Semi-supervised Spectral Clustering for Image Set Classification

Arif Mahmood, Ajmal Mian, Robyn Owens
School of Computer Science and Software Engineering, The University of Western Australia
{arif.mahmood, ajmal.mian, robyn.owens}@ uwa.edu.au

Abstract

We present an image set classification algorithm based on unsupervised clustering of labeled training and unlabeled test data where labels are only used in the stopping criterion. The probability distribution of each class over the set of clusters is used to define a true set based similarity measure. To this end, we propose an iterative sparse spectral clustering algorithm. In each iteration, a proximity matrix is efficiently recomputed to better represent the local subspace structure. Initial clusters capture the global data structure and finer clusters at the later stages capture the subtle class differences not visible at the global scale. Image sets are compactly represented with multiple Grassmannian manifolds which are subsequently embedded in Euclidean space with the proposed spectral clustering algorithm. We also propose an efficient eigenvector solver which not only reduces the computational cost of spectral clustering by many folds but also improves the clustering quality and final classification results. Experiments on five standard datasets and comparison with seven existing techniques show the efficacy of our algorithm.

1. Introduction

Image set based object classification has recently received significant research interest [1, 2, 5, 10, 12, 17, 20, 28, 29, 30] due to its higher potential for accuracy and robustness compared to single image based approaches. An image set contains more appearance details such as multiple views, illumination variations and time lapsed changes. These variations complement each other resulting in a better object representation. However, set based classification also introduces many new challenges. For example, in the presence of large intra set variations, efficient representation turns out to be a difficult problem. Considering face recognition, it is well known that the images of different identities in the same pose are more similar compared to the images of the same identity in different poses (Fig. 1). Other important challenges include efficient utilization of all available information in a set to exploit intra-set similarities and inter-set dissimilarities.

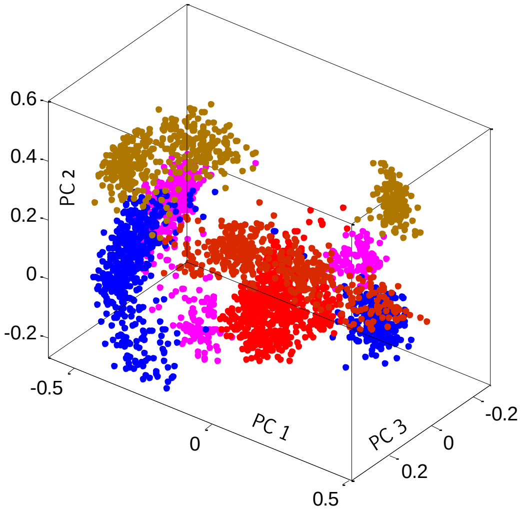

Figure 1. Projection on the top 3 principal directions of 5 subjects (in different colors) from the CMU Mobo face dataset. The natural data clusters do not follow the class labels and the underlying face subspaces are neither independent nor disjoint.

Many existing image set classification techniques are variants of the nearest neighbor (NN) algorithm where the NN distance is measured under some constraint such as representing sets with affine or convex hulls [8], regularized affine hull [3], or using the sparsity constraint to find the nearest points between image sets [30]. Since NN techniques utilize only a small part of the available data, they are more vulnerable to outliers.

At the other end of the spectrum are algorithms that represent the holistic set structure, generally as a linear subspace, and compute similarity as canonical correlations or principle angles [26]. However, the global structure may be a non-linear complex manifold and representing it with a single subspace will lead to incorrect classification [2]. Discriminant analysis has been used to force the class boundaries by finding a space where each class is more compact while different classes are apart. Due to multi-modal nature of the sets, such an optimization may not scale the inter class distances appropriately (Fig. 1). In the middle of the spectrum are algorithms [2, 24, 28] that divide an image set into multiple local clusters (local subspaces or manifolds) and measure cluster-to-cluster distance [24, 28]. Chen et

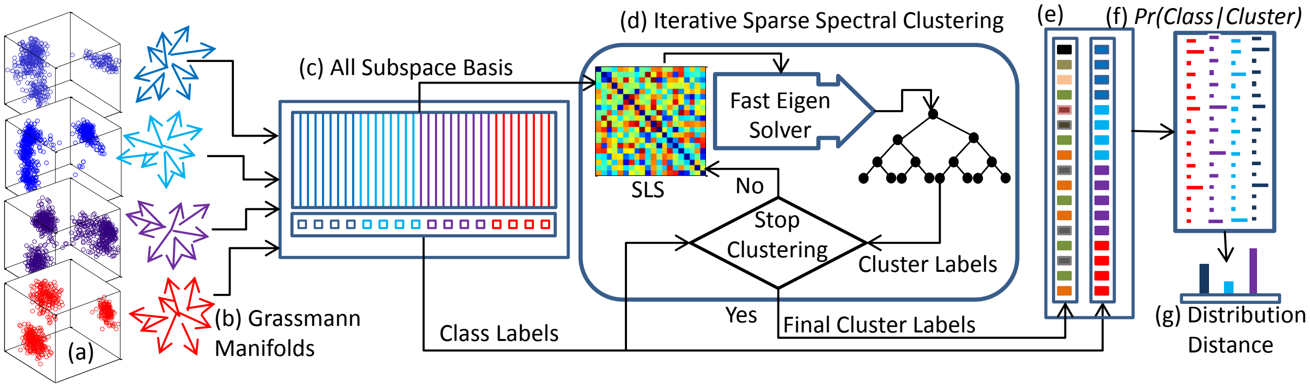

Figure 2. Proposed algorithm:(a) Face manifolds of 4 subjects from CMU Mobo dataset in PCA space. (b) Each set transformed to a Grassmannian manifold. (c) Data matrix and class label vector. (d) Iterative Sparse Spectral Clustering Algorithm: Proximity matrix based on Sparse Least Squares, a novel fast Eigenvector solver and supervised termination criterion. (e) Each class element assigned a latent cluster label. (f) Probability distribution of each class over the set of clusters. (g) Class probability distribution based distance measure.

al. [2] achieved improved performance by computing the distance between different locally linear subspaces. In all these techniques, classification is dominated by either a few data points, only one cluster, one local subspace, or one basis of the global set structure, while the rest of the image set data or local structure variations are ignored.

We propose a new framework in which the final classification decision is based on all data/clusters/subspaces contained in all image sets. Therefore, classification decisions are global compared to the existing local decisions. For this purpose, we apply unsupervised clustering on all data contained in the gallery and the probe sets, without enforcing set boundaries. We find natural data clusters based on the true data characteristics without using labels. Set labels are used only to determine the required number of clusters i.e. in the termination criterion. The probability distribution of each class over the set of natural data clusters is computed. Classification is performed by measuring distances between the probability distributions of different classes.

The proposed framework is generic and applicable to any unsupervised clustering algorithm. However, we propose an improved version of the Sparse Subspace Clustering (SSC) algorithm [6] for our framework. SSC yields good results if different classes span independent or disjoint subspaces and is not directly applicable to the image-set classification problem where the subspaces of different classes are neither independent nor disjoint. We propose various improvements over the SSC algorithm. We obtain the proximity matrix by sparse least squares with more emphasis on the reconstruction error minimization than the sparsity induction. We perform clustering iteratively and in each iteration divide a parent cluster into small number of child clusters using the *NCut* objective function. Coarser clusters capture the global data variations while the finer clusters, in the later iterations, capture the subtle local variations not visible at the global level. This coarse to fine scheme allows discrimination between globally similar data elements.

The proposed clustering algorithm can be directly used with the set samples however, we represent each set with a Grassmannian manifold and perform clustering on the manifold basis matrices. This strategy reduces computational complexity and increases discrimination and robustness to different types of noise in the image sets. Using Grassmann manifolds, we make an ensemble of spectral classifiers which further increases accuracy and gives reliability (confidence) of the label assignment.

Our final contribution is a fast eigenvector solver based on the group power iterations method. The proposed algorithm iteratively finds all eigenvectors simultaneously and terminates when the signs of the required eigenvector coefficients become stable. This is many times faster than the Matlab SVDS implementation and yields clusters of the same or better quality. Experiments are performed on three standard face image-sets (Honda, Mobo and Youtube Celebrities), an object categorization (ETH 80), and Cambridge hand gesture datasets. Results are compared to seven state of the art algorithms. The proposed technique achieves the highest accuracy on all datasets. The maximum improvement is observed on the most challenging Youtube dataset where our algorithm achieves 11.4% higher accuracy than the best reported.

2. Image Set Classification by Semi-supervised Clustering

Most image-set classification algorithms try to enforce the set boundaries in a supervised way i.e. label assignment to the training data is often manual and based on the real world semantics or prior information instead of the underlying data characteristics. For example, all images of the same subject are pre-assigned the same label despite that the intra subject dissimilarity may exceed that of inter subject. Therefore, the intrinsic data clusters may not align well with the imposed class boundaries.

Wang et. al. [24, 28] computed multiple local clusters (sub-sets) within the set boundaries. Subsets are individually matched and a gallery class containing a subset maximally similar to a query subset becomes the label winner. This may address the problem of within set dissimilarity but most samples do not play any role in the label estimation.

We propose to compute the set similarities based on the probability distribution of each data-set over an exhaustive number of natural data clusters. We propose unsupervised clustering to be performed on all data, the probe and all the training sets combined without considering labels. By doing so, we get natural clusters based on the inherent data characteristics. Clusters are allowed to be formed across two or more gallery classes. Once an appropriate number of clusters are obtained, we use labels to compute class probability distribution over the set of clusters.

Let n_k be the number of natural clusters and n_c be the number of gallery classes. For a class c_i having n_i data points, let $\mathbf{p}_i \in \mathcal{R}^{n_k}$ be the distribution over all clusters: $\sum_{k=1}^{n_k} p_i[k] = 1$ and $1 \geq p_i[k] = n_i[k]/n_i \geq 0$, where $n_i[k]$ are the data points of class c_i in the k-th cluster. Our basic framework does not put any condition on n_k however, we argue by Lemma 2.1 that an optimal number of clusters exist and can be found for the task of set label estimation. The derivation of Lemma 2.1 is based on the notion of 'conditional orthogonality' and 'indivisibility' of clusters as defined below. We assume that classes c_i and c_j belong to the gallery ($\mathcal{G}$) with known labels while c_p is the probe set with unknown label.

Conditional Orthogonality Distribution of class c_i is conditionally orthogonal to the distribution of a class c_j w.r.t the distribution of probe set c_p if

$$(\mathbf{p}_i \perp_c \mathbf{p}_j)_{\mathbf{p}_p} := \langle\!\langle (\mathbf{p}_i \wedge \mathbf{p}_p), (\mathbf{p}_j \wedge \mathbf{p}_p) \rangle\!\rangle = 0 \quad \forall (c_i, c_j) \in \mathcal{G}, \tag{1}$$

where $\wedge$ is the logical AND operation and $\langle\!\langle \cdot \rangle\!\rangle$ is the inner product. If both operands of $\wedge$ are non-zero then the result will be 1 otherwise result will be 0.

Indivisible Cluster A cluster k^* is indivisible from the probe set label estimation perspective if

$$p_i[k^*] \wedge p_j[k^*] \wedge p_p[k^*] = 0 \quad \forall (c_i, c_j) \in \mathcal{G}. \tag{2}$$

A cluster is 'indivisible' if either exactly one gallery class has non zero probability over that cluster or the probability of probe set is zero.

Lemma 2.1 *Optimal number of clusters for the set labeling problem are the minimum number of clusters such that all gallery class distributions become orthogonal to each other w.r.t the probe set distribution.*

$$n_k^* \triangleq \min_{n_k} (\boldsymbol{p}_i \perp_c \boldsymbol{p}_j)_{\boldsymbol{p}_p} \qquad \forall c_i, c_j \in \mathcal{G} \tag{3}$$

We only discuss an informal proof of Lemma 2.1. The condition (3) ensures all clusters are indivisible, therefore increasing the number of clusters beyond n_k^* will not yield more discrimination. For $n_k < n_k^*$, there will be some clusters having overlap between class distributions, hence reducing the discrimination. Thus n_k^* are the optimal number of clusters for the task at hand.

Once the optimality condition is satisfied, all clusters will be *indivisible*. Existing measures may then be used for the computation of distances between class-cluster distributions. For this purpose, we consider Bhattacharyya and Hellinger distance measures. We also propose a modified Bhattacharyya distance which we empirically found more discriminative than the existing measures.
Bhattacharyya distance ($\mathcal{B}_{i,p}$) between a class $c_i \in \mathcal{G}$ and the probe-set c_p, having $p_i(k)$ and $p_p(k)$ probability over the k-th cluster is

$$\mathcal{B}_{i,p} = -\ln \sum_{k=1}^{n_k^*} \sqrt{p_i(k) p_p(k)}. \tag{4}$$

In (4), $\ln(0) := 0$, therefore $0 \leq \mathcal{B}_{i,p} \leq 1$.

Hellinger distance is the ℓ_2 norm distance between two probability distributions

$$\mathcal{H}_{i,p} = \frac{1}{\sqrt{2}} \sqrt{\sum_{k=1}^{n_k^*} (\sqrt{p_i[k]} - \sqrt{p_p[k]})^2}, \tag{5}$$

Modified Bhattacharyya Distance ($\mathcal{B}^M_{i,p}$) of gallery class c_i with probe class c_p is given by

$$\mathcal{B}^M_{i,p} = -\ln \langle\!\langle \sqrt{\mathbf{p}_i}, \sqrt{\mathbf{p}_i} \cdot (\mathbf{p}_i \wedge \mathbf{p}_p) \rangle\!\rangle \langle\!\langle \sqrt{\mathbf{p}_p}, \sqrt{\mathbf{p}_p} \cdot (\mathbf{p}_i \wedge \mathbf{p}_p) \rangle\!\rangle, \tag{6}$$

where $(\cdot)$ in this definition is a point-wise multiplication operator and have precedence over inner product.

Lemma 2.2 *Bhattacharyya distance is upper bounded by the modified Bhattacharyya distance: $\mathcal{B}^M_{i,p} \geq \mathcal{B}_{i,p}$*

Proof simply follows from Cauchy Schwartz inequality:

$$\mathcal{B}^M_{i,p} \geq -\ln \langle\!\langle \sqrt{\mathbf{p}_i}, \sqrt{\mathbf{p}_p} \rangle\!\rangle.$$

At the extreme cases, when the angle between the two distributions is 0 or 90^o, $\mathcal{B}^M_{i,p} = \mathcal{B}_{i,p}$, while for all other cases $\mathcal{B}^M_{i,p} > \mathcal{B}_{i,p}$. Note that the factors forced to be zero in $\mathcal{B}^M_{i,p}$ by introducing the $\wedge$ operator are automatically canceled in $\mathcal{B}_{i,p}$. Performance of the three measures was experimentally compared and we observed that $\mathcal{B}^M_{i,p}$ achieves the highest accuracy.

3. Spectral Clustering

The basic idea is to divide data points into different clusters using the spectrum of proximity matrix which represents an undirected weighted graph. Each data point corresponds to a vertex and edge weights correspond to similarity between the two points. Let $\mathcal{G} = \{X_i\}_{i=1}^{g} \in \mathcal{R}^{l\times n_g}$ be the collection of data points in the gallery sets. Here, $n_g = \sum_{i=1}^{g} n_i$ are the total number of data points in the gallery. The i-th image set has n_i data points each of dimension l. The gallery contains g sets and n_c classes: $g \geq n_c$, $X_i = \{x_j\}_{j=1}^{n_i} \in \mathcal{R}^{l\times n_i}$. Each data point x_j could be a feature vector or simply the pixel values.

Let $c_p = \{x_i\}_{i=1}^{n_p} \in \mathcal{R}^{l\times n_p}$ be the probe-set with a dummy label n_c+1. We make a global data matrix by appending all gallery sets and the probe set: $\mathcal{D} = [\mathcal{G} \;\; c_p] \in \mathcal{R}^{l\times n_d}$, where $n_d = n_g + n_p$. The affinity matrix $A \in \mathcal{R}^{n_d\times n_d}$ is computed as

$$A_{i,j} = \begin{cases} \exp \frac{-\|x_i - x_j\|_2^2}{2\sigma^2} \text{ if } i \neq j \\ 0 \text{ if } i = j. \end{cases} \tag{7}$$

From A, a degree matrix D is computed

$$D(i,j) = \begin{cases} \sum_{i=1}^{n_d} A(i,j) \text{ if } i = j \\ 0 \text{ if } i \neq j, \end{cases} \tag{8}$$

Using A and D, a Laplacian matrix L is computed

$$L_w = D^{-1/2} A D^{-1/2}. \tag{9}$$

Let $E = \{e_i\}_{i=1}^{n_c}$ be the matrix of n_c smallest eigenvectors of L_w. The eigenvectors of the Laplacian matrix embed the graph vertices into a Euclidean space where NN approach can be used for clustering. Therefore, the rows of E are unit normalized and grouped into n_c clusters using kNN.

3.1. Proximity Matrix As Sparse Least Squares

Often high dimensional data sets lie on low dimensional manifolds. In such cases, the Euclidean distance based proximity matrix is not an effective way to represent the geometric relationships among the data points. A more viable option is the sparse representation of data which has been used for many tasks including label propagation [4], dimensionality reduction, image segmentation and face recognition [11]. Alhamifar and Vidal [6] have recently proposed sparse subspace clustering which can discriminate data lying on independent and disjoint subspaces.

A vector can only be represented as a linear combination of other vectors spanning the same subspace. Therefore, the proximity matrices based on linear decomposition of data points lead to subspace based clustering. Representing a data point x_i as a linear combination of the remaining data points $\hat{\mathcal{D}} = \mathcal{D}/x_i$ ensures that zero coefficients will only correspond to the points spanning different subspaces. Such a decomposition can be computed with least squares

$$\alpha_i = (\hat{\mathcal{D}}^\top \hat{\mathcal{D}})^{-1} \hat{\mathcal{D}}^\top x_i, \tag{10}$$

where α are the linear coefficients. We are mainly concerned with the face space which is neither independent nor disjoint across different subjects. Therefore, introducing a sparsity constraint on α ensures that the linear coefficients from less relevant subspaces will be forced to zero:

$$\alpha_i^* := \min_{\alpha_i} \left(\|x_i - \hat{\mathcal{D}}\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 \right) . \tag{11}$$

The first term minimizes the reconstruction error while the second term induces sparsity. This process is repeated for all data points and the corresponding α_i are appended as columns in a matrix $S = \{\alpha_i\}_{i=1}^{n_d} \in \mathcal{R}^{n_d\times n_d}$. Some of the α coefficients may be negative and in general $S(i,j) \neq S(j,i)$. Therefore, a symmetric sparse LS proximity matrix is computed as $A = |S| + |S^\top|$ for spectral clustering.

3.2. Iterative Hierarchical Sparse Spectral Clusters

Conventionally, sparse spectral clustering is performed by simultaneous partitioning of the graph into n^k clusters [6]. We argue that iterative hierarchical clustering has many advantages. In each iteration, we divide the graph into very few partitions/clusters (four in our implementation). If a cluster is not indivisible (2), we recompute the local sparse LS based proximity matrix only for that cluster and then re-compute the eigenvectors. Note that we do not reuse a part of the initial proximity matrix, because the sparse LS gives a different matrix due to reduced number of candidates. The new matrix highlights only local connectivity as opposed to global connectivity at the highest level. Coarser clusters capture the large global variations while the finer clusters, obtained later, capture the subtle inter class differences which may not be visible at the global scale. As a result, we are able to locally differentiate between data points which were globally similar. Moreover, as we explain next, in the case of simultaneous partitioning, an approximation error accumulates which adversely affects the cluster quality, whereas the iterative hierarchical approach enables us to obtain high quality clusters.

Let a_l and a_r be the two disjoint partitions. We want to find a graph cut that minimizes the sum of edge weights $\sum_{i\in|a_l|, j\in|a_r|} A(i,j)$ across the cut. MinCut is easy to implement but it may give unbalanced partitions. In the extreme case, minCut may separate only one node from the remaining graph. To ensure balanced partitions, we minimize the normalized cut $NCut$ objective function [25, 7]

$$\frac{1}{V_{a_l}} \sum_{i\in|a_l|, j\in|a_r|} A(i,j) + \frac{1}{V_{a_r}} \sum_{i\in|a_l|, j\in|a_r|} A(i,j), \tag{12}$$

where V_{a_l} is the sum of all edge weights attached to the vertices in a_l. The objective function increases with the decrease in the number of nodes in a partition. Unfortunately, the minCut using the $NCut$ objective function is NP hard and only approximate solutions can be computed using the spectral embedding approach. It has been shown in [25] that the eigenvector corresponding to the second smallest eigenvalue (e_2) of

$$L_{sym} = D^{-1/2}(D - A)D^{-1/2} \tag{13}$$

provides an approximate solution to a relaxed $NCut$ problem. All data points corresponding to $e_2(i) \geq 0$ are assigned to a_l while the remaining ones to a_r. Similarly, e_3 can further divide each cluster into two more clusters and the higher order eigenvectors can be used for further partitioning the graph into smaller clusters. However, the approximation error accumulates degrading the clustering quality.

Although good quality clusters can be obtained using the iterative approach, it requires more computations because the proximity matrix and eigenvector computations must be repeated at each iteration. If the data matrix $\mathcal{D}$ is divided into four balanced partitions each time, the size of the new matrices are 16 times smaller. Since the complexity of eigenvector solvers is $O(n_d^3)$, the complexity reduction in the next iteration is $O((\frac{n_d}{4})^3))$ for each sub-problem. The depth of the recursive tree is $\log_4(n_d)$, however the proposed supervised stopping criteria does not let the iterations to continue until the end, rather the process stops as soon as all clusters are indivisible. A computational complexity analysis of the recursion tree reveals that the overall complexity of the eigenvector computations remains $O(n_d^3)$.

4. Computational Cost Reduction

We propose two approaches for computational complexity reduction of spectral clustering. The first approach reduces the data by embedding each image set on a Grassmann manifold and then using the manifold basis vectors to represent the set. The second approach is a fast eigenvector solver which quickly computes approximate eigenvectors using an early termination criterion of sign changes.

4.1. Data Reduction by Grassmann Manifolds

Eigenvectors computation of large Laplacian matrices $L_{sym} \in \mathcal{R}^{n_d \times n_d}$ incurs high computational cost. Often a part of the data matrix $\mathcal{D}$ or some columns from L_{sym} are sampled and eigenvectors are computed for the sampled data and extrapolated for the rest [7, 22, 18]. Approximate approaches provide sufficient accuracy for computing the most significant eigenvectors but are not as accurate for the least significant eigenvectors which are actually required in spectral clustering. In contrast, we propose to represent each set by a compact representation and clustering to be performed on the representation instead of the original data.

Our choice of image set representation is motivated from linear subspace based image-set representations [28, 26]. These subspaces may be considered as points on Grassmannian manifolds [9, 10]. While others performed discriminant analysis on the Grassmannian manifolds or computed manifold to manifold distances, we perform sparse spectral clustering on the Grassmannian manifolds.

A set of λ-dimensional linear subspaces of $\mathcal{R}^n$, $n = \min(l, n_j)$ and $\lambda \leq n$, is termed the Grassmann manifold $Grass(\lambda, n)$. An element $\mathcal{Y}$ of $Grass(\lambda, n)$ is a λ-dimensional subspace which can be specified by a set of λ vectors: $Y = \{y_1, ..., y_\lambda\} \in \mathcal{R}^{l \times \lambda}$ and $\mathcal{Y}$ is the set of all their linear combinations. For each data element of the image set, we compute a set of basis Y and the set of all such Y matrices is termed as the non-compact Stiefel manifold $ST(\lambda, n) := \{Y \in R^{l \times \lambda} : \text{rank}\ (\mathcal{Y}) = \lambda.\}$. We arrange all the Y matrices in a basis matrix B which is capable of representing each data point in the image set by just using λ of its columns. For the i-th data point in j-th image set $x_j^i \in X_j$, having B_j as the basis matrix, $x_j^i = B_j \alpha_j^i$, where α_j^i is the set of linear parameters with $|\alpha_j^i|_o = \lambda$. For the case of known B_j, we can find a matrix $\alpha_j = \{\alpha_j^1, \alpha_j^2, \cdots \alpha_j^{n_i}\}$ such that the residue approaches zero

$$\min_{\alpha_j}(\sum_{i=1}^{n_i} \|x_j^i - B_j \alpha_i^j\|_2^2) \quad \text{s.t.}\ \|\alpha_j\|_o \leq \lambda\ . \tag{14}$$

Since ℓ_1 can approximate ℓ_o, we can estimate both α_j and B_j iteratively by using the following objective function [16]

$$\min_{\alpha_j, B_j} \left(\frac{1}{n_j} \sum_{i=1}^{n_j} \frac{1}{2} \|x_j^i - B_j \alpha_j^i\|_2^2 \quad \text{s.t.}\ \|\alpha_j^i\|_1 \leq \lambda \right)\ . \tag{15}$$

The solution is obtained by randomly initializing B_j and computing α_j, then fixing α_j and computing B_j. The size of $B_j \in \mathcal{R}^{l \times \Lambda}$ is significantly smaller than the corresponding image set $X_j \in R^{l \times n_j}$. Representing each image set with Λ basis vectors reduces the data matrix from $n_d \times n_d$ to $\Lambda g \times \Lambda g$, which significantly reduces the computational cost. Additionally, we observe that this representation also provides significant increase in the accuracy of the proposed clustering algorithm because the underlying subspaces of each class are robustly captured in B leaving out noise.

4.2. Fast Approximate Eigenvector Solver

The size of Laplacian matrix still grows with the number of subjects. To reduce the cost of eigenvector computations, we propose a fast group power iteration method which finds all the eigenvectors and eigenvalues simultaneously, given sufficient number of iterations. The proposed algorithm always converges and has good numerical stability.

The signs of the e_2 eigenvector coefficients (corresponding to the minimum non-zero eigenvalue) of L_{sym} provides an approximate solution to a relaxed $NCut$ problem [25]. Therefore, if we get the correct signs with approximate magnitude of the eigenvectors, we will still get the same quality of clusters. We exploit this fact in our proposed eigenvector solver and terminate the iterations when the number of sign changes falls below a threshold.

Power iterations algorithm is a fundamental technique for eigenvalues and eigenvectors computation [31]. A random vector $\mathbf{v}$ is repeatedly multiplied by the matrix L and normalized. After k iterations $A\mathbf{v}^{(k)} = \lambda\, \mathbf{v}^{(k)}$, where $\mathbf{v}^{(k)}$ is the most dominant eigenvector of L and λ the corresponding eigenvalue. To calculate the next eigenvector, the same process is repeated on the deflated matrix L.

Group power iteration method can be used for the computation of all eigenvectors simultaneously. A random matrix V_r is iteratively multiplied by L. After each multiplication, a unit normalization step is required by division with $V_r^\top V_r$ which converges to a diagonal matrix as V_r converges to an orthonormal matrix. To stop all columns from converging to the most dominant eigenvector, an orthogonalization step is also required. We use QR decomposition for this purpose: $V_r^{(k)} R_r^{(k)} = V_r^{(k)}$. A simple group power iteration method will have poor convergence properties [31]. To overcome this, we apply convergence from the left and the right sides simultaneously by computing the left and the right eigenvectors (see Algorithm 1). We observe that Algorithm 1 always converges to the correct solution and has better numerical properties than the group power iteration.

For the proposed spectral clustering, only the signs are important. Therefore, in Algorithm 1, we replace the eigenvalue convergence based termination criterion with stabilization of sign changes criterion between consecutive iterations as follows

$$\Delta S = \sum (V_r^{(k)} > 0) \oplus (V_r^{(k-1)} > 0), \qquad (16)$$

where $\oplus$ is the XOR operator. $0 \oplus 1 = 1$, $1 \oplus 0 = 1$, $0 \oplus 0 = 0$, $1 \oplus 1 = 0$. We empirically found that most of the signs become stable after very few iterations (≤ 4).

5. Ensemble of Spectral Classifiers

Representing image sets with Grassmannian manifolds facilitates formulation of an ensemble of spectral classifiers. Different random initializations of B_j^0 in (15) may converge to different solutions resulting in multiple image set representations. In addition to that, we also vary the dimensionality of manifolds and compute a set of manifolds for each class. The proposed spectral classifier is independently applied to the manifolds of the same dimensionality over all classes and the inter class distances are estimated. We fuse the set of distances using mode fusion and also by sum rule. In mode fusion, the probe set label is estimated individually for each classifier and the label with the maximum frequency is selected. In sum fusion, minimum distance over the cumulative distance vector defines the probe set label.

6. Experimental Evaluation

Evaluations are performed for image-set based face recognition, object categorization and gesture recognition. The SPAMS [21] package is used for sparse coding. Comparisons are performed with Discriminant Canonical Correlation (DCC) [26], Manifold-Manifold Distance (MMD) [28], Manifold Discriminant Analysis (MDA) [24], linear Affine and Convex Hull based Image Set Distance (AHISD, CHISD) [8], Sparse Approximated Nearest Points [30], and Covariance Discriminative Learning (CDL) [27]. The same experimental protocol is used for all algorithms. The codes of [8, 26, 28, 30] were provided by the original authors.

Algorithm 1 Fast Eigen Solver: Group Power Iteration

```
Input: $L \in \mathcal{R}^{n\times n}, \epsilon_L$
Output: $U, V$ {Eigenvectors}, $\Lambda$ {Eigenvalues}
  $V_r^{(0)} = I(n,n)$ {Identity matrix}
  $\Lambda^{(0)} = L$, $\delta\Lambda = 1$
  while $\delta\Lambda \geq \epsilon_L$ do
    $V_l^{(k)} \leftarrow L^\top V_r^{(k-1)}$
    $V_l^{(k)} \leftarrow V_l^{(k)} / (V_l^{(k)\top} V_l^{(k)})$
    $V_l^{(k)} R_l^{(k)} \leftarrow V_l^{(k)}$ {left qr decomposition}
    $V_r^{(k)} \leftarrow V_l^{(k)}$
    $V_r^{(k)} \leftarrow V_r^{(k)} / (V_r^{(k)\top} V_r^{(k)})$
    $V_r^{(k)} R_r^{(k)} \leftarrow V_r^{(k)}$ {right qr decomposition}
    $\Lambda^{(k)} \leftarrow V_r^{(k)} L V_l^{(k)\top}$
    $\delta\Lambda = \|\mathrm{diag}(\Lambda^{(k)} - \Lambda^{(k-1)})\|_2$
  end while
  $U \leftarrow V_r^{(k)}, V \leftarrow V_l^{(k)}$
```

6.1. Face Recognition using Image-sets

Our first dataset is the You-tube Celebrities [19] which is very challenging and includes 1910 very low resolution videos (of 47 subjects) containing motion blur, high compression, pose and expression variations (Fig. 3c). Faces were automatically detected, tracked and cropped to 30×30 gray-scale images. Due to tracking failures, our sets contained fewer images (8 to 400 per set) than the total number of video frames. The proposed algorithm performed best on HOG features. Five-fold cross validation experiments were performed where 3 image sets were randomly selected for training and the remaining 6 for testing.

Each image-set was represented by two manifold-sets by using $\lambda = \{1, 2\}$ in (15). Each set has 8 classifiers with dimensionality increasing from 1 to 8. Recognition rates of the ensembles of each dimensionality are compared with both fusion schemes in Fig. 4a. No-fusion case shows the average accuracy of the individual classifiers over the five folds. Note that mode fusion achieves the highest accuracy and performs better than sum fusion because error does not get accumulated. Hierarchical clustering is compared with one-step clustering in 4b which shows the superiority of the hierarchical approach. The performance of the proposed eigenvector solver is compared with the Matlab sparse SVD

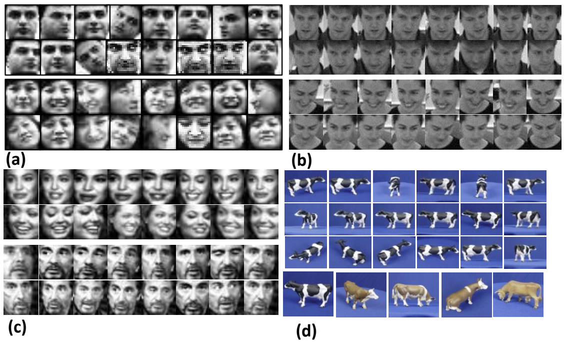

Figure 3. Two example image-sets from (a) Honda, (b) CMU Mobo and (c) You-tube Celebrities datasets. (d) ETH 80 dataset.

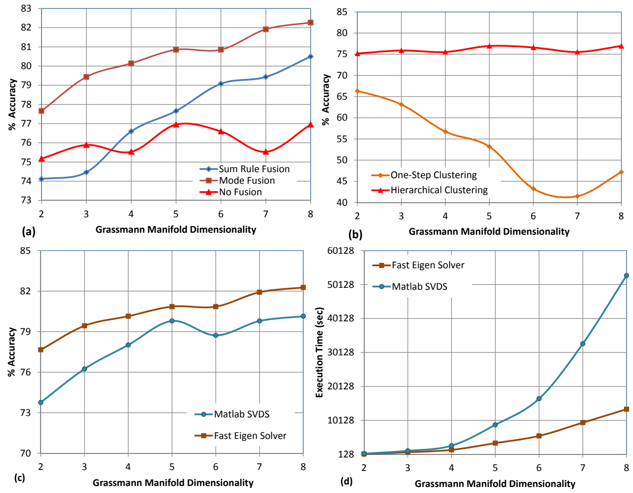

Figure 4. You-tube dataset: (a) Comparison of different spectral ensemble fusion schemes. (b) Comparison of one step and the iterative clustering (No Fusion). (c-d) Accuracy and execution time comparison of the proposed fast eigen-solver with Matlab SVDS.

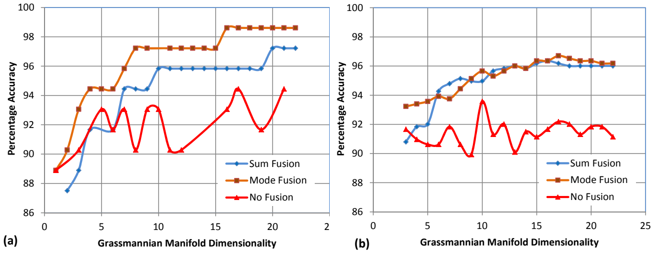

Figure 5. Mobo dataset: Comparison of (a) hierarchical with (b) one-step spectral clustering.

(svds) in Fig. 4c. We observe more accuracy due to better quality embedding of Grassmannian Manifolds in the Euclidean space by the proposed solver. In terms of execution time, our eigenvector solver is around 4 times faster than SVDS (Fig. 4d). This demonstrates the efficacy of our eigenvector solver for the purpose of spectral clustering. The maximum accuracy of our algorithm is 83% and the average accuracy is 76.4±5.7% using $\mathcal{B}^M_{i,p}$ (6) distance measure (1). To the best of our knowledge, this is the highest accuracy to date reported on this dataset.

Table 1. Average recognition rates over 10-folds on Honda, MoBo, & ETH80, 5-fold on YouTube and 1-fold on Cambrage dataset.

	Honda	MoBo	ETH80	You-Tube	Cambr.
DCC	94.7±1.3	93.6±1.8	90.9±5.3	53.9±4.7	65.0
MMD	94.9±1.2	93.2±1.7	85.7±8.3	54.0±3.7	58.1
MDA	97.4±0.9	97.1±1.0	80.50±6.81	55.1 ±4.5	20.9
AHISD	† 89.7±1.9	97.4±0.8	74.76±3.3	60.7±5.2	18.1
CHISD	† 92.3±2.1	96.4±1.0	71.0±3.9	60.4±5.9	18.3
SANP	93.1±3.4	96.9±0.6	72.4±5.0	65.0±5.53	22.5
CDL	**100±0.0**	95.8±2.0	89.2±6.8	⋆62.2±5.1	73.4
Prop.	**100±0.0**	**98.0±0.9**	**91.5±3.8**	**76.4±5.7**	**83.05**

⋆ CDL results are on different folds therefore, the accuracy is less than that reported by [27]. †The accuracy of AHISD and CHISD is less than in [8] due to smaller image sizes.

The second dataset is CMU Mobo [23] containing 96 videos of 24 subjects. Face images were re-sized to 40 × 40 and LBP features were computed using circular (8, 1) neighborhoods extracted from 8 × 8 gray scale patches [8]. We performed 10-fold experiments by randomly selecting one image-set per subject as training and the remaining 3 as probe. We achieved a maximum accuracy of 100% and average 98.0±0.93% (Table 1) which is the highest reported so far.

Our final dataset is Honda/UCSD [13] containing 59 videos of 20 subjects with varying poses and expressions. Histogram equalized 20×20 gray scale face image pixel values were used as features [28]. We performed 10-fold experiments by randomly selecting one set per subject as gallery and the remaining 39 as probes. An ensemble of 15 spectral classifiers obtained 100% accuracy on all folds.

6.2. Object Categorization & Gesture Recognition

For object categorization, we use the ETH-80 dataset containing images of 8 object categories each with 10 different objects. Each object has 41 images taken at different views forming an image-set. We use 20×20 intensity images for classifying an image-set of an object into a category. ETH-80 is a challenging database because it has fewer images per set and significant appearance variations across objects of the same class. For each class, 5 random image sets are used for training and the remaining 5 for testing. We achieved an average recognition rate of 91.5±3.8% using an ensemble of 15 spectral classifiers.

The Cambridge Hand Gesture dataset [15] contains 900 image-sets of 9 gesture classes with large intra-class variations. Each class has 100 image sets, divided into two parts, 81-100 used as gallery and 1-80 as probes [14]. Pixel values of 20 × 20 gray scale images are used as feature vectors. Using an ensemble of 9 spectral classifiers, we obtained an accuracy of 83.05% which is higher than the other algorithms.

6.3. Robustness to Outliers

We performed robustness experiments in a setting similar to [8]. Honda dataset was modified to have 100 randomly selected images per set. In the first experiment, each

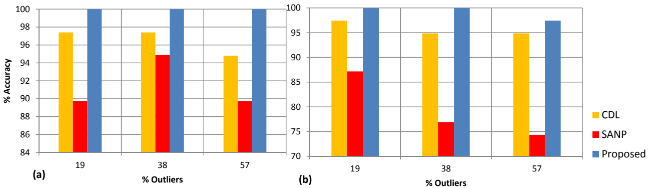

Figure 6. Robustness to outliers experiment: (a) Corrupted gallery case (b) Corrupted probe case.

gallery set was corrupted by adding 1 to 3 random images from each other gallery set resulting in 19%, 38% and 57% outliers. Our algorithm achieved 100% accuracy for all three cases. In the second experiment, the *probe* set was corrupted by adding 1 to 3 random images from each gallery set. In this case, our algorithm achieved {100%, 100%, 97.43%} recognition rates. Our algorithm outperformed all others. Fig. 6 compares our algorithm to the nearest 2 competitors in both experiments i.e. CDL [27], SANP [30].

7. Conclusion

We presented an iterative sparse spectral clustering algorithm for robust image-set classification. Each image-set is represented with Grassmannian manifolds of increasing dimensionality to facilitate the use of an ensemble of spectral classifiers. An important contribution is a fast eigenvector solver which makes spectral clustering more efficient in general. Instead of eigenvalue error, we minimize the sign changes in our group power iteration algorithm which provides significant speedup and better quality clusters.

8. Acknowledgements

This research was supported by ARC grants DP1096801 and DP110102399. We thank T. Kim and R. Wang for the codes of DCC and MMD and the cropped faces of Honda data. We thank H. Cevikalp for the LBP features of Mobo data and Y. Hu for the MDA implementation.

References

[1] A. Mahmood and A. Mian. Hierarchical sparse spectral clustering for image set classification. In *BMVC*, 2012. 1

[2] S. Chen, C. Sanderson, M. T. Harandi, and B. C. Lovell. Improved image set classification via joint sparse approximated nearest subspaces. In *CVPR*, 2013. 1, 2

[3] M. Yang, P. Zhu, L. Gool, and L. Zhang. Face recognition based on regularized nearest points between image sets. In *FG*, 2013. 1

[4] H. Cheng, Z. Liu, and L. Yang. Sparsity induced similarity measure for label propagation. In *ICCV*, 2009. 4

[5] Z. Cui, S. Shan, H. Zhang, S. Lao, X. Chen. Image sets align -ment for video based face recognition. *CVPR*, 2012. 1

[6] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *TPAMI*, 2013. 2, 4

[7] C. Fowlkes, S. Belongie, F. Chung, J. Malik. Spectral grouping using the Nystrom method. *TPAMI*, 2004. 4, 5

[8] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573, 2010. 1, 6, 7

[9] J. Hamm and D. Lee. Grassmann discriminant analysis: a unifying view on subspace learning. In *ICML*, 2008. 5

[10] M. Harandi, C. Sanderson, S. Shirazi, B. Lovell. Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching *CVPR*, 2011. 1, 5

[11] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma. Robust face recognition via sparse representation. *TPAMI*, 31(2):210–227, 2009. 4

[12] Y. Chen, V. Patel, P. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In *ECCV*. 2012. 1

[13] K. Lee, J. Ho, M. Yang and D. Kriegman. Video based face recognition using probabilistic appearance manifolds. In *CVPR*, 2003. 7

[14] T. Kim and R. Cipolla. Canonical correlation analysis of video volume tensors for action categorization and detection. *TPAMI*, 31(8), 2009. 7

[15] T. Kim, K. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *CVPR*, 2007. 7

[16] H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *NIPS*, 2007. 5

[17] H Li, G Hua, Z Lin, and J Yang Probabilistic elastic matching for pose variant face verification. In *CVPR*, 2013. 1

[18] M. Li, X. Lian, J. K., B. Lu. Time and space efficient spectral clustering via column sampling. In *CVPR*, 2011. 5

[19] M. Kim, S. Kumar, V. Pavlovic and H. Rowley. Face tracking and recognition with visual constraints in real world videos. In *CVPR*, 2008. 6

[20] M. Nishiyama, M. Yuasa, T. Shibata, T. Wakasugi, T. Kawahara and Yamaguchi. Recognizing faces of moving people by hierarchical image-Set matching. In *CVPR*, 2007. 1

[21] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*, 2009. 6

[22] X. Peng, L. Zhang, and Z. Yi. Scalable sparse subspace clustering. *CVPR*, 2013. 5

[23] R. Gross and J. Shi. The CMU Motion of Body (MoBo) Database. Technical Report CMU-RI-TR-01-18, 2001. 7

[24] R. Wang and X. Chen. Manifold discriminant analysis. In *CVPR*, 2009. 1, 3, 6

[25] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 2000. 4, 5

[26] T. Kim, O. Arandjelovic and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *TPAMI*, 2007. 1, 5, 6

[27] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, 2012. 6, 7, 8

[28] R. Wang, S. Shan, X. Chen, Q. Dai, and W. Gao. Manifold-manifold distance and its application to face recognition with image sets. *TIP*, 2012. 1, 3, 5, 6, 7

[29] B Wu, Y Zh, B Hu, Q Ji Constrained clustering and its application to face clustering in videos. In *CVPR*, 2013. 1

[30] Yiqun Hu and Ajmal S. Mian and Robyn Owens. Face recognition using sparse approximated nearest points between image sets. *TPAMI*, 2012. 1, 6, 8

[31] G. H. Golub and C. V. Loan. Matrix Computations. *The John Hopkins University Press*, 1996. 6