_i v_i) - (\mu_j + U_j v_j)\|_2^2 \\ &\quad \cdot \exp\left\{\frac{1}{2}[v_i^T U_i^T \Sigma_i^{-1} U_i v_i + v_j^T U_j^T \Sigma_j^{-1} U_j v_j]\right\} \\ &= \|Uv + \mu_i - \mu_j\|_2^2 \cdot \exp\left(\frac{1}{2} v^T A v\right) \end{aligned} \tag{12}$$

where $v = \begin{pmatrix} v_i \\ v_j \end{pmatrix}$, $U = (U_i, -U_j)$, $A = \begin{pmatrix} U_i^T \Sigma_i^{-1} U_i & \\ & U_j^T \Sigma_j^{-1} U_j \end{pmatrix}$.

C. Solving and Optimization

To solve the optimization problem $v^* = \arg\min_v J(v)$, we first derive the gradient of $J(v)$ with respect to variable v:

$$\begin{aligned} \frac{\partial J(v)}{\partial v} =& \frac{\|Uv + \mu_i - \mu_j\|_2^2 \exp\left(\frac{1}{2} v^T A v\right)}{2} (A + A^T) v \\ &+ 2 \exp\left(\frac{1}{2} v^T A v\right) U^T (Uv + \mu_i - \mu_j) \end{aligned} \tag{13}$$

Then we can exploit the output computed by (6) as an initial value and solve the optimization problem by Conjugate Gradient (CG) method. Thus we find a pair of ProNNs, and the dissimilarity between sets can be defined as Euclidean distance between this pair of ProNNs. Finally, classification can be conducted by a simple Nearest Neighbor (NN) classifier.

As shown in Fig. 1, with an additional consideration of the statistical structure of image sets, the ProNN method filters out the outlier in set 1 and forms a more accurate affine hull model with shrinking region (bounded by dashed red lines) which yields the correct match of image set 1 to set 3.

III. Discussion

Here we give a discussion about similarities and differences between our approach and some related works.

A. Relation to other nearest neighbor search methods

Compared with other nearest neighbor search methods, such as Affine Hull based Image Set Distance (AHISD) [15], Convex Hull based Image Set Distance (CHISD) [15], Sparse Approximated Nearest Point (SANP) [16] and Regularized Nearest Points (RNP) [18], our approach also basically boils down to computing distance between some samples which are defined nearest. But we consider both nearest distance and the statistical structure of image data in each set, which makes our method more robust against outliers.

B. Relation to other statistical model based methods

Compared with other statistical model based methods, such as Gaussian Density Matching (GDM) [1], Manifold Density Method (MDM) [2] and Covariance Discriminative Learning(CDL) [3], our approach also tends to model the statistical structure of each image set. But we exploit the proved efficient affine hull model to explicitly represent the unseen appearances which do not appear in image sets. Therefore, our probabilistic model can characterize the underlying appearances in each set and simultaneously represent the probability distribution of these underlying appearances.

IV. Experiments

In this section, we conduct experiments under two different kinds of situations to evaluate the performance of our approach when image sets are contaminated by containing face detection false-alarm images or by including images from other classes respectively.

A. Datasets and Settings

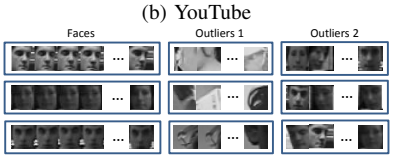
We evaluate the experimental performance of different methods on three widely studied public datasets: Honda/UCSD [19], YouTube Celebrities [20] and Multiple Biometric Grand Challenge (MBGC) [21], [22].

The Honda/UCSD database contains 59 video sequences of 20 different persons. Each video covers large pose changing and expression variation, and the length of the videos varies from 12 to 645 frames. The detected faces are resized to 20×20 gray-scale images and pre-processed by histogram equalization in order to eliminate lighting effects, following the similar settings in previous works [3], [4], [11], [16]. Note that for more accurate evaluation, we conduct 5-fold cross validation experiments and take one image set from each subject as the gallery, and the rest sets as probes.

The YouTube Celebrities database is a quite challenging and widely used video face dataset collected in real world condition. It consists of 1910 video clips of 47 celebrities. The clips have different numbers of frames and are mostly low resolution and highly compressed. The detected faces are resized and pre-processed similarly with those on Honda. Following the setting in [3], [11], we conduct 10-fold cross

(a) Honda

(b) YouTube



(c) MBGC

Fig. 3: Examples of contaminated sets. Note that "Outliers 1" denotes detection errors, and "Outliers 2" denotes faces from other classes.

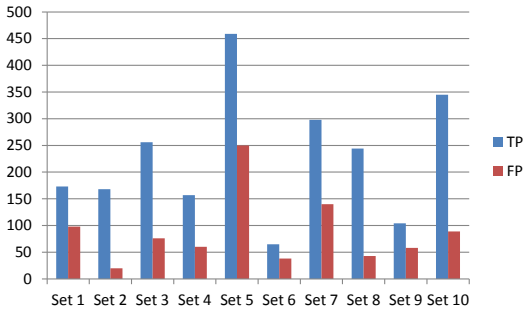


Fig. 4: The number of True Positive (TP) and False Positive (FP) face images in 10 randomly picked sets from YouTube contaminated by detection errors.

validation experiments, and in each fold, each subject has 3 randomly chosen image sets for gallery, and 6 for probes.

The MBGC database consists of 143 subjects walking towards a camera in a variety of illumination conditions, and the number of videos per subject ranges from 1 to 5. We take subsets of the database containing at least k sequences per subject as S_k, $k = 2, 3, 4$. Following the similar settings in [24], [25], we resize the gray-scale face images to 100×100 and conduct leave-one-out testing.

B. Comparison Results and Analysis

We compare our approach with several state-of-the-art image set classification methods developed in recent years, including Affine Hull based Image Set Distance (AHISD) [15], Convex Hull based Image Set Distance (CHISD) [15], Sparse Approximated Nearest Point (SANP) [16], Discriminant Canonical Correlations (DCC) [7], Manifold Discriminant Analysis (MDA) [11] and Covariance Discriminative Learning(CDL) [3].

For fair comparison, we obtained the source code of all methods from the original authors, and the important parameters followed the same recommendations in the original references. Specifically, in AHISD and CHISD, we used their linear version and retained 95% data energy by PCA. The error penalty parameter in CHISD was set to $C = 100$ as in [15]. For SANP, the parameters were the same as [16]. For DCC, we divided the single gallery image set of each object in Honda/UCSD into two subsets randomly which is the same as [7] for computing within-class scatter matrix. In MDA, the parameters were the same as [10]. For CDL, we used KDA for discriminative learning following the same setting as [3].

In order to verify the robustness, we evaluate experimental performance under two kinds of situations where different kinds of noise are added artificially. One situation is on image sets contaminated by face detection errors which is quite commonly encountered in practical conditions, the other one is on image sets contaminated by images from other classes to simulate group images or multiple faces in a video frame. The examples of contaminated sets in Honda/UCSD, YouTube Celebrities and MBGC are shown in Fig. 3.

1) Comparison results on sets contaminated by detection errors: On the one hand, we use the fast multi-pose face detection system in [23] to detect faces in frames of the three databases, and thus the image sets contain both real faces and false alarms as well. We randomly pick 10 sets from YouTube and compare the number of True Positive (TP) and False Positive (FP) face images of each set in Fig. 4. From Fig. 4, the average portion of FP is as high as about 28% of the total number of face detector output (TP + FP).

Besides, we remove these detection errors in Honda/UCSD and YouTube manually to obtain clean image sets. Results of the experiments conducted on these clean sets are shown in Tab. I. It is obvious that ProNN achieves very competitive performance with the state-of-the-art methods and performs much better than the baseline affine hull method AHISD.

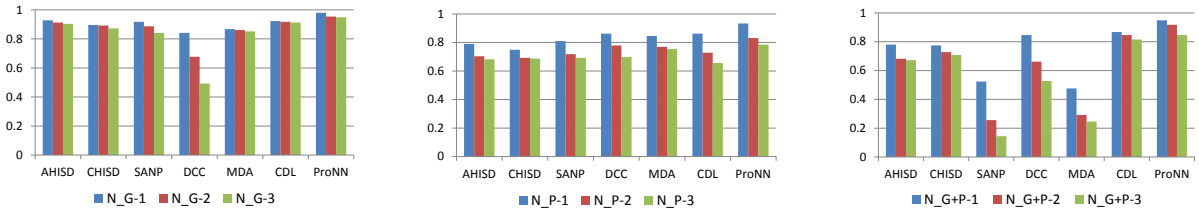
The recognition results on face image sets contaminated by face detection errors are also shown in Tab. I. Note that each reported rate is an average over multiple-fold trials. The images of these databases are complex and have large variations, especially in MBGC because of the low resolution, high compression ratio and large illumination, pose and expression changes of face images. Therefore, face recognition on MBGC contaminated by detection errors is relatively difficult and challenging, which leads to commonly low identification rates by all the comparison methods in our experiments. Nevertheless, from Tab. I, it is obvious that our approach still performs reasonably well.

Compared with results of AHISD, CHISD and SANP, which all model image sets similarly with our approach, but do not consider structure of data, our ProNN performs much better than these competitors on all of the three databases. Moreover, our approach, as an unsupervised method, has achieved comparable performance with the three supervised methods, DCC, MDA and CDL.

2) Comparison results on sets contaminated by images from other classes: On the other hand, we conduct simulation experiments to evaluate the performance of our approach when image sets contain group photos. Specifically, with the clean image sets of Honda/UCSD and YouTube

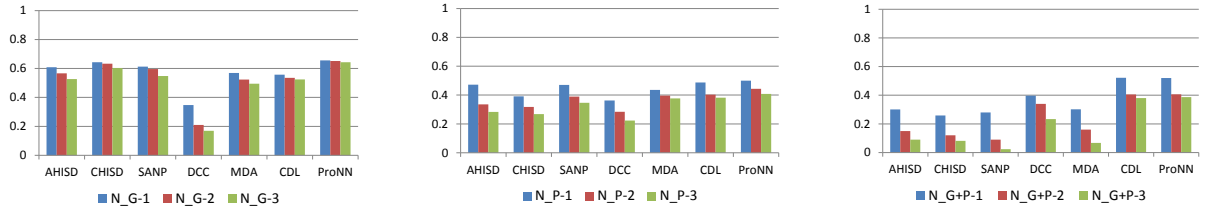
TABLE I: Identification rates on face image sets contaminated by face detection errors.

Scenes \ Methods		AHISD [15]	CHISD [15]	SANP [16]	DCC [7]	MDA [11]	CDL [3]	**ProNN**
Honda	Clean	0.892	0.908	0.928	0.867	0.933	0.969	**0.995**
	Noise	0.651	0.662	0.601	0.682	0.657	0.742	**0.764**
YouTube	Clean	0.637	0.665	0.684	0.668	0.670	**0.697**	0.671
	Noise	0.317	0.312	0.276	0.477	0.482	**0.499**	0.493
MBGC	S_1	0.181	0.193	0.247	0.174	0.257	0.292	**0.382**
	S_2	0.182	0.200	0.263	0.182	0.291	0.346	**0.394**
	S_3	0.204	0.222	0.275	0.188	0.324	0.371	**0.426**



(a) N_G (b) N_P (c) N_G+P

Fig. 5: Identification rates on the Honda/UCSD database contaminated by images from other classes. Here, "N_G" is the experimental scenario where only gallery is contaminated, "N_P" denotes that only probe is contaminated, and "N_G+P" denotes both are contaminated. The gallery and/or probe sets are contaminated by 1, 2 or 3 images (denoted as "-1/-2/-3" in the figure) from each of all the other classes. Note that each reported rate is an average over multiple folds.



(a) N_G (b) N_P (c) N_G+P

Fig. 6: Identification rates on the YouTube Celebrities database contaminated by images from other classes. Here, all the denotations are the same as that in Fig. 5.

TABLE II: Computation time (seconds) of different methods on the Honda/UCSD database for training and testing.

	AHISD [15]	CHISD [15]	SANP [16]	DCC [7]	MDA [11]	CDL [3]	**ProNN**
Training	N/A	N/A	N/A	3.48	4.91	1.88	N/A
Testing	5.57	6.57	61.03	1.47	6.44	1.19	7.41

Celebrities database, we corrupt them by adding images from other classes, as done in [3]. The recognition results are shown in Fig. 5 and Fig. 6. Note that each reported rate is an average over multiple-fold trials. Here, "N_G" is the experimental setting where only gallery face image sets are contaminated, "N_P" denotes that only probe face image sets are contaminated, and "N_G+P" denotes they are both contaminated. The gallery and/or probe face image sets are contaminated by 1, 2 or 3 images (denoted as "-1/-2/-3" in the figure) from each of the image sets belonging to other classes.

From the comparison results shown in Fig. 5 and Fig. 6, it is obvious that our method gives better performance across different types of noise on both databases except in N_G+P with 1 image/class on YouTube where CDL obtains a higher rate of 52.2% compared with 52.0% of our approach. Note that CDL also models statistical nature of image sets. From the comparison results, it can be concluded that robustness against outliers can be enhanced by taking the statistical structure of data into consideration.

Comparing the accuracies when different number of images per class are added to contaminate the gallery and/or the probe set, we can find that our ProNN method has a smaller decline scope as the number of outliers increases, and it performs robust classification even when a relatively large number of noises are contained.

Lastly, we compared the computational complexity of different methods on the Honda/UCSD database, where each fold contains 20 training face image sets and 39 testing face image sets belonging to 20 subjects. The experiments are conducted on an Intel i7-3770, 3.40 GHz PC. Since our approach is not a discriminative method, training stage is

not involved. For testing, we report the time of matching one probe image set against 20 gallery image sets. The time cost is shown in Tab. II. We can see that the computational time of our approach is comparable to the state-of-the-art methods.

V. Conclusions

This paper has proposed a Probabilistic Nearest Neighbor (ProNN) search method for robust image set classification. We attempt to enhance the robustness against outliers by considering both nearest distance and the statistical structure of each face image set. We represent image sets by the affine hull models to account for the underlying appearances and simultaneously characterize the statistical structure of these appearances by making a natural assumption that they follow some distribution. In our current study, the distribution function is simply assumed as a Gaussian distribution and was estimated by samples contained in the image sets. By minimizing the derived objective function using Conjugate Gradient (CG) algorithm, we search for ProNNs, which enhances the robustness against outliers and better characterizes the dissimilarity between image sets.

To evaluate the performance of our approach, we have conducted extensive experiments under simulation situations when image sets are contaminated with different types of noises respectively. The experiments have demonstrated the superiority of the proposed approach over state-of-the-art methods.

Currently we only make a simple prior assumption that images are generated i.i.d from some distribution which might not be always met in real applications. In the future, we will explore more flexible models to characterize the statistical distribution of the image set, such as kernel density estimation. Furthermore, the simple yet appealing idea of probabilistic set modeling can also be extended to work with other nearest neighbor search methods, and even to be incorporated into more sophisticated set models.

VI. Acknowledgments

This work is partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contracts Nos. 61222211, 61379083, and 61390511.

References

[1] G. Shakhnarovich, J. W. Fisher and T. Darrell, "Face recognition from long-term observations", *IEEE European Conference on Computer Vision (ECCV)*, 2002, pp. 851–865.

[2] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla and T. Darrell, "Face recognition with image sets using manifold density divergence", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 581–588.

[3] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2496–2503.

[4] J. Lu, G. Wang and P. Moulin, "Image Set Classification Using Holistic Multiple Order Statistics Features and Localized Multi-kernel Metric Learning", *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 329–336.

[5] R. Vemulapalli, J. K. Pillai, and R. Chellappa, "Kernel learning for extrinsic classification of manifold features", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1782-1789.

[6] O. Yamaguchi, K. Fukui and K. Maeda, "Face recognition using temporal image sequence", *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 1998, pp. 318–323.

[7] T. Kim, J. Kittler and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations", *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2007, pp. 1005–1018.

[8] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: a unifying view on subspace-based learning", *Proceedings of the 25th international conference on Machine learning (ICML)*. 2008, pp. 376–383

[9] M. T. Harandi, C. Sanderson, S. Shirazi and B. C. Lovell, "Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 2705–2712

[10] R. Wang, S. Shan, X. Chen and W. Gao, "Manifold-manifold distance with application to face recognition based on image set", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1-8.

[11] R. Wang, and X. Chen, "Manifold discriminant analysis", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 429–436.

[12] Z. Cui, S. Shan, H. Zhang, S. Lao and X. Chen, "Image sets alignment for Video-Based Face Recognition", *in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2626-2633.

[13] S. Chen, C. Sanderson, M. Harandi and B. C. Lovell, "Improved Image Set Classification via Joint Sparse Approximated Nearest Subspaces", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 452–459.

[14] M. Hayat, M. Bennamoun, and S. An, "Learning non-linear reconstruction models for image set classification", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1915–1922.

[15] H. Cevikalp and B. Triggs, "Face recognition based on image sets", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2567–2573.

[16] Y. Hu, A. S. Mian and R. Owens, "Sparse approximated nearest points for image set classification", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 121–128.

[17] S. Chen, A. Wiliem, C. Sanderson and B. C. Lovell, "Matching Image Sets via Adaptive Multi Convex Hull", arXiv:1403.0320.

[18] M. Yang, P. Zhu, L. V. Gool and L. Zhang, "Face recognition based on regularized nearest points between image sets", *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1-7.

[19] K. C. Lee, J. Ho, M. H. Yang and D. Kriegman, "Video-Based Face Recognition Using Probabilistic Appearance Manifolds", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003, pp. 313–320.

[20] M. Kim, S. Kumar, V. Pavlovic and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[21] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scruggs, A. J. O'Toole, D. Bolme, K. W. Bowyer, B. A. Draper, G. H. Givens, Y. M. Lui, H. Sahibzada, J. A. Scallan III, and S. Weimer, "Overview of the multiple biometrics grand challenge", *International Conference on Biometrics*, 2009.

[22] National Institute of Standards and Technology: "Multiple biomertic grand challenge (MBGC)."
http://www.nist.gov/itl/iad/ig/mbgc.cfm

[23] P. Viola and M. J. Jones, "Robust Real-Time Face Detection", *International Journal of Computer Vision (IJCV)*, 2008, no. 2, pp. 137–154.

[24] Y. C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. "Dictionary-based face recognition from video", *IEEE European Conference on Computer Vision (ECCV)*, 2012, pp. 766-779.

[25] Y. C. Chen, V. M. Patel, S. Shekhar, R. Chellappa and P. J. Phillips. "Video-based face recognition via joint sparse representation", *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1-8.