Improved Image Set Classification via Joint Sparse Approximated Nearest Subspaces

Shaokang Chen, Conrad Sanderson, Mehrtash T. Harandi, Brian C. Lovell

University of Queensland, School of ITEE, QLD 4072, Australia
NICTA, GPO Box 2434, Brisbane, QLD 4001 Australia
Queensland University of Technology, Brisbane, QLD 4000, Australia

Abstract

Existing multi-model approaches for image set classification extract local models by clustering each image set individually only once, with fixed clusters used for matching with other image sets. However, this may result in the two closest clusters to represent different characteristics of an object, due to different undesirable environmental conditions (such as variations in illumination and pose). To address this problem, we propose to constrain the clustering of each query image set by forcing the clusters to have resemblance to the clusters in the gallery image sets. We first define a Frobenius norm distance between subspaces over Grassmann manifolds based on reconstruction error. We then extract local linear subspaces from a gallery image set via sparse representation. For each local linear subspace, we adaptively construct the corresponding closest subspace from the samples of a probe image set by joint sparse representation. We show that by minimising the sparse representation reconstruction error, we approach the nearest point on a Grassmann manifold. Experiments on Honda, ETH-80 and Cambridge-Gesture datasets show that the proposed method consistently outperforms several other recent techniques, such as Affine Hull based Image Set Distance (AHISD), Sparse Approximated Nearest Points (SANP) and Manifold Discriminant Analysis (MDA).

1. Introduction

Image set classification approaches can be categorised into two general classes: parametric and non-parametric methods. The former utilise parametric distributions [2, 3, 17] to represent image sets. The similarity between the estimated parameters of the distributions can be considered as a distance measure between two sets. However, the estimated parameters might be dissimilar if the training and test data sets of the same subject have weak statistical correlations [14, 28].

Non-parametric methods can be grouped into two classes: single-model and multi-model methods. Single-model methods can be further divided into two groups: single linear subspace methods and affine hull methods. Single linear subspace methods [14, 29] use principal angles to measure the difference between two subspaces. As the similarity of data structures is used for comparing sets, subspace approaches can be robust to noise and relatively small number of samples [29, 28]. However, subspace methods consider the structure of all data samples without selecting optimal subsets for classification. Affine hull approaches [4, 12] use geometric distances to compare sets, such as the closest points between two affine hulls by least squares optimisation. As such, these methods adaptively choose optimal samples to obtain the distance between sets, allowing for a degree of intra-class variations [12]. However, as only distances between certain samples are used, structural information is largely ignored. Furthermore, deterioration in discrimination performance can occur if the nearest points between two hulls are outliers or noisy.

Multi-model approaches generate multiple local linear models by clustering to improve recognition accuracy [9, 27, 28]. In [9], Locally Linear Embedding [21] and k-means clustering are used to extract several representative exemplars. The maximal linear patches technique is used to extract local linear models in [27, 28]. For two sets with m and n local models, the minimum distance between their local models determines the set-to-set distance, which is acquired by $m \times n$ local model comparisons.

A limitation of current multi-model approaches is that each set is clustered individually only once, resulting in fixed clusters of each set being used for classification. These clusters may not be optimal for discrimination, as undesirable environmental conditions (such as variations in illumination and pose) may result in the two closest clusters representing two different characteristics of an object.

Consider that each cluster can be interpreted as representing a particular physical property of an object. For example, let us assume we have two face image sets of the same person, representing two different conditions. The clusters in the first set represent various poses, while the clusters in the second set represent varying illumination (where the illumination is different to the illumination present in the first set). As the two sets of clusters capture two different variations, matching two image sets based on cluster matching may result in a non-frontal face (eg. rotated or tilted) being compared against a frontal face.

Contributions. To address the above problem, we propose to constrain the clustering of each query image set by forcing the clusters to have resemblance to the clusters in gallery image sets, while simultaneously using structural information (similar to single linear subspace methods) and selecting a subset of samples (similar to affine hull methods).

Consider two sets to be compared. The proposed approach first uses sparse approximation to extract local linear subspaces from the first set. Each local linear subspace is then represented as a reference point on a Grassmann manifold. For each reference point, we approximate its closest point on the manifold from a group of points of the second set. Instead of searching through all the points, we apply joint sparse approximation to solve the search problem. We prove that by minimising the joint sparse representation error, we are approaching the nearest point to the reference point on the Grassmann manifold. The average distance of the closest points from the second set to the corresponding reference points of the first set is taken to indicate the distance between the two sets. We term the proposed approach as Sparse Approximated Nearest Subspaces (SANS). Fig. 1 shows a conceptual illustration of the proposed approach.

Comparisons on three benchmark datasets for face, hand gesture and object classification show that the proposed method consistently outperforms several recent techniques. To our knowledge, this is the first paper to show the link between joint sparse approximation and Grassmann manifolds, and the proposed method is the first that adaptively constructs the closest subspace to a reference subspace from the samples of a set.

We continue the paper as follows. In Section 2, we briefly overview sparse representation and Grassmann manifolds. We then define a Frobenius norm distance between subspaces over Grassmann manifold in Section 3. The proposed approach is discussed in detail in Section 4, followed by empirical evaluations and comparisons with other methods in Section 5. The main findings and possible future research directions are summarised in Section 6.

2. Mathematical Preliminaries

This section overviews sparse representation as well as Grassmann manifolds, serving as a ground for further developments. More rigorous treatment of sparse representation can be found in [5, 6], while manifolds and related topics are covered in [1, 7, 11].

2.1. Sparse Representation

Sparse representation is based on the observation that natural signals can be concisely represented if the signal basis is properly selected. Consider a single measurement vector (SMV) $\boldsymbol{x} \in \mathbb{R}^n$, which requires n numbers for representation in the spatial domain. If the basis of the space is carefully selected, $\boldsymbol{x}$ can be represented with d atoms (with $d < n$), where each atom is an entry in a dictionary. Assume a dictionary $\boldsymbol{D}$ can represent all possible measurements of the signal. The sparse representation of $\boldsymbol{x}$ can be achieved by solving the following ℓ_0-norm optimisation:

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_0 , \quad \text{s.t.} \quad \boldsymbol{x} = \boldsymbol{D}\boldsymbol{w} \tag{1}$$

where $\|\boldsymbol{w}\|_0$ is the ℓ_0-norm that counts the number of non-zero elements in $\boldsymbol{w}$. Greedy pursuit methods iteratively approximates the sparse solution by finding the local optimal at each iteration to solve the equivalent feasible problem [25]:

$$\min_{\boldsymbol{w}} ||\boldsymbol{x} - \boldsymbol{D}\boldsymbol{w}||_2, \quad \text{s.t.} \quad \|\boldsymbol{w}\|_0 \leq \alpha. \tag{2}$$

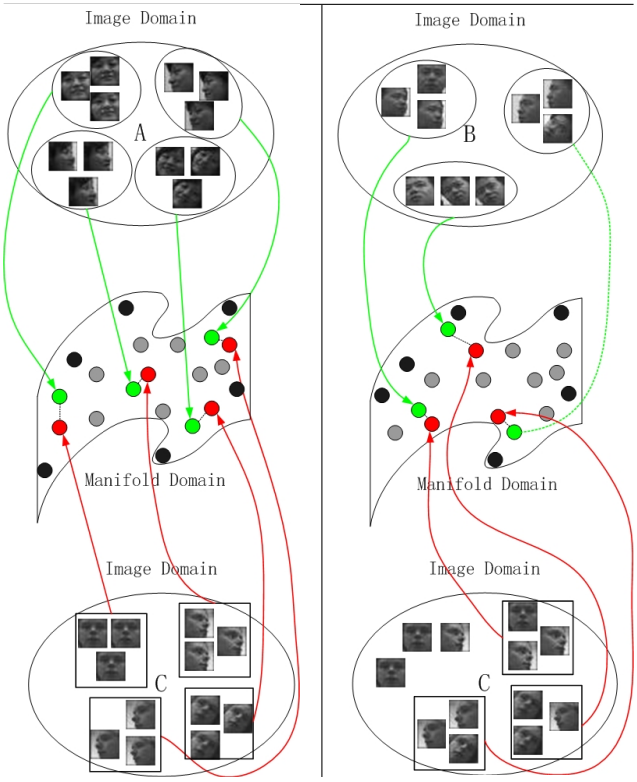

Figure 1. Conceptual illustration of the proposed approach. Image sets A and B are separately clustered. The green dots indicate the corresponding point on a Grassmann manifold using images in the cluster. The black points indicate other Grassmann points using subsets of images from image set A or B. Query set C is separately clustered according to the clusters of sets A and B. Set C is divided into 4 clusters during comparison with set A. The red dots show that images in set C are adaptively clustered such that the nearest Grassmann point can be constructed corresponding to the reference green points on the manifold. In this way, the corresponding nearest clusters in set A and C capture similar variations. The gray points indicate other Grassmann points using subsets of images from set C. When comparing with set B, set C is adaptively clustered into 3 clusters.

Sparse representation has been extended from SMV to multiple measurement vectors (MMV) [23, 24, 26], also known as joint sparse representation (JSR). In MMV, multiple vectors are simultaneously reconstructed using the same basis. Given a matrix $\boldsymbol{X}$ composed from a set of column vectors, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m]$, and a dictionary $\boldsymbol{D}$, JSR solves the following optimisation problem:

$$\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_{p,r}, \quad \text{s.t.} \quad \boldsymbol{X} = \boldsymbol{D}\boldsymbol{W} \tag{3}$$

where $\|\boldsymbol{W}\|_{p,r}$ is the matrix norm defined as [23]:

$$\|\boldsymbol{W}\|_{p,r} = \left\{ \sum_{i=1}^{n} \left(\sum_{j=1}^{m} |w_{i,j}|^p \right)^{\frac{r}{p}} \right\}^{\frac{1}{r}} = \left\{ \sum_{i=1}^{n} \left\| \boldsymbol{w}^{[i]} \right\|_p^r \right\}^{\frac{1}{r}} \tag{4}$$

with $\boldsymbol{w}^{[i]}$ representing the i-th row of $\boldsymbol{W}$. A typical choice of p is 2 or ∞ [26]. Following the ℓ_0-norm optimisation in Eqn. (2), solution of Eqn. (3) can be approximated by [5]:

$$\min_{\boldsymbol{W}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{W}\|_F^2, \quad \text{s.t.} \quad \|\boldsymbol{W}\|_{2,0} \leq \alpha, \tag{5}$$

where $\|\boldsymbol{W}\|_{2,0}$ counts the number of rows in $\boldsymbol{W}$ that contain non-zero elements.

2.2. Grassmann Manifolds

Manifold analysis has been extensively studied with success in various disciplines, such as activity recognition and pedestrian detection [10, 22]. A manifold can be considered as a low dimensional smooth surface embedded in a higher dimensional space. At each point of the manifold, it is locally similar to Euclidean space. In this paper we focus on a particular class of manifolds, known as Grassmann manifolds.

A Grassmann manifold $G_{D,m}$ is a set of m-dimensional linear subspaces of $\mathbb{R}^D$. A point in $G_{D,m}$ can be represented by an orthonormal matrix with a size of $D \times m$. The matrix representation of a Grassmann point is not unique, ie. two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ represent the same point if the subspaces spanned by the column vectors of the two matrices are the same. The distance between two Grassmann points is the length of the shortest geodesic connecting two points, which can be obtained via [7]:

$$d_G(\boldsymbol{A}, \boldsymbol{B}) = \|\Theta\|_2 \tag{6}$$

where $\Theta = [\theta_1, \theta_2, \cdots, \theta_p]$ is the principal angle vector, ie.

$$\cos(\theta_i) = \max_{\boldsymbol{a}_i \in \boldsymbol{A},\ \boldsymbol{b}_j \in \boldsymbol{B}} \boldsymbol{a}_i' \boldsymbol{b}_j \tag{7}$$

subject to $\boldsymbol{a}_i'\boldsymbol{a}_i = \boldsymbol{b}_i'\boldsymbol{b}_i = 1$, $\boldsymbol{a}_i'\boldsymbol{a}_j = \boldsymbol{b}_i'\boldsymbol{b}_j = 0$, $i \neq j$. The principal angles have the property of $\theta_i \in [0, \pi/2]$ and can be computed through singular value decomposition of $\boldsymbol{A}'\boldsymbol{B}$ [16].

Grassmann manifolds provide a straightforward way to solve image set matching problems. A set with m images of D pixels can be transformed directly to a point on $G_{D,m}$. Thus the image set classification problem can be transferred to a point classification problem on Grassmann manifolds.

3. Residual Distance on Grassmann Manifold

Following the form of JSR, we define a Frobenius norm distance, named residual distance, between two subspaces over a Grassmann manifold. For two subspaces S_a and S_b, the distance between subspaces is defined as the summation of distance from the unit vectors of orthonormal basis of the subspace S_a to the subspace S_b. That is

$$D(S_a, S_b) = ||\boldsymbol{U_a} - \boldsymbol{U_b}\boldsymbol{U_b}'\boldsymbol{U_a}||_F^2, \tag{8}$$

where $\boldsymbol{U_a}$ and $\boldsymbol{U_b}$ are the orthonormal basis of S_a and S_b individually. The distance $D(S_a, S_b)$ is also the reconstruction error of $\boldsymbol{U_a}$ represented by the basis $\boldsymbol{U_b}$.

This residual distance is the $l2$ norm of the sine of principal angles given in Eqn (6) and is proved to be a form of projection distance over Grassmann manifolds [10].

4. Sparse Approximated Nearest Subspaces

We now propose the approach to find the nearest subspace over Grassmann manifolds by minimising the residual distance. The proposed method consists of three main components, which are explained in detail in the following sub-sections.

1. **Local linear subspace extraction.** Images in a gallery image set are grouped based on sparse representation, in order to extract multiple local linear subspaces.
2. **Nearest subspace approximation.** For each local linear subspace from a gallery image set, the approximated nearest subspace is adaptively constructed from the samples of the query image set. Joint sparse representation is applied to approximate the nearest subspace.
3. **Distance calculation.** The average distance of all the closest subspace pairs is considered as the distance between two sets.

4.1. Local Linear Subspace Extraction

Given an image set I_a represented by matrix $\boldsymbol{X}_a = \left[\boldsymbol{x}_1^a, \boldsymbol{x}_2^a, \ldots, \boldsymbol{x}_{N_a}^a\right]$, where each column vector represents an image of I_a, we can create a total of $N_m^a = \frac{N_a!}{m!(N_a-m)!}$ subspaces of rank m from the available N_a sample images. A collection of all these subspaces S_m^a is called the m-order subspace set of I_a. We note that not all of the subspaces can precisely represent the variations of the object and hence only some of the subspaces should be used for classification.

Single measurement vector (SMV) sparse representation is applied to create and select local linear subspaces that can accurately represent real samples from the image set. This is in contrast to affine hull based methods [4, 12], where the nearest points are synthetic samples generated through linear combination of real samples. For each sample image $\boldsymbol{x}_i^a$ from set I_a we use the remaining images $\boldsymbol{D}_k^a = \left[\boldsymbol{x}_1^a, \boldsymbol{x}_2^a, \ldots, \boldsymbol{x}_{i-1}^a, \boldsymbol{x}_{i+1}^a, \boldsymbol{x}_{N_a}^a\right]$ to reconstruct sample $\boldsymbol{x}_i^a$ sparsely as per Eqn. (2). We specify the number of atoms

m to use by choosing the atoms corresponding to the largest m absolutes of coefficients w acquired via Eqn. (2). Let us construct matrix $\boldsymbol{M}_k^a = \left[\boldsymbol{x}_{k_1}^a, \boldsymbol{x}_{k_2}^a, \ldots, \boldsymbol{x}_{k_m}^a\right]$ containing m column vectors of the selected atoms. $\boldsymbol{M}_k^a$ can be used to represent a subspace s_k^a. The distance between the sample point $\boldsymbol{x}_k^a$ and the subspace s_k^a can be calculated by the reconstruction error $r_k^a = \|\boldsymbol{M}_k^a \boldsymbol{w}_k^a - \boldsymbol{x}_k^a\|_2$, where $\boldsymbol{w}_k^a$ are the coefficients of the selected atoms.

From a manifold point of view, each subspace of order m can be represented as a point on Grassmann manifold $G_{D,m}$. For each image $x_i^a \in I_a$, the subspace s_k^a constructed by SMV sparse representation is represented as a point over the manifold. By setting a threshold ϵ on the reconstruction error, the SMV sparse representation can select representative subspaces s_k^a that can linearly represent real samples x_k^a with an error smaller than threshold ϵ as following:

$$\widetilde{S}_m^a = \{s_k^a\}, \forall r_k^a < \epsilon, k \in [1, N_a], \tag{9}$$

where $\widetilde{S}_m^a$ is a set of selected points on the manifold. This filtering step significantly reduces the number of points from N_m^a to less than N_a. This type of subspace extraction is motivated by [8], where SMV sparse representation is used to cluster linear subspaces.

4.2. Nearest Subspace Approximation

After extracting local linear subspaces, traditional multi-model approaches use fixed subspaces (clusters) of each set for classification. In contrast, we propose to adaptively cluster the query image set via considering the clusters from a gallery image set. To match image sets I_a and I_b, we first extract the local linear subspace set $\widetilde{S}_m^a$ for image set I_a as per Eqn. (9). Then for each extracted local linear subspace, we find its corresponding nearest subspace from the m-order subspace set S_m^b of I_b. From manifold point of view, for image set I_b with N_b images, there are $N_m^b = \frac{N_b!}{m!(N_b-m)!}$ points on the same manifold. For each point s_k^a, we need to find its closest point from N_m^b points of set I_b on the manifold.

Instead of searching through all the points on the manifold, we apply joint sparse approximation [24, 26] to solve this challenging NP-hard search problem. We first generate the orthonormal basis U_a of subspace s_k^a. We then treat all the samples in I_b as elements in dictionary and apply joint sparse representation to find the optimal solution via Eqn. (5) by specifying the number of active atoms m. Given orthonormal basis $\boldsymbol{U}_a$ from s_k^a, we find m samples from matrix $\boldsymbol{X}_b$, representing image set I_b, that give minimal sparse representation error[1]:

$$\min_W \|\boldsymbol{U}_a - \boldsymbol{X}_b \boldsymbol{W}\|_F^2, \text{s.t.} ||W||_{2,0} \leq m. \tag{10}$$

[1]Note that a rotated basis $\boldsymbol{U_a R_a}$ may have slightly different solution to $\boldsymbol{U_a}$ due to the limitation of the approximated solution for joint sparse representation.

Assume matrix $\tilde{\boldsymbol{X}}_{\boldsymbol{b}}$ is formed by the m samples selected by equation 10 and $\tilde{\boldsymbol{W}}$ is the corresponding non-zero elements from $\boldsymbol{W}$. Thus, the reconstruction error is

$$E_k = \left\|\boldsymbol{U_a} - \tilde{\boldsymbol{X}}_{\boldsymbol{b}} \tilde{\boldsymbol{W}}\right\|_F^2. \tag{11}$$

The samples $\tilde{\boldsymbol{X}}_{\boldsymbol{b}}$ can be also be used to construct a subspace s_k^b with orthonormal basis $\boldsymbol{U_b}$. The reconstruction error can be rephrased as equation 8. Thus the reconstruction error can be used as a measure of distance $D(s_k^a, s_k^b) = E_k$ between two subspaces on Grassmann manifold. By minimising the error, the nearest subspace over Grassmann manifolds is approached.

4.3. Distance Calculation

We have shown above how to approximate the nearest subspace s_k^b from S_m^b, given a specific subspace s_k^a from the m-order subspace set $\widetilde{S}_m^a$. As we generate N_c local linear subspaces from I_a and find their corresponding nearest subspaces from I_b, the distance between two image sets I_a and I_b is defined as the average distance of the nearest subspace pairs:

$$\widehat{D}\left(I_a, I_b\right) = \frac{1}{N_c} \sum\nolimits_{k=1}^{N_c} D\left(s_k^a, s_k^b\right), \quad k \in [1, N_c] \tag{12}$$

4.4. Complexity Analysis

The complexity of the proposed SANS method is dependant on the complexity of joint sparse representation (JSR). Given two image sets with n_c and n_d samples, the complexity of JSR is $O(n_c n_d m)$, where m is the number of active atoms used. Thus, the complexity of SANS is $O(N_c m n_d m)$, where N_c is the number of local linear subspaces generated. By controlling m and the reconstruction threshold (to limit N_c), the time complexity can be constrained.

5. Experiments

The proposed approach was first evaluated on synthetic data to investigate the accuracy of nearest subspace approximation, followed by a performance comparison against previous state-of-the-art methods on three image set recognition tasks: face, gesture and object recognition.

5.1. Synthetic Data

We randomly generated m sample points in $\mathbb{R}^n$ (n = 100) to construct a reference subspace S_{ref} with rank m. $N >> m$ sample points are randomly generated in $\mathbb{R}^n$ as a dictionary. The proposed nearest subspace approximation (NSA) approach is used to find m samples from the dictionary to construct the approximated nearest subspace S_{app} and is compared with the actual nearest subspace S_{act} found by a brute force method. The relative difference ratio $r = \frac{|D(S_{ref}, S_{app}) - D(S_{ref}, S_{act})|}{D(S_{ref}, S_{act})}$ and the percentage of S_{app} in the top k nearest subspaces of S_{ref} are considered as the measurements of performance. The results are

Table 1. Accuracy of the proposed nearest subspace approximation (NSA) on synthetic data. m is the number of samples used to construct the reference subspace S_{ref}. N is the number of samples in dictionary. The total number of subspaces for each search is C_N^m. 'mean rank' is the average ranking of the approximated subspace S_{app} in all subspaces. The percentage that S_{app} is in the top k nearest subspaces of S_{ref} is shown for $k = 1, k = 5, k = 0.01 \times C_N^m$, and $k = 0.05 \times C_N^m$. Ratio r measures the relative difference of distance as $r = |D(S_{ref}, S_{app}) - D(S_{ref}, S_{act})|/D(S_{ref}, S_{act})$, where S_{act} is the actual nearest subspace of S_{ref} found by a brute force method.

dictionary size $N = 20$									
num. of	total num. of	mean	S_{app} in the top k nearest subspaces				ratio	time (ms)	
samples m	subspaces C_N^m	rank	k=1	k=10	k=1%	k=5%	r	NSA	brute force
2	190	6.33	32%	80%	45%	80%	0.014	0.9	33
3	1140	16.7	22%	59%	63%	91%	0.012	1.1	236
4	4845	44.7	18%	52%	83%	96%	0.008	1.3	1099
dictionary size $N = 100$									
num. of	total num. of	mean	S_{app} in the top k nearest subspaces				ratio	time (ms)	
samples m	subspaces C_N^m	rank	k=1	k=10	k=1%	k=5%	r	NSA	brute force
2	4950	21.9	23%	60%	83%	100%	0.022	4	894
3	161700	375.3	7%	22%	96%	100%	0.018	4	32450

achieved based on the average of 1000 tests for dictionary size $N = 20$ and $N = 100$ separately.

Table 1 shows how close the approximated nearest subspace is to the actual nearest subspace. The average relative difference ratio r is less than 1.5%. The ratio is insensitive to the number of samples m of reference subspaces. However, it is affected by the dictionary size. This is expected as increasing the dictionary size, the total number of subspaces is exponentially increased, while the ratio is only increased slightly. Evaluating the performance from the point of view of the ranking for the approximated nearest subspace, most of the approximated subspaces are in the top 1% closest subspaces and almost all of the approximated subspaces are in the top 5% closest subspaces. The proposed approach can find maximally 32% actual nearest subspaces when dictionary size is small. In the worst case, at least 7% actual nearest subspaces are found when the total number of subspaces is huge ($> 160,000$). The calculation time of the proposed method is nearly constant and takes only several milliseconds, disregarding the number of samples and the dictionary size. In contrast, the brute force method to find the actual nearest subspace is hundreds or even thousands of times slower.

5.2. Image Set Recognition Tasks

We used the Honda/UCSD dataset [17] for the face recognition task, the ETH-80 dataset [18] for the object recognition task and Cambridge-Gesture dataset [15] for hand gesture recognition task. We will first briefly overview the datasets used in the experiments (Section 5.2.1), followed by a description and discussion of the experiments (Section 5.2.2).

5.2.1 Datasets

Honda/UCSD consists of 59 videos of 20 subjects. There are pose, illumination and expression variations across the sequences for each subject. As in [28], face images from each frame of Honda/UCSD dataset were cropped and resized to 20×20. We followed [12, 27] to conduct 10-fold cross validations by randomly selecting one sequence for each subject for training and using the rest for testing.

ETH-80 contains 8 object categories. Each category includes 10 object subcategories (eg. various dogs), with each subcategory having 41 orientations. We resized the images to 32×32 and treated each subcategory as an image set. For each category, we selected each subcategory in turn for training and the remaining 9 for testing. In total, 80 image sets were used for training and 720 for testing.

The Cambridge-Gesture dataset includes 900 video sequences for nine gestures. For each gesture, the 100 videos are further divided into five illumination sets. Following the protocol of [19], the first four sets are used for test set and the fifth set is the training set. All images are resized to 20×20 and we select the middle 32 frames from each video sequence as in [19].

On the Honda/UCSD dataset, we used three configurations of training and testing images: randomly chosen 50, randomly chosen 100, and all images. If the number of images in a set is smaller than the number specified, then all the images are selected. Using a subset of images partly simulates real-world situations where a face detector and/or tracker may fail on some frames. On ETH-80 and Cambridge-Gesture datasets, we used all raw images for classification, while on Honda/UCSD we used two types of images: raw and normalised via histogram equalisation. Histogram equalisation provides some compensation to illumination variations, and hence it can mask the limitations of the matching algorithms. As such, the raw image type provides a more challenging comparison.

5.2.2 Comparative Evaluation and Discussion

The proposed method was compared against five recent algorithms: Affine Hull based Image Set Distance (AHISD) [4], Convex Hull based Image Set Distance (CHISD) [4],

Table 2. Performance of the proposed SANS method with varying parameters m and ϵ on Honda/UCSD dataset using 100 raw images per set.

$\epsilon = 0.01$						
m	10	15	20	25	30	35
accuracy	93.6	94.1	93.3	92.8	93.8	92.3
$m = 15$						
ϵ	0.005	0.01	0.02	0.03	0.04	0.05
accuracy	92.8	94.1	92.3	93.6	93.3	93.3

Table 3. Comparison of the proposed method with the component techniques involved, such as Joint Sparse Representation (JSR), Grassmann Manifolds (GM) and Local Linear Subspace (LLS) extraction. The results were obtained on the Honda/UCSD dataset. 'h.e.' indicates that the images were pre-processed with histogram equalisation.

num. of images	image type	JSR	JSR+LLS	GM Eqn. (6)	GM Eqn. (8)	**proposed SANS**
50	raw	83.8	87.7	88.7	90.0	**92.3**
	h.e.	87.4	89.5	94.4	94.1	**95.6**
100	raw	88.2	93.8	87.9	90.0	**93.8**
	h.e.	89.2	90.8	94.9	94.8	**96.7**
all	raw	87.7	93.6	88.7	92.1	**94.1**
	h.e.	85.4	93.8	90.8	95.1	**96.4**

Sparse Approximated Nearest Points (SANP) [12], Mutual Subspace Method (MSM) [29] and Manifold Discriminant Analysis (MDA) [27].

AHISD, CHISD and SANP are nearest point based methods, which find the closest points between two hulls. MSM and MDA are subspace based methods which model image sets as linear subspaces. Except for SANP, we obtained the implementations of all methods from the original authors. We also compare with two component techniques involved in the proposed SANS method: Joint Sparse Representation (JSR) technique (Eqn. (5)) and Grassmann Manifold (GM) technique on two distances (Eqn. (6) and Eqn. (8)).

The proposed SANS model has only two parameters: the number of active atoms m and the sparse representation threshold ϵ. Preliminary experiments suggested that $m \in [10, 30]$ and $\epsilon \in [0.01, 0.05]$ resulted in satisfactory performance. The performance of SANS is not sensitive to both parameters in the range specified above. Table 2 shows the results of varying m and ϵ on Honda/UCSD dataset, using 100 raw images per set. The performance of SANS is very stable and is consistently better than other methods shown in Table 4. To avoid the effect of duplication samples on local linear subspace extraction due to the limitation of sparse representation, we remove the duplication of samples in each image set individually.

Table 3 shows the comparison of the proposed SANS method with component techniques, including Joint Sparse Representation (JSR), Grassmann Manifolds (GM) and local linear subspace (LLS) extraction. The proposed SANS method can be considered as GM+LLS. SANS always performs the best compared to each individual technique. By applying local linear subspace extraction, the performance of both JSR and GM is improved. For all the methods, performance on histogram equalised images is slightly better than raw images.

Table 4. Performance comparison with other methods on the Honda/UCDS dataset. 'h.e.' indicates that the images were pre-processed with histogram equalisation.

num. of images	image type	AHISD [4]	CHISD [4]	SANP [2] [12]	MSM [29]	MDA [27]	**proposed SANS**
50	raw	68.4	69.7	71.0	84.9	71.0	**92.3**
	h.e.	94.6	92.8	93.1	93.8	88.7	**95.6**
100	raw	65.9	66.9	68.7	84.4	72.1	**93.8**
	h.e.	92.1	93.1	94.4	92.1	87.2	**96.7**
all	raw	64.1	61.5	71.1	84.4	74.4	**94.1**
	h.e.	90.7	91.3	94.9	90.8	96.2	**96.4**

The comparison with other state-of-the-art methods is shown in Table 4. The proposed SANS method obtains the highest accuracy in all cases, with considerable improvements over other methods on raw images. As the Honda/UCSD dataset contains considerable illumination variations, histogram equalisation is required by AHISD, CHISD, SANP and MDA to obtain good performance. For these four methods, there is about 20 percentage points difference between the performance on raw and normalised images. In contrast, the proposed SANS method is considerably more robust, obtaining high performance for both raw and normalised images.

AHISD, CHISD and SANP are all based on the nearest point distance between subspaces, which is inevitably sensitive to the illumination variations. If two image sets are taken in different illumination conditions, the distance between points on two subspaces will be rather large, leading to a deterioration in classification performance. While MDA clusters images to construct local linear models and learns a more discriminant embedding space, the distances between local models/subspaces are based on the Euclidean distance between the center points of models. Thus the distances is also sensitive to illumination variations.

In contrast, MSM, JSR, GM and the proposed SANS exploit structural similarities between subspaces (eg. principal angles), which are more robust to noise (such as illumination variations). It has been previously shown that for holistic face representations, illumination variations lie in a low

[2]The performance of SANP on histogram equalised images is slightly different from the results reported in [12]. The difference might be due to a different face detector being used and/or the random selection of images. Minor performance variations of this nature on the Honda/UCSD dataset have also been observed for MDA in [12, 27].

Table 5. Results on the Cambridge-Gesture dataset [15].

	PM [19]	TCCA [13]	DCCA [14]	**proposed SANS**
Set 1	89	81	63	**90**
Set 2	86	81	61	**89**
Set 3	89	78	65	**91**
Set 4	87	86	69	**89**
Average	88	82	65	**90**

dimensional linear subspace [20]. Sparse representation approaches allow the use of several atoms to linearly represent any sample lying in the same subspace. In other words, if there are several images of a person's face taken under varying illumination conditions, the subspace constructed from these images can be used to represent many possible illumination conditions. The Grassmann manifold approach treats all samples lying in the same subspace as one point on a Grassmann manifold, suggesting that illumination variations do not affect the point. The robustness of SANS also comes from being able to exploit the variations present in the training data by local linear subspace (LLS) extraction and the adaptively constructed nearest subspaces. Multiple local linear subspaces can be extracted from a gallery image set that represent variations of a subject. For a given local linear subspace, SANS finds the closest subspace from the subspace set of the query image set, which represents a similar variation. As an example, Fig. 2 shows the sample images of an extracted local linear subspace as well as the sample images of the constructed nearest subspaces.

Fig. 3 illustrates the results obtained on the ETH-80 dataset. In this test, all methods perform worse than on the Honda/UCSD dataset. ETH-80 is more challenging as it has much less images per set, significant appearance differences across subjects of the same class, and larger view angle variations within each image set. Nevertheless, the proposed SANS method dramatically outperforms other methods by more than 20 percentage points. We note that the performance of MDA on ETH-80 is lower than that reported in [27], as our setup is more challenging. Compared to [27], where 5 sets are used for training, we use only one set. The average time cost to compare two image sets is shown in Table 6.

Figure 2. **(a)** Sample images of an extracted local linear subspace from a gallery image set. **(b)** Sample images of the constructed nearest subspace from a query image set of the same class. **(c)** Sample images of the constructed nearest subspace from a query image set of a different class.

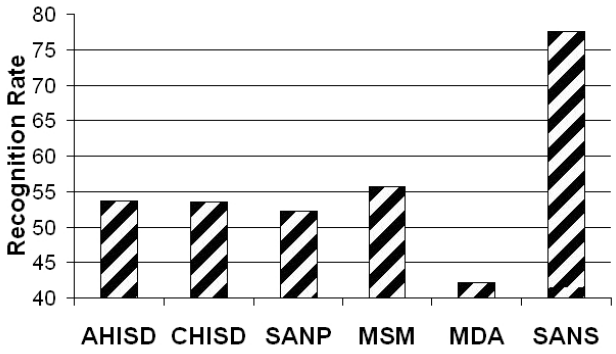

Figure 3. Results on the ETH-80 dataset [18].

Table 6. Comparison of average time cost to compare two image sets with 100 images per set.

Methods	AHISD	CHISD	SANP	MSM	MDA	SANS
Time (ms)	15.3	936.1	65.9	22.3	11.2	35.3

Table 5 shows the results of the proposed method compared with three recent approaches for action classification on the Cambridge-Gesture dataset. Nearest point based methods, such as AHISD, CHISD and SANP, and the multi-model method MDA perform poorly in this dataset with less than 30% accuracy on average, due to the significant illumination variation. The proposed SANS method still performs the best compared to state-of-the-art action classification methods, including Product Manifolds (PM) [19], Tensor Canonical Correlation Analysis (TCCA) [13] and Discriminative Canonical Correlation Analysis (DCCA) [14] methods.

6. Main Findings and Future Directions

We have proposed a novel approach to approximate nearest subspaces over Grassmann manifolds. To this end, we first defined a residual distance over Grassmann manifolds. Single measurement vector sparse representation is then employed to create local linear subspaces from a gallery image set, followed by joint sparse representation to approximate the corresponding nearest subspaces from the probe image set. We have shown that by minimising the joint sparse reconstruction error, the nearest subspace on a Grassmann manifold is approached. The average distance of nearest subspace pairs is defined a new distance between two image sets.

In contrast to single linear subspace methods, the pro-

posed Sparse Approximated Nearest Subspaces (SANS) method extracts multiple local linear subspaces using a subset of samples. Unlike affine hull based approaches, SANS compares structural similarity between local linear subspaces. Distinct to multi-model based methods, SANS utilises the subspaces (clusters) of one image set to adaptively cluster the samples of another image set by constructing the corresponding closest subspaces without complete pairwise local subspace comparisons.

Comparative evaluations on synthetic data show that the proposed method can approximate the nearest subspaces with small errors. Further experiments on three recognition tasks show that the proposed approach consistently outperforms several recent methods (AHISD [4], CHISD [4], SANP [12] and MDA [27]), especially in cases of large image variations and limited number of samples. The experiments also indicate that subspace structural similarity based methods generally perform better than nearest point based methods for image sets with variations in illumination.

Future avenues of research include random rotation of orthogonal basis for more robust nearest subspace approximation and learning more discriminative embedding spaces for manifolds [10, 27]. The proposed nearest subspace approximation can also be extended to use other multi-model approaches for local model extraction, such as Manifold to Manifold Distance (MMD) [28], Manifold Discriminant Analysis (MDA) [27] or Local Linear Embedding (LLE) with $k-$means clustering [9].

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, USA, 2008.

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

[3] F. Cardinaux, C. Sanderson, and S. Bengio. User authentication via adapted statistical models of face images. *IEEE Transactions on Signal Processing*, 54(1):361–373, 2006.

[4] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 2567–2573, 2010.

[5] J. Chen and X. Huo. Theoretical results on sparse representations of multiple-measurement vectors. *IEEE Transactions on Signal Processing*, 54(12):4634–4643, 2006.

[6] D. L. Donoho. Compressed sensing. *IEEE Trans. Information Theory*, 52(4):1289–1306, 2006.

[7] A. Edelman, T. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[8] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 2790–2797, 2009.

[9] A. Hadid and M. Pietikäinen. Manifold learning for video-to-video face recognition. In *Biometric ID Management and Multimodal Communication, Lecture Notes in Computer Science*, volume 5707, pages 9–16, 2009.

[10] J. Hamm and D. D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Int. Conf. Machine Learning (ICML)*, pages 376–383, 2008.

[11] M. Harandi, C. Sanderson, R. Hartley, and B. Lovell. Sparse coding and dictionary learning for symmetric positive definite matrices: a kernel approach. In *European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science*, volume 7573, pages 216–229, 2012.

[12] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 121–128, 2011.

[13] T.-K. Kim and R. Cipolla. Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 31(8):1415–1428, 2009.

[14] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 29(6):1005–1018, 2007.

[15] T.-K. Kim, S.-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007.

[16] A. V. Knyazev and M. E. Argentati. Principal angles between subspaces in an A-based scalar product: Algorithms and perturbation estimates. *SIAM J. Scientific Computing*, 23(6):2008–2040, 2002.

[17] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 313–320, 2003.

[18] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 409–415 (vol. 2), 2003.

[19] Y. M. Lui and J. R. Beveridge. Action classification on product manifolds. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010.

[20] R. Ramamoorthi. Analytic PCA construction for theoretical analysis of lighting variability in images of a lambertian object. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 24(10):1322–1333, 2002.

[21] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[22] A. Sanin, C. Sanderson, M. Harandi, and B. Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 103–110, 2013.

[23] L. Sun, J. Liu, J. Chen, and J. Ye. Efficient recovery of jointly sparse vectors. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1812–1820. 2009.

[24] J. A. Tropp. Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Processing*, 86(3):589–602, 2006.

[25] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

[26] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing*, 86(3):572–588, 2006.

[27] R. Wang and X. Chen. Manifold discriminant analysis. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 429–436, 2009.

[28] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[29] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *IEEE Int. Conf. Automatic Face and Gesture Recognition (AFGR)*, pages 318–323, 1998.