# Sparse Approximated Nearest Points for Image Set Classification

Yiqun Hu Ajmal S. Mian Robyn Owens
School of Computer Science & Software Engineering
The University of Western Australia
{yiqun, ajmal}@csse.uwa.edu.au, robyn.owens@uwa.edu.au

## Abstract

*Classification based on image sets has recently attracted great research interest as it holds more promise than single image based classification. In this paper, we propose an efficient and robust algorithm for image set classification. An image set is represented as a triplet: a number of image samples, their mean and an affine hull model. The affine hull model is used to account for unseen appearances in the form of affine combinations of sample images. We introduce a novel between-set distance called Sparse Approximated Nearest Point (SANP) distance. Unlike existing methods, the dissimilarity of two sets is measured as the distance between their nearest points, which can be sparsely approximated from the image samples of their respective set. Different from standard sparse modeling of a single image, this novel sparse formulation for the image set enforces sparsity on the sample coefficients rather than the model coefficients and jointly optimizes the nearest points as well as their sparse approximations. A convex formulation for searching the optimal SANP between two sets is proposed and the accelerated proximal gradient method is adapted to efficiently solve this optimization. Experimental evaluation was performed on the Honda, MoBo and Youtube datasets. Comparison with existing techniques shows that our method consistently achieves better results.*

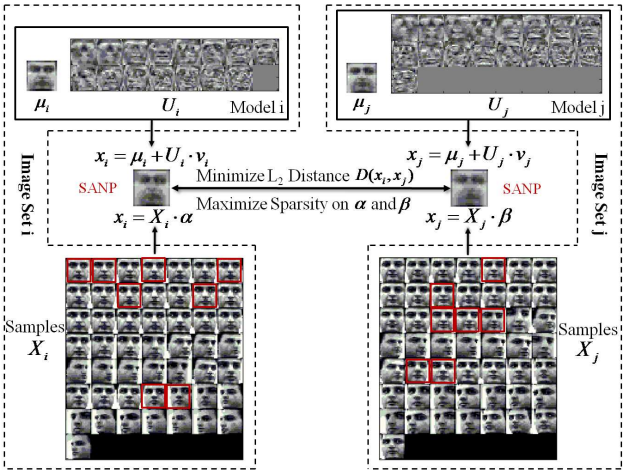


Figure 1. Sparse Approximated Nearest Points (SANPs) of two image sets. Given the affine hull models $(\mu_i, U_i)$ and $(\mu_j, U_j)$ of two image sets, the points on each set can be represented as a linear combination of bases plus the mean image. They can also be represented as the linear combination of sample images. The SANPs are dynamically chosen by the joint optimization which simultaneously searches for sparse approximated points (maximize sparsity of sample coefficients) that are the nearest (minimize distance) between the two sets. The optimal SANPs of the two image sets are shown in the center, each of which is sparsely approximated by the sample images marked with red boxes.

## 1. Introduction

In image set classification, each class is represented by one or more image sets and a query image set is assigned the label of the gallery set that is the nearest to it using some distance criterion. For the specific case of human faces, each set comprises a different number of facial images under arbitrary poses, illumination conditions and expressions. Image set classification is a generalization of video-based classification [15, 20, 8], which focuses on exploiting the temporal relationship between the images with the priori that individual images are consecutive video frames. However, in image set based classification, the images of a set may manifest large view-point and illumination changes and non-rigid deformations without any temporal relationship.

Classification based on image sets has recently attracted growing interest in the computer vision community [23, 28, 9, 12, 25, 19, 18, 29, 5] because it holds more promise compared to single image based classification. The problem of image set classification naturally arises in a wide range of applications including video surveillance, classification based on images from different views using multiple cameras, relevant pictures of a personal album and classification based on long term observations. Within a set, individual images either share the common semantic relationship or complement the appearance variations of the subject. Im-

age set data pose new challenges to the visual classification task. The main challenge relates to modelling the image set in order to exploit the semantic knowledge between individual images. Traditional classification models e.g. SVM, k-Nearest Neighbor Classifier based on single sample, cannot address this issue.

In this paper, we propose a novel algorithm for image set classification. In the proposed method, every image set is represented as a triplet including the sample images, their mean and the affine hull model of its sample images to cover all possible affine combinations of sample images. Such a loose representation of the affine hull is capable of accounting for the unseen appearances of any affine combination of sample images which do not appear in the samples. On the downside, it also introduces a challenge for matching different image sets. The image sets of different classes are more likely to intersect due to the over-large space of their affine hulls. To address this issue, we introduce the Sparse Approximated Nearest Points (SANP) for computing the between-set distance. SANPs of two image sets are defined as a pair of nearest points on the sets that can be sparsely approximated by the sample images of the respective set individually.

The search for SANP of two image sets is formulated as a partial $L_1$ norm regularized convex optimization. Fig 1 illustrates the formulation of SANP optimization for two image sets. This novel formulation is different from the sparse modeling of single images in two aspects. First, the nearest points to be sparsely approximated are the unknowns which means that we need to jointly optimize the nearest points and their sparse approximations. Second, the sparsity is enforced on the sample coefficients instead of model coefficients in our formulation. We show how recent advances in first-order optimization techniques can be adapted to solve this optimization, leading to a fast, scalable algorithm. Once the SANPs are found, the between-set distance is then defined using these points and the Nearest Neighbor (NN) classifier is deployed to assign the query set to the class of its nearest neighbor. Experiments on three benchmark datasets and comparison with existing techniques [25, 19, 18, 5] show that the proposed method consistently gives better results.

### 1.1. Related Work

Image set classification techniques can be categorized based on two criterion: firstly, how the image sets are represented and secondly, how the between-set distance is defined. As far as image set representation is concerned, existing techniques include parametric and non-parametric representations. Parametric model-based representations [4, 8, 14, 29] use some parametric distributions to represent an image set with the parameters estimated from the set data itself. A limitation of these techniques is that if the set data does not have strong statistical correlations for parameter estimation, the estimated model cannot well characterize the image set [25, 19]. Non-parametric model-free methods attempt to represent an image set as a linear subspace [15, 13, 25, 9], mixture of subspaces [28, 12], or nonlinear manifolds [2, 19, 18, 27]. Without any assumption on data distribution, it has been shown that these model-free representations inherit many favorable properties.

Existing techniques can also be differentiated based on the second criterion of between-set distance. However, the between set distance is usually defined specifically for certain image set representations. For example, for parametric representations, the between-set distance is calculated by measuring the similarity between the corresponding distributions of their parameters. Kullback-Leibler divergence [14] is an example of this category. For non-parametric representations, two types of distances have been proposed. The first one defines the between set distance using some of the set samples. For example, a simple method for calculating the between-set distance is to measure the distance between the sample means of the two sets [19]. Another example is to use geometric distances (distances of closest point approach) [5] to compare different image sets. Unlike the mean difference, this method adaptively selects different samples to calculate the between set distance for different image sets. Thus it is able to better handle intra-class variations. Given two image sets, the closest points are obtained by minimizing the distance between them through least square optimization. The between-set distance is then defined as the distance between these two closest points.

The second type of distance for non-parametric representations compares different image sets by analyzing their model structures instead of the sample data. Canonical Correlation Analysis (CCA) [6] is one of the techniques for calculating subspace similarity. It finds $d$ *principal angles* $0 \leq \theta_1 \leq \ldots \leq \theta_d \leq \pi/2$ between the subspaces of two sets, which are the smallest angles between any vector in the first set and any vector in the second set. The between-set similarity is then defined as the sum of canonical correlations, which are the cosines of principal angles. Different extensions have been proposed for using CCA to match image sets. For example, kernelized CCA [10] and localized CCA [30] have been proposed for image set-based classification. Boosting techniques have been applied on principal angles for improving the classification performance of CCA [26, 24]. Other methods for subspace similarity analysis include the Mutual Subspace Method (MSM) [15], Orthogonal Subspace Method (OSM) and their variants [13, 23, 9, 29, 22]. These techniques are further extended by integrating discriminant learning [25] and online learning [23, 22].

One common aspect about the above techniques is that they either measure the distance between certain samples of

the two sets or the similarity between their structures. On the other hand, the proposed technique in this paper tries to utilize both structure information and the image samples.

## 2. Image Set Representation

In this paper, we propose a joint representation for image sets which consists of different numbers of images. Denote $X_c = [x_{c,1}, x_{c,2}, \cdots, x_{c,N_c}]$ as the data matrix of the $c^{th}$ image set, where $x_{c,i}$ is a feature vector of the $i^{th}$ image. The feature of an individual image can be simply the high dimensional array of pixel values or any other features e.g. Local Binary Pattern (LBP) [21] of the image. The joint representation, besides using the sample data $X_c$, constructs a linear model to approximate the structure of the image set in high-dimensional feature space. We model an image set as an affine hull of the set data [5]:

$$AH_c = \{x = \sum_{i=1}^{N_c} \alpha_{c,i} \cdot x_{c,i} \mid \sum_{i=1}^{N_c} \alpha_{c,i} = 1\}. \quad (1)$$

This affine hull can also be represented by another parametric form using the sample mean $\mu_c = \frac{1}{N_c}\sum_{i=1}^{N_c} x_{c,i}$ as a reference point to represent every data:

$$AH_c = \{x = \mu_c + U_c v_c \mid v_c \in \mathcal{R}^l\}, \quad (2)$$

where the $l$ columns of $U_c$ are the orthonormal bases obtained from the Singular Value Decomposition (SVD) of the centered data matrix $\bar{X}_c = [x_{c,1} - \mu_c, x_{c,2} - \mu_c, \ldots, x_{c,N_c} - \mu_c]$. Different from other tight representations e.g. convex hull, any affine combination of sample images in the set is accommodated in this representation even when the combination does not appear in the samples of the set. Such a loose representation is particularly appealing in the context of small set size because the unseen data belonging to the image set can be better modeled. However, this loose representation also brings challenges for calculating the distance between two image sets. The affine hulls of image sets are likely to be over-large, which results in the intersection of multiple affine hulls. In this paper, we represent an image set as a triplet $(\mu_c, U_c, X_c)$ by including both structure information and sample images. As we will show in the next section, the information of sample images can be utilized to eliminate the ambiguity of the over-large space of the affine hulls. This joint representation of image set is useful for improving the robustness of matching image sets.

## 3. Sparse Approximated Nearest Points

Existing methods [19, 5] directly search the nearest points in the complete space of two image sets without any additional constraints. These points could be very noisy and vulnerable to outliers. This issue is especially serious in our case because we use loose affine hulls to model image sets. Even for two image sets of different classes, it is possible to find two nearest points with very small distance. This can degrade the classification performance. To overcome this problem, we propose *Sparse Approximated Nearest Points (SANP)* to measure the dissimilarity between two image sets. SANPs are the two points, one on each individual set, which satisfy the following constraints:

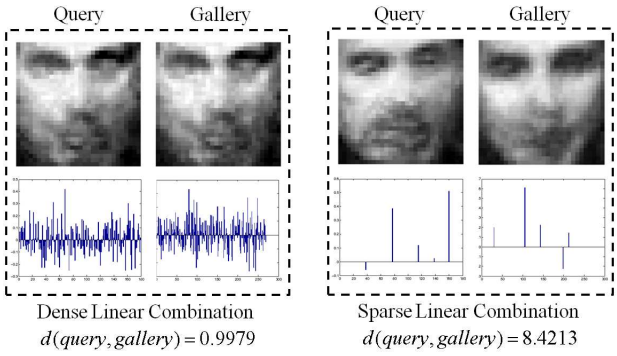


Figure 2. Illustration of matching two image sets of different subjects. Nearest points of the sets (a) with dense approximation and (b) with sparse approximation. First row shows the images of nearest points on the sets and second row shows the sample coefficients used for approximation from samples.

- The Euclidean distance between these two points should be small;
- Each of the two points should be able to be approximated by a sparse combination of sample images in the corresponding image set.

Note that the second constraint improves the discriminative power of the SANPs. Given two image sets of different classes (subjects), the nearest points between these two sets using a dense combination of all sample images could be very close. For example, the nearest images in Fig 2(a) (top row) are very close but they deviate significantly from the sample images of the respective set i.e. they neither look like the query nor the gallery. Alternatively, using only a sparse combination of a few samples, the minimum distance between points of the same two sets (correctly) becomes large (e.g. the images in Fig 2(b) are approximated by the linear combination of 5 samples). From a geometric point of view, the affine hull of an image set is formed from sample images which lie on the facets of the hull. The constraint of sparse approximation enforces the SANPs to be close to some facet(s) of the affine hull and consequently close to some sample image(s) on those facet(s). With this constraint, the spurious nearest points of image sets of different classes can be avoided.

### 3.1. Convex Formulation

To find the SANPs of two image sets which are optimal in terms of the above two criteria, we propose a convex formulation. Given the data matrices $X_i$ and $X_j$ of two image sets, their corresponding affine hull representations are $(\mu_i, U_i)$ and $(\mu_j, U_j)$. We first define several functions as follows:

$$\begin{aligned} F_{v_i,v_j} &= |(\mu_i + U_i \cdot v_i) - (\mu_j + U_j \cdot v_j)|_2^2 \\ G_{v_i,\alpha} &= |(\mu_i + U_i \cdot v_i) - X_i \cdot \alpha|_2^2 \\ Q_{v_j,\beta} &= |(\mu_j + U_j \cdot v_j) - X_j \cdot \beta|_2^2. \end{aligned} \quad (3)$$

The optimal model coefficients $\{v_i^*, v_j^*\}$ and sample coefficients $\{\alpha^*, \beta^*\}$ of SANPs are obtained by optimizing the following unconstrained problem:

$$\min_{v_i,v_j,\alpha,\beta} F_{v_i,v_j} + \lambda_1(G_{v_i,\alpha} + Q_{v_j,\beta}) + \lambda_2|\alpha|_1 + \lambda_3|\beta|_1, \quad (4)$$

where the first term is to keep the distance between SANPs $x_i = \mu_i + U_i \cdot v_i$ and point $x_j = \mu_j + U_j \cdot v_j$ small. The second term is to preserve the individual fidelities between these two points and their sample approximations. The last two terms enforce the approximations to be sparse. $\lambda_1$, $\lambda_2$ and $\lambda_3$ are the trade-off weights to control the relative importance of different terms. The value of $\lambda_1$ is fixed as $0.01$ for all the experiments conducted in this paper. For $\lambda_2$ and $\lambda_3$, we design an automatic mechanism to control the relative sparsity of $\alpha$ and $\beta$. Notice that if $\lambda_2 >= \lambda_2^* = \max(|2\lambda_1 \cdot (X_i^T \mu_i)|)$, the zero vector is optimal for $\alpha$ at zero. Similarly, if $\lambda_3 >= \lambda_3^* = \max(|2\lambda_1 \cdot (X_j^T \mu_j)|)$, the zero vector is optimal for $\beta$ at zero. We adaptively set $\lambda_2 = 0.1 \cdot \lambda_2^*$ and $\lambda_3 = 0.1 \cdot \lambda_3^*$ for all experiments.

To the best of our knowledge, this is the first time that sparse modeling has been formulated to match two image sets. Note that we do not enforce the sparsity on the model coefficients $v_i$ and $v_j$, because the bases $U_i$/$U_j$ obtained from SVD do not align with the sample data points. Instead, we enforce the sparsity property on the sample coefficients $\alpha$ and $\beta$, which imply that each nearest point is sparsely approximated by the combination of a few sample images. Different from sparse modeling of single image classification [7], our formulation jointly optimizes the nearest points between two image sets and their sparse approximations from samples.

## 4. Efficient Optimization

In this section, we provide an efficient solution to the optimization problem in (4) which is summarized in Algorithm 1. The objective function in (4) is a composite model consisting of a smooth function and a non-smooth function. The smooth part corresponds to $f(v1, v2, \alpha, \beta) = F_{v_i,v_j} + \lambda_1(G_{v_i,\alpha} + Q_{v_j,\beta})$ and the non-smooth part is $g(\alpha, \beta) = \lambda_2|\alpha|_1 + \lambda_3|\beta|_1$. Obviously, $g(\alpha, \beta)$ is a convex function with respect to $\alpha$ and $\beta$. It can also be proved that the smooth function $f(v1, v2, \alpha, \beta)$ is jointly convex with respect to all its variables. Hence, the objective function in (4) is convex and the global minimum solution can be obtained. In the rest of this section, we adapt the Accelerated Proximal Gradient (APG) methods [31, 1] to solve this optimization problem, which can achieve the optimal convergence rate of first order methods.

**Algorithm 1** Optimization of SANPs

**Require:** $(X_i, \mu_i, U_i)$, $(X_j, \mu_j, U_j)$

1: Set $v_i^1 = v_i^0 = \mathbf{0}$, $v_j^1 = v_j^0 = \mathbf{0}$, $\alpha^1 = \alpha^0 = \mathbf{0}$, $\beta^1 = \beta^0 = \mathbf{0}$, $t_0 = 0$, $t_1 = 1$, $k = 1$, $L = L_0 = 100$, $\eta = 1.1$, $\lambda_1 = 0.01$, $\lambda_2 = 0.1 \cdot max(|2\lambda_1 \cdot (X_i^T \mu_i)|)$ and $\lambda_3 = 0.1 \cdot max(|2\lambda_1 \cdot (X_j^T \mu_j)|)$.

2: **while** not converged **do**

3: compute the proximal points:
$y_{v_i}^k = v_i^k + \frac{t^{k-1}-1}{t^k}(v_i^k - v_i^{k-1})$;
$y_{v_j}^k = v_j^k + \frac{t^{k-1}-1}{t^k}(v_j^k - v_j^{k-1})$;
$y_\alpha^k = \alpha^k + \frac{t^{k-1}-1}{t^k}(\alpha^k - \alpha^{k-1})$;
$y_\beta^k = \beta^k + \frac{t^{k-1}-1}{t^k}(\beta^k - \beta^{k-1})$;

4: calculate gradient:
$\bigtriangledown f_{v_i} = (2 + 2\lambda_1)U_i^T U_i y_{v_i}^k - 2U_i^T U_j v_j^{k-1} - 2U_i^T \mu_j + (2 + 2\lambda_1)U_i^T \mu_i - 2\lambda_1 U_i^T X_i \alpha^{k-1}$;
$\bigtriangledown f_{v_j} = (2 + 2\lambda_1)U_j^T U_j y_{v_j}^k - 2U_j^T U_i v_i^{k-1} - 2U_j^T \mu_i + (2 + 2\lambda_1)U_j^T \mu_j - 2\lambda_1 U_j^T X_j \beta^{k-1}$;
$\bigtriangledown f_\alpha = 2\lambda_1 X_i^T X_i y_\alpha^k - 2\lambda_1 X_i^T \mu_i - 2\lambda_1 X_i^T U_i v_i^{k-1}$;
$\bigtriangledown f_\beta = 2\lambda_1 X_j^T X_j y_\beta^k - 2\lambda_1 X_j^T \mu_j - 2\lambda_1 X_j^T U_j v_j^{k-1}$;

5: optimize proximal regularization:
$v_i^{k+1} = y_{v_i}^k - \frac{1}{L} \bigtriangledown f_{v_i}$; $v_j^{k+1} = y_{v_j}^k - \frac{1}{L} \bigtriangledown f_{v_j}$;
$\alpha^{k+1} = \tau_{\frac{\lambda_2}{L}}(y_\alpha^k - \frac{1}{L} \bigtriangledown f_\alpha)$;
$\beta^{k+1} = \tau_{\frac{\lambda_3}{L}}(y_\beta^k - \frac{1}{L} \bigtriangledown f_\beta)$;

6: If $F_{v_i^{k+1},v_j^{k+1}} + \lambda_1 \cdot (G_{v_i^{k+1},\alpha^{k+1}} + Q_{v_j^{k+1},\beta^{k+1}}) > P_L$, update $L = \eta L$ and go to Step 5;

7: stepsize update:
$t^{k+1} = \frac{1+\sqrt{4(t^k)^2+1}}{2}$;

8: **end while**

9: **Output:** optimal solution $(v_i^*, v_j^*, \alpha^*, \beta^*)$ to (4)

The gradient method [31, 1] was used to minimize the composite function $f(w) + g(w)$ by extending the equivalence relationship between gradient step and the proximal regularization of the linearized function $f$ at $w_{k-1}$ to the composite function $f(w) + g(w)$. The corresponding iterative scheme is as follows. At every iteration $k$, the new solution $w_k$ is obtained by solving the following proximal regularization problem from the solution $w_{k-1}$ at the previ-

ous iteration:

$$w_k = \arg\min_w \{P_L(w, w_{k-1}) + g(w)\}, \quad (5)$$

where
$P_L(w, w_{k-1}) =$

$$f(w_{k-1}) + < \bigtriangledown f(w_{k-1}), w - w_{k-1} > + \frac{L}{2}||w - w_{k-1}||^2. \quad (6)$$

When $g(w) = \lambda|w|_1$, the optimal $w_k$ of (5) can be efficiently obtained by the soft-thresholding operators at every iteration as follows:

$$\tau_\alpha(x)_i = (|x_i| - \alpha)_+ sgn(x_i), \quad (7)$$

where $(x)_+ = max(0, x)$ and $sgn(x)$ returns the sign of $x$. APG methods [31, 1] improve the convergence rate of the gradient method from $o(\frac{1}{k})$ to $o(\frac{1}{k^2})$ by carefully selecting a sequence of points $Y^k$ for proximal regularization instead of directly using the point in the previous iteration (Step 1 in Algorithm 1).

The composite objective function (4) of our SANP optimization is different from the standard one in the non-smooth part, where $L_1$ norm only relates to some optimization variables ($\alpha$ and $\beta$). Because the objective function is separable, the proximal regularization of SANP optimization at every iteration still can be solved efficiently: $v_i$ and $v_j$ are directly updated from proximal points in the negative gradient direction since they are independent of the non-smooth part; $\alpha$ and $\beta$ are updated using the soft-thresholding operator (7) with the thresholding value of $\frac{\lambda}{L}$ (Step 5 in Algorithm 1). The stepsize $L$ is related to the Lipschitz constant of $\bigtriangledown f$, which unfortunately is unknown. We adaptively select the stepsize using the backtracking rule [1]. Given the initial $L = L_0$ and some $\eta > 1$, we keep updating $L = \eta L$ until $P_L$ between the solutions of iteration $k+1$ and $k$ is larger than $F_{v_i,v_j} + \lambda_1 \cdot (G_{v_i,\alpha} + Q_{v_j,\beta})$ at iteration $k+1$ (Step 6 in Algorithm 1).

### 4.1. Convergence Rate Analysis

Following the more general results in [1], it can be proven that the sequence $p^k = (v_i^k, v_j^k, \alpha^k, \beta^k)$ generated by Algorithm 1 converges to the global solution $p^* = (v_i^*, v_j^*, \alpha^*, \beta^*)$ of the function (4) with a non-asymptotic convergence rate of $O(\frac{1}{k^2})$, where $k$ is the iteration number. Compared to the general gradient method whose convergence rate is $O(\frac{1}{k})$, this convergence rate is optimal for the first-order optimization methods. Actually, it can be shown that

$$F(p^k) - F(p^*) \leq \frac{2\eta L(f)||p^k - p^*||^2}{(k+1)^2}, \quad (8)$$

where $\eta > 1$ is the constant for backtracking the update of stepsize and $L(f)$ is the Lipschitz constant of $\bigtriangledown f$. To achieve the $\varepsilon$-optimal solution (i.e. a $\tilde{p}$ such that $F(\tilde{p}) - F(p^*) \leq \varepsilon$), the number of required iterations is at most $\lceil \frac{C}{\sqrt{\varepsilon}} - 1 \rceil$, where $C = \sqrt{2\eta L(f)||p^0 - p^*||^2}$. Fig 3 plots the values of the objective function (4) over iterations when computing the SANPs of two image sets. The algorithm quickly converges after $40$ iterations in $0.8$ second using a Matlab implementation on a $2.3$GHz machine.

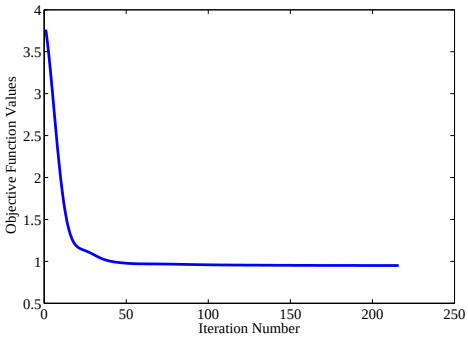


Figure 3. Illustration of fast convergence of SANP optimization.

## 5. Experimental Evaluation

We evaluate the proposed method on the task of face recognition based on image sets. Once the SANPs are found, the nearest neighbor classifier is used for recognition. For every query set, the most similar image set in the gallery is searched by finding the minimum between-set distances based on the SANPs of two image sets. We define the between-set distance as follows

$$D(c_i, c_j) = (d_i + d_j) \cdot [F_{v_i^*, v_j^*} + \lambda_1 (G_{v_i^*, \alpha^*} + Q_{v_j^*, \beta^*})], \quad (9)$$

where $(v_i^*, v_j^*, \alpha^*, \beta^*)$ is the optimal solution of (4) and $d_i$ and $d_j$ are the dimensions of the affine hulls of $c_i$ and $c_j$, respectively. Multiplication with the factor $(d_i + d_j)$ is performed to eliminate the bias to larger image sets. The bias occurs because, when calculating the distance to the larger image sets, the error of the least square function $F_{v_i,v_j}$, which is the projection of $\mu_j - \mu_i$ onto the null space of $[U_i, -U_j]$, will be smaller since the dimension of the null space is reduced. In the extreme case, if $d_i + d_j$ is larger than the feature dimension, a zero minimum distance can be obtained even when the two image sets are very different. Multiplication with $(d_i + d_j)$ ensures that a small between-set distance is only obtained when the distance between SANPs and the dimensions of sets are both small.

### 5.1. Experiment Setup

**Dataset Configuration:** We used the Honda/UCSD [8], CMU Mobo [17] and Youtube Celebrities [11] datasets in our experiments. Honda/UCSD dataset contains $59$ video sequences of $20$ different subjects. Different poses and expressions appear across different sequences of each subject.

Each video sequence corresponds to an image set. The faces in every frame are detected using [16] and then resized to gray-scale images of size $20 \times 20$ as in [18]. The lengths of the sets vary from $12$ to $645$. Histogram equalization is the only pre-processing step used to minimize the illumination variations. For this dataset, we directly vectorize the raw pixels of the resized images to form the columns of data matrix $X$.

Mobo (Motion of Body) dataset [17] was originally created for human pose identification. There are $96$ sequences of $24$ subjects walking on a treadmill. Multiple cameras were used to capture videos of four walking patterns: slow, fast, inclined and carrying a ball. For each subject, $4$ video sequences are collected each of which corresponds to a walking pattern. The faces are detected in every frame as before using [16] and then resized to $40 \times 40$ gray-scale images. For this dataset, we use the Local Binary Pattern (LBP) [21] as the features of individual images. The uniform LBP histogram using circular $(8, 1)$ neighborhoods is extracted from the $8 \times 8$ squares of gray-scale images.

We also provide experimental results on Youtube Celebrities [11], which is a large video dataset collected for face tracking and recognition. $1910$ video sequences of $47$ celebrities (actors, actresses and politicians) are collected from Youtube. The clips contain different number of frames (from $8$ to $400$) which are mostly low resolution and highly compressed. This database [11] only provides the cropped face in the first frame. Therefore, we apply [3] to track faces and resize them to $30 \times 30$ gray-scale images. The pixel values are used as features. This dataset introduces more challenging situations for image set classification because of two reasons. First, the video sequences exhibit larger variations in pose, illumination and expressions. Second, the low quality of frames, due to the high compression rate, introduces tracking errors and noises in the cropped faces. Without enforcing facial constraints as in [11], the cropped faces we used in this paper contain larger tracking errors than the face images from [11], which makes our experimental setting even more challenging.

**Comparison with Existing Methods:** We compare the proposed method with several image set classification methods lately proposed in the literature. They include Discriminant Canonical Correlation Analysis (DCC) [25], Manifold-to-Manifold Distance (MMD) [19], Manifold Discriminant Analysis (MDA) [18], Linear version of Affine Hull based Image Set Distance (AHISD) [5] as well as Convex Hull based Image Set Distance (CHISD) [5]. Here, AHISD can be regarded as a baseline method which finds the nearest neighbors without the sparsity constraint. Note that [25, 19, 18, 5] have conducted extensive comparisons with exemplar-based methods e.g. Linear Discriminant Analysis (LDA) and Marginal Fisher Analysis (MFA) have shown that set-based methods generally outperform exemplar-based methods. Due to this reason and paucity of space, we do not provide comparison with exemplar-based methods.

The standard implementations of all methods from the original authors are used except MDA. We carefully implement the MDA algorithm since it is not publicly available. The important parameters of different methods are carefully optimized as follows: For DCC, the dimension of the embedding space is set to $100$. The subspace dimensions are set to $10$ which preserves $90\%$ energy and the corresponding $10$ maximum canonical correlations are used to define set similarity. For MMD and MDA, the parameters are configured according to [19, 18]. Specifically, the ratio between Euclidean distance and geodesic distance is optimized for different datasets (i.e. 2.0 for Honda, 5.0 for Mobo and 2.0 for Youtube dataset[1]). The maximum canonical correlation is used in defining MMD. For MDA, the number of between-class NN local models and the dimension of MDA embedding space are tuned for different datasets as specified in [18]. The number of connected nearest neighbors for computing geodesic distance in both MMD and MDA is fixed to its default value i.e. $12$. There is no parameter setting for AHISD. For CHISD, we set the error penalty parameter to be the same value as in [5] ($C = 100$ for gray-scale features and $C = 50$ for LBP in linear SVM). Both methods apply PCA to preserve $90\%$ energy as before.

### 5.2. Results and Analysis

On the Honda dataset, we use the standard training/testing configuration provided in [8]: $20$ sequences are used for training and the remaining $39$ sequences for testing. We report results using all frames as well as with a limited number of frames. Specifically, we conduct the experiments by setting an upper bound $M$ of maximum set length to $100$ and $50$. In case a set contains fewer than $M$ images, all images are used for classification. Such situations often occur in real-world applications, for example the tracking of a face may fail for a long sequence and only the first part of the sequence is available for classification. Moreover, classification based on smaller sets can also be more efficient. Table 1 summarizes the identification rates of all methods. We can see that the proposed method achieves the best overall performance in different situations. When the whole sequences are used, it achieves perfect classification. When the sets are reduced, our method achieves the $2^{nd}$ highest performance. It is interesting to notice that the performances of discriminant learning methods (DCC and MDA) degrade more heavily due to the reduction of training data. Geometric models (AHISD and CHISD) perform more consistently across different set lengths with lower accuracy. Note that the accuracies of AHISD and CHISD are

[1]The optimal parameter for Mobo dataset is different because the LBP histograms are used in this case.

Table 1. Identification rates on Honda/UCSD Dataset

| Set Length/Methods | DCC [25] | MMD [19] | MDA [18] | AHISD (linear) [5] | CHISD (linear) [5] | Our method |
|---|---|---|---|---|---|---|
| 50 frames | 76.92% | 69.23% | 74.36% | **87.18%** | 82.05% | 84.62% |
| 100 frames | 84.62% | 87.18% | **94.87%** | 84.62% | 84.62% | 92.31% |
| Full Length | 94.87% | 94.87% | 97.44% | 89.74% | 92.31% | **100%** |
| Average | 85.47% | 83.76% | 88.89% | 87.18% | 86.33% | **92.31%** |

Table 2. Average identification rates and the standard deviations of different methods on CMU Mobo dataset.

| Methods | Average Performance |
|---|---|
| DCC [25] | $91.53 \pm 1.66\%$ |
| MMD [19] | $89.72 \pm 3.48\%$ |
| MDA [18] | $95.97 \pm 1.90\%$ |
| AHISD [5] | $94.58 \pm 2.57\%$ |
| CHISD [5] | $96.52 \pm 1.18\%$ |
| Our method | $\mathbf{97.08\% \pm 1.03}$ |

Table 3. Average identification rates and the standard deviations on Youtube dataset for five-fold cross validation experiments.

| Methods | Average Performance |
|---|---|
| DCC [25] | $53.90 \pm 4.68\%$ |
| MMD [19] | $54.04 \pm 3.69\%$ |
| MDA [18] | $55.11 \pm 4.55\%$ |
| AHISD [5] | $60.71 \pm 5.24\%$ |
| CHISD [5] | $60.42 \pm 5.95\%$ |
| Our method | $\mathbf{65.03 \pm 5.74\%}$ |

lower than those reported in [5] because the images are resized to $20 \times 20$ instead of $40 \times 40$. The results are obtained by the implementation provided by the authors of [5].

On CMU Mobo dataset, one sequence per subject is randomly selected for training and the remaining are used for testing. We conduct $10$ experiments by repeating the random selection of training/testing data and report the average identification rates and standard deviations of different methods. The results summarized in Table 2 show that the proposed method consistently achieves the best performance (highest classification rate and smallest standard deviation). It is worth mentioning that our method is generic and gives good performance across different types of features e.g. pixel values or LBP features. Table 1 and 2 show that our method consistently achieves good results using pixel values (Honda and Youtube) and LBP features (Mobo). On the other hand, other methods may achieve good results using one feature and degraded performance using another feature. For example, MDA achieves the second best overall performance on the Honda dataset using pixel values and CHISD achieves the second best performance on Mobo dataset using LBP histograms.

On the Youtube Celebrities dataset, we conduct five-fold cross validation experiments. The whole dataset is equally divided into five folds. In each fold, $3$ image sets are randomly selected for training and $6$ are selected for testing. The average identification rates and the associated standard deviations of different methods are summarized in Table 3. Because the videos are captured from real world in low quality and broad appearance variations are covered in this dataset, all methods achieve lower recognition rates compared to the other two datasets. Notice that the results of some methods are relatively lower than those reported in [18] because our experimental setting is more challenging, the automatically cropped faces contain larger tracking errors and the data distribution of training/testing in $5$-fold cross validation is broader than [18]. It is shown that our method again achieves the best performance using the same set of parameters used in previous experiments.

Table 4. Comparison of our proposed method with sparse modeling for single image [7] on Youtube dataset ($47$ subjects).

| | Our method | Wright et al. [7] |
|---|---|---|
| Identification Rate | 65.03% | 63.12% |
| Matching Time per set | 55.64s | 336.33s |

We also compare the performance and computational complexity of our proposed method and the sparse modeling method for single image classification [7] on the Youtube dataset. The technique in [7] can be extended from a single image to multiple images for image set classification. Given a query set, all sample images are sparsely represented as a linear combination of the images of all gallery sets and the image set is assigned to the class with the minimum reconstruction error of all its sample images as in [7]. Table 4 shows the advantages of our proposed sparse modeling for image set classification. Our method not only achieves better performance but is also more efficient. The accuracy of our method comes from the fact that it dynamically finds the nearest points (SANPs), which correspond to images that may not have appeared in the set samples. On the other hand, Wright et al. [7] rely completely upon the sparse representations of the original samples. Our

method is more efficient because it optimizes SANPs based on smaller individual gallery sets (small dictionary) compared to [7] where the query image is approximated from the complete gallery (i.e. a much larger dictionary). Moreover, a straight forward extension of [7] to the image set classification problem requires sparse approximations of all samples in the query set whereas our method requires the sparse approximations of SANPs only.

## 6. Conclusion and Discussion

We proposed a novel sparse formulation for image set classification. An image set is represented as a triplet including the sample images, their mean and their affine hull model. We introduced the Sparse Approximate Nearest Points (SANP) to measure the between-set dissimilarity. Unlike the sparse model of a single image, the sparsity is enforced on sample coefficients rather than the model coefficients of the proposed SANP. The optimization of SANP jointly minimizes the distance and maximizes the sparsity of the nearest points using a scalable accelerated proximal gradient method. We conducted a thorough experimental evaluation on three benchmark datasets for face recognition based on image sets and compared the results to the existing state-of-the-art. Using the same fixed set of parameters, our method consistently achieves the best performances across all experiments as well as features while the performances of other methods fluctuate even with tuned parameters on different datasets/features.

## Acknowledgements

This research was supported by ARC grants DP1096801 and DP0881813. We thank T.-K. Kim for sharing the source code of DCC and R. Wang for sharing the source code of MMD and the cropped faces of the Honda/UCSD dataset. We also thank H. Cevikalp for sharing the source code of AHISD/CHISD and providing the LBP features for the Mobo dataset.

## References

[1] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imaging Sciences*, 2(1), 2009.

[2] A. W. Fitzgibbon and A. Zisserman. Joint Manifold Distance: A New Approach to Appearance based Clustering. In *CVPR*, 2003.

[3] D. Ross, J. Lim, R.-S. Lin and M.-H. Yang. Incremental Learning for Robust Visual Tracking. *IJCV*, 77(1-3), 2008.

[4] G. Shakhnnarvovich, J. W. Fisher and T.Darrell. Face Recognition From Long-Term Observations. In *ECCV*, 2002.

[5] H. Cevikalp and B. Triggs. Face Recognition Based on Image Sets. In *CVPR*, 2010.

[6] H. Hotelling. Relations between Two Sets of Variates. *Biometrika*, 28(3-4), 1936.

[7] J. Wright, A. Yang, A. Ganesh, S. Sastry and Y. Ma. Robust Face Recognition via Sparse Representation. *TPAMI*, 31(2), 2009.

[8] K.-C. Lee, J. Ho, M.-H. Yang and D. Kriegman. Video-Based Face Recognition Using Probabilistic Appearance Manifolds. In *CVPR*, 2003.

[9] K. Fukui and O. Yamaguchi. The Kernel Orthogonal Mutual Subspace Method and Its Application to 3D Object Recognition. In *ACCV*, 2007.

[10] L. Wolf and A. Shashua. Learning over Sets using Kernel Principal Angles. *JMLR*, 4(10), 2003.

[11] M. Kim, S. Kumar, V. Pavlovic and H. Rowley. Face Tracking and Recognition with Visual Constraints in Real-World Videos. In *CVPR*, 2008.

[12] M. Nishiyama, M. Yuasa, T. Shibata, T. Wakasugi, T. Kawahara and O. Yamaguchi. Recognizing Faces of Moving People by Hierarchical Image-Set Matching. In *CVPR*, 2007.

[13] M. Nishiyama, O. Yamaguchi and K. Fukui. Face Recognition with the Multiple Constrained Mutual Subspace Method. In *AVBPA*, 2005.

[14] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla and T. Darrell. Face Recognition with Image Sets Using Manifold Density Divergence. In *CVPR*, 2005.

[15] O. Yamaguchi, K. Fukui and K.-I. Maeda. Face Recognition Using Temporal Image Sequence. In *AFGR*, 1998.

[16] P. Viola and M. J. Jones. Robust Real-Time Face Detection. *IJCV*, 57(2), 2004.

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

[18] R. Wang and X. Chen. Manifold Discriminant Analysis. In *CVPR*, 2009.

[19] R. Wang, S. Shan, X. Chen and W. Gao. Manifold-Manifold Distance with Application to Face Recognition based on Image Set. In *CVPR*, 2008.

[20] S. Zhou and R. Chellappa. Probabilistic Human Recognition from Video. In *ECCV*, 2002.

[21] T. Ahonen, A. Hadid and M. Pietikainen. Face Description with Local Binary Patterns: Application to Face Recognition. *TPAMI*, 28(12), 2006.

[22] T.-K. Kim and R. Cipolla. On-line Learning for Maximizing Orthogonality Between Subspaces and Its Application to Image Set-based Face Recognition. *TIP*, 19(4), 2009.

[23] T.-K. Kim, J. Kittler and R. Cipolla. Incremental Learning of Locally Orthogonal Subspaces for Set-based Object Recognition. In *BMVC*, 2006.

[24] T.-K. Kim, O. Arandjelovic and R. Cipolla. Boosted Manifold Principal Angles for Image Set-Based Recognition. *Pattern Recognition*, 40(9), 2007.

[25] T.-K. Kim, O. Arandjelovic and R. Cipolla. Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations. *TPAMI*, 29(6), 2007.

[26] T.-K. Kim, O. Arandjelovic, R. Cipolla. Learning Over Sets Using Boosted Manifold Principle Angles (BoMPA). In *BMVC*, 2005.

[27] T. Wang and P. Shi. Kernel Grassmannian Distances and Discriminant Analysis for Face Recognition from Image Sets. *PRL*, 30(13), 2009.

[28] W. Fan and D.-Y. Yeung. Locally Linear Models on Face Appearance Manifolds with Application to Dual-Subspace Based Classification. In *CVPR*, 2006.

[29] X. Li, K. Fukui and N. Zheng. Boosting Constrained Mutual Subspace Method for Robust Image-Set Based Object Recognition. In *IJCAI*, 2009.

[30] X. Li, K. Fukui and N. Zheng. Image-Set based Face Recognition Using Boosted Global and Local Principal Angles. In *ACCV*, 2009.

[31] Y. Nesterov. Gradient methods for minimizing composite objective function. *CORE Discussion Paper*, 2007.