Hierarchical Sparse Spectral Clustering for Image Set Classification

Arif Mahmood and Ajmal S. Main
{arifm,ajmal}@csse.uwa.edu.au

Computer Science and Software Engineering, The University of Western Australia, Crawley WA 6009

Abstract

We present a structural matching technique for robust classification based on image sets. In set based classification, a probe set is matched with a number of gallery sets and assigned the label of the most similar set. We represent each image set by a *sparse* dictionary and compute a similarity matrix by matching all the dictionary atoms of the gallery and probe sets. The similarity matrix comprises the sparse coding coefficients and forms a fully connected directed graph. The nodes of the graph are the dictionary atoms and the edges are the sparse coefficients. The graph is converted to an undirected graph with positive edge weights and spectral clustering is used to cut the graph into two balanced partitions using the normalized cut algorithm. This process is repeated until the graph reduces to critical and non-critical partitions. A critical partition contains atoms with the same gallery label along with one or more probe atoms whereas a non-critical partition either consists of only probe atoms or atoms with multiple gallery labels with no probe atom. Using the critical partitions, we define a novel set based similarity measure and assign the probe set the label of the gallery set with maximum similarity. The proposed algorithm is applied to image set based face recognition using two standard databases. Comparison with existing techniques shows the validity and robustness of our algorithm in the presence of outlier images.

1 Introduction

In image set based classification, each training class is represented by one or more image sets and each set contains multiple images with the same label. The query set also consists of multiple images with the same but unknown label. The query image set is assigned the label of the nearest class using some similarity criterion. Although, nearest neighbor techniques can still be applied to image set classification, they do not fully exploit the within set structure which offers additional information and is robust to outliers. Set-to-set matching is preferred because an image set offers significantly more information compared to a single image. Multiple images in a set compliment the appearance variations of a subject. Considering the example of face recognition, an image set may contain arbitrary pose and expression variations of the same person.

Image set classification is a generalization of video based classification [6, 10]. The main difference is that in videos, the adjacent frames have minor variation, are temporally related and usually acquired under similar illumination with the same sensor. On the other hand, no such assumptions are made in image sets. In fact, the images of a set may come

 http://dx.doi.org/10.5244/C.26.51

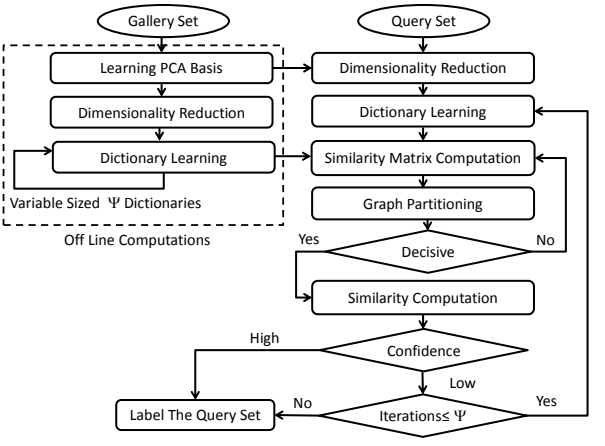

Figure 1: Block diagram of the Hierarchical Sparse Spectral Clustering (HSSC) algorithm.

from completely different sources such as personal photo albums, images collected from the Internet or multiple surveillance cameras.

Classification based on image sets has recently gained significant interest from the research community [1, 12, 16, 19, 22]. Although classification based on image sets offer more potential in terms of accuracy and robustness, it also introduces two main challenges. Firstly, the representation of an image set is challenging because of the large within-set variations and the lack of any semantic information. For example, in 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. Therefore, one of the major challenge posed by the image set classification is to efficiently exploit within-set similarities and across-set dissimilarities. One obvious approach to overcome this challenge is to divide each set into smaller subsets representable in a more compact and unambiguous way.

Image set classification techniques can be broadly divided into sample based (nearest neighbor) or structural based. Sample based techniques measure the distance between certain samples of the sets. For example, between the set centers or their nearest neighbor samples. More sophisticated techniques measure the distance between nearest neighbors of two sets under some constraints. For example, Cevikalp and Triggs [5] represent sets with affine hulls and use various bounds to constrain the search for nearest neighbors. Hu et al. [22] used the sparsity constraint to find the nearest points between two sets. Sample based techniques offer significant classification accuracy, however, they are vulnerable to outliers. For example, if a query set contains a single outlier image which is closer to a different gallery set, it will be misclassified based on that sample alone.

Structural techniques learn the underlying structure of a set, for example with one or more linear subspaces, and measure structural similarities. Manifold-manifold distance (MMD) [16] clusters the images of each set into multiple linear subspaces. The ratio between Euclidean and geodesic distance is used as a criterion to cluster the images into the same or distinct subspaces. The similarity between two sets is defined as the canonical correlation between the nearest linear subspaces. However, the authors of [16] also use the nearest neighbor as an additional criterion in their similarity measure. Kim et al. [20] per-

formed discriminative learning using canonical correlations. More specifically, a discriminative function is learned that maximizes the within-class and minimizes the between-class canonical correlations. Manifold Discriminant Analysis (MDA) [15] learns an embedding space where manifolds of different classes are better separable. Manifolds are represented by multiple linear subspaces similar to [16]. The scatter within a linear subspace is minimized and the scatter between the linear subspaces of different classes is maximized. Similarity between two sets is computed as the pairwise distances between the local linear models of their manifolds in the MDA embedding space. Structural techniques are sensitive to noise within the sets. For example, two sets of the same identity can result in very different linear subspaces or manifolds.

We propose Hierarchical Sparse Spectral Clustering (HSSC), a structure based image set classification algorithm which is robust to noise and outliers. The query and the gallery sets are represented by *sparse* dictionaries and an unsupervised clustering of the dictionary atoms is performed irrespective of their labels. The dictionary atoms are spectrally clustered into two partitions. Atoms from the same gallery or probe set can end up in either cluster. Clusters containing atoms from the query set and multiple gallery sets are non-decisive. Therefore, they are further divided until all clusters become decisive. Decisive clusters are of two types, critical and non-critical. A non-critical cluster contains either samples from only the query set or no sample from the query set. These clusters reduce the search space and are of no further use. On the other hand, a critical cluster contains some samples from the query set and a *single* gallery set. Based on the distribution of gallery and query set atoms in all critical clusters, we define a new similarity measure for set classification. We also define a confidence measure based on which we repeat the classification process with varying dictionary sizes until the probe set's identity is found with high confidence (see Fig. 1).

Experiments are performed on the Honda/UCSD [7] and CMU Mobo data [14] for face recognition based on image sets. Comparison with existing techniques shows the efficacy of the proposed algorithm. We also test robustness to outliers by mixing an increasing number of imposter images in the probe set. The proposed algorithm demonstrates significant robustness by achieving 100% recognition rate on the Honda database in the presence of up to 9 imposters selected randomly from a random gallery set and mixed with the probe set.

1.1 Contributions

To the best of our knowledge, no set classification technique similar to HSSC exists. The closest approach is the sparse subspace clustering by Elhamifar and Vidal [4] which is only meant for linear and affine subspace clustering as opposed to classification. A direct application of [4] to image set based classification is possible only if the bases of the gallery sets are linearly independent, which is never the case. If the bases were linearly independent, the classes would be linearly separable and hence classification would be trivial in that case. In difficult classification problems such as face recognition, different classes share a common structure that is why generic face detection algorithms can detect all faces.

Our main contribution is that we simultaneously perform clustering and matching. Rather than imposing constraints on the number or dimensionality of the clusters, we constrain the contents of the final clusters such that they favor un-ambiguous classification. As opposed to a fixed clustering of the gallery sets, the clustering is guided by the probe set during matching which offers more robustness to noise and variations between different sets of the same identity. More precisely, the probe set plays a role in deciding how many clusters should be

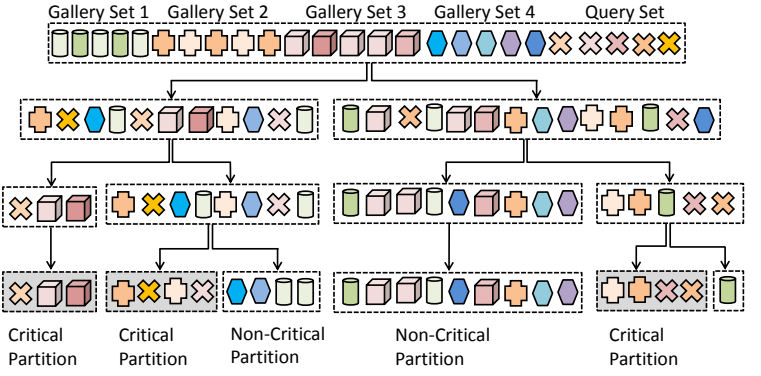

Figure 2: HSSC of a query and four gallery sets. The first clustering resulted in two non-decisive clusters, L and R. L is further divided into LL (critical) and LR (non-decisive) clusters. LR is further divided into LRL (critical) and LRR (non-critical). R is divided into RL (non-critical) and RR (non-decisive). RR is further divided into RRL (critical) and RRR (non-critical). The process stops when all clusters are either critical or non-critical.

there. Moreover, we perform hierarchical spectral clustering irrespective of the labels which is more accurate and robust to outliers. Finally, we define a novel structure based similarity measure for set classification along with a confidence measure.

2 Hierarchical Sparse Spectral Clustering (HSSC)

A schematic diagram of the HSSC algorithm is shown in Fig. 1. In the following subsections details of each step are given.

2.1 Dimensionality Reduction

The intrinsic data dimensionality in the image sets is often less than the apparent dimensions. Therefore, we reduce the data dimensionality using PCA whose basis is computed from the training (gallery) sets. Let $G = \{X_i\}_{i=1}^{g} \in \mathcal{R}^{l \times N}$ be the collection of the images (or features) of all gallery sets in vectorized form. Here, $N = \sum_{i=1}^{g} n_i$ and the i^{th} image set has n_i vectors each of dimension l. The gallery contains g image sets where each set $X_i = \{x_j\}_{j=1}^{n_i} \in \mathcal{R}^{l \times n_i}$. Each column x_j of X_i could be a feature vector (such as LBP features [18]) of the image or simply the pixel values. Note that the number of vectors n_i may vary across image sets.

The gallery vectors are mean centered and their covariance matrix is computed as $\mathcal{C} = GG^T \in \mathcal{R}^{l \times l}$. Then eigenvectors and eigenvalues of $\mathcal{C}$ are computed and the most significant m eigenvectors $E = \{e_i\}_{i=1}^{m}$ corresponding to the largest m eigenvalues $V = \{v_i\}_{i=1}^{m}$ are selected such that $\sum_{i=1}^{m} v_i^2 / \sum_{i=1}^{l} v_i^2 \geq 0.999$ i.e. 99.9% energy is retained. The gallery images are projected on the selected eigenvectors E for dimensionality reduction: $G_r = E^T G$. Each projected gallery vector is then normalized to unit magnitude.

The query image set X_q is also centered with respect to its mean, projected on the same basis E: $\widehat{X}_t = E^T \widehat{X}_t$ and normalized to unit magnitude.

2.2 Sparse Dictionary Learning

For each gallery set, we pre-compute *sparse* dictionaries of varying sizes. The spectral clustering starts using the smallest size dictionary and chooses the next size up only if the classification confidence is low. Note that the choice of *sparse*, small size, dictionary as well as pre-computation of variable size dictionaries is done to speed up the online matching process. A *sparse* dictionary must be able to represent all images in an image set as a sparse linear combination of its atoms.

Given an image set $X_i = \{x_j\}_{j=1}^{n_i} \in \mathcal{R}^{m \times n_i}$, its dictionary $D_i \in \mathcal{R}^{m \times p_i}$ should be able to minimize the cost function $\frac{1}{n}\sum_{j=1}^{n_i} f(x_j, D_i)$ [2]. Each column of D_i is called an atom and represents a basis vector for image set X_i. Unlike PCA basis, the dictionary atoms need not be orthogonal. We use the convex ℓ_1 formulation of the Lasso as the cost function [8]

$$f(x_j, D_i) = \min_{\alpha_i} \frac{1}{2}||x_j - D_i\alpha_i||_2^2 + \lambda||\alpha_i||_1 \ , \tag{1}$$

and use the Least Angle Regression (LARS) algorithm [21] to solve it. Here, $\alpha_i = \{\alpha_{ij}\}_{j=1}^{p} \in \mathcal{R}^{p_i}$ is a vector of sparse linear coefficients and λ is a regularization parameter. Substituting the value of $f(x_j, D_i)$ in the cost function

$$\min_{\alpha_i, D_i} \Big(\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{2}||x_j - D_i\alpha_i||_2^2 + \lambda||\alpha_i||_1 \Big) \ . \tag{2}$$

This cost function is not convex for unknown D_i and α_i. However, it is convex with respect to one variable if the other is known. Therefore, the solution is obtained by alternating between D_i and α_i [8]. Based on Equation 2, an efficient dictionary learning algorithm has been proposed by [11] which we use in the proposed HSSC algorithm.

2.3 Sparse Similarity Matrix Computation

We start by choosing the smallest size gallery dictionaries, each with p_i atoms per gallery set. During online matching, a dictionary with p_q atoms is learned for the query (probe) image set. Let $D_G = \{D_i\}_{i=1}^{g} \in \mathcal{R}^{m \times P}$, where $P = \sum_{i=1}^{g} p_i$ be the learned dictionaries for the gallery image sets and $D_q \in \mathcal{R}^{m \times p_q}$ be the dictionary for the query image set.

Each dictionary atom in D_G inherits a label from its parent image set whereas a test label t is assigned to each atom in D_q. Let $L_G = \{L_i\}_{i=1}^{g} \in \mathcal{R}^{P}$ be the labels of the gallery image sets, where $L_i = \{l_j\}_{j=1}^{p_i} \in \mathcal{R}^{p_i}$ are the labels for each image set such that all images in the same set get the same label. Let $L_q = \{l_j\}_{j=1}^{p_q} \in \mathcal{R}^{p_q}$ be the label for the query image set. We append dictionaries D_G and D_q in an array $D_s = [D_G|D_q] \in \mathcal{R}^{m \times (P+p_q)}$ as well as their labels $L_s = [L_G|L_q] \in \mathcal{R}^{P+p_q}$. However, the labels are not used at this stage.

As a similarity measure, we compute the sparse coefficients required to represent a particular dictionary atom as a linear combination of the remaining atoms similar to [4]. More precisely, we take one atom $d_i \in \mathcal{R}^m$ out of D_s and represent it as a sparse linear combination of the remaining dictionary $D'_s = \{D_s\} - \{d_i\} + \{0_i\} \in \mathcal{R}^{m \times (P+p_q)}$. In D'_s, 0_i is the column of zeros placed as column i to maintain the original matrix size. We use a fast implementation of LARS [3] to find the sparse coding coefficients. The sparse coding coefficients α_i of d_i are computed as

$$\min_{\alpha_i} ||d_i - D'_s\alpha_i||_2^2 \ \ s.t. \ \ ||\alpha_i||_1 \leq \lambda \ . \tag{3}$$

We append all α_i as columns in the similarity matrix $S = \{\alpha_i\}_{i=1}^{P+p_q} \in \mathcal{R}^{(P+p_q)\times(P+p_q)}$ and apply spectral clustering to group dictionary atoms into clusters.

2.4 Graph Laplacian Computation

We can consider each dictionary atom in D_s as a node in a fully connected graph G, and the sparse linear coefficients given in S as the edge weights connecting any two nodes in G. Thus the similarity matrix S forms an adjacency matrix for G, which is a directed graph because in general $S(i,j) \neq S(j,i)$. As some of the coefficients in S may be negative, we take the absolute of all values and to make the graph G undirected we add the edge weight $S(i,j)$ with the edge weight $S(j,i)$. The modified adjacency matrix for the resulting undirected graph having all positive weights is given by

$$A = abs(S) + abs(S^T) \; . \tag{4}$$

In order to apply spectral clustering [9] on the graph represented by matrix A, we first compute the degree matrix of the graph, which is a diagonal matrix containing the degree of each vortex i at the diagonal position (i,i). The degree matrix D is given by

$$D(i,j) = \begin{cases} \sum_{i=1}^{P+p_q} A(i,j) & \text{if } i = j \\ 0 & \text{if } i \neq j \; . \end{cases}$$

Next, using D and A, we compute the un-normalized graph Laplacian matrix by simple subtraction, $L = D - A$. The graph Laplacian is useful for clustering graph A into different partitions by computing the eigenvectors of L. We also use normalized graph Laplacian as recommended by [17]:

$$L_w = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \tag{5}$$

2.5 Graph Partitioning

We partition the graph represented by the adjacency matrix A into two disjoint partitions A_L and A_R, such that the sum of edge weights across the cut is minimum. The loss function may be written as $\sum_{i\in|A_L|, j\in|A_R|} A(i,j)$, where $|A_L|$ and $|A_R|$ represent the number of nodes in each partition. It is easy to implement the minCut algorithm, however, it may yield unbalanced partitions. In the extreme case, a simple minCut implementation may separate only one node from the rest of the graph. In order to ensure that both partitions are balanced, we minimize the normalized cut *NCut* objective function [17]

$$\frac{1}{V_{A_L}} \sum_{i\in|A_L|, j\in|A_R|} A(i,j) + \frac{1}{V_{A_R}} \sum_{i\in|A_L|, j\in|A_R|} A(i,j), \tag{6}$$

where V_{A_L} is the sum of all edge weights attached to the vertices in A_L. As the number of nodes starts reducing in any partition, the objective function starts increasing.

Unfortunately, the minCut using the *NCut* objective function turns out to be an NP hard problem and only approximate solutions can be computed by using spectral clustering. It has been shown in [17] that the second eigenvector of Lw may provide an approximate solution to a relaxed *NCut* problem. The second eigenvector of L_w may be used to partition the graph by using the simple criteria

$$\begin{cases} v_i \in A_L & \text{if } e_2(i) \geq 0 \\ v_i \in A_R & \text{if } e_2(i) < 0 \end{cases}$$

where e_2 is the second eigenvector of L_w. The higher order eigenvectors of L_w may also be used for further partitioning the graph into smaller clusters. However, the approximation error gets accumulated and the clustering performance degrades. It is, therefore, recommended to recompute L_w for each partition and repeat the same process hierarchically to get smaller spectral clusters [17].

We recursively perform binary partitioning of the graph G based on *NCut* until we get only decisive clusters. Decisive clusters are either critical or non-critical. Critical clusters contain atoms from only one gallery set along with query atoms. Non-critical clusters either contain only query atoms or no query atom.

2.6 Set Based Similarity Computation

Let the number of critical clusters be n_c, and $S_q \in \mathcal{R}^{n_c}$ contain the number of query atoms in each n_c. Let $S_G \in \mathcal{R}^{n_c}$ represent the number of gallery atoms in each cluster and $L_G \in \mathcal{R}^{n_c}$ represent the label of the gallery atoms in each cluster. Using these three arrays, we compute two more, gallery count array $G_c \in \mathcal{R}^g$ and query count array $Q_c \in \mathcal{R}^g$. In G_c and Q_c, each index represents one gallery image set label. Each value in G_c represents the number of atoms of that label which are found in critical clusters and each value in Q_c represents the number of query atoms in critical clusters accompanied by any atom of that particular label. We pointwise multiply these two arrays and get the final similarity score of query set with each gallery set as

$$\rho(i) = G_c(i)Q_c(i)\ , \tag{7}$$

where $\rho \in \mathcal{R}^g$ represents the similarity between the query and each gallery set. The label assigned to the query set is the index of ρ exhibiting maximum similarity $\rho^1_{max} = \max(\rho)$

$$L_q = i\ , \quad \text{s.t.} \quad \rho(i) == \rho^1_{max}\ . \tag{8}$$

From the sparseness of the similarity array ρ, we compute the match confidence. We observe that for larger sized dictionaries, if there is only one peak in ρ with maximum energy E_{max}, the confidence index is 100%. Such matches are always correct. The maximum energy peak at $\rho(i)$ will be $E_{max} = p_i p_q$, where p_i is the number of dictionary atoms in the i^{th} gallery image set and p_q is the number of dictionary atoms in the query image set. A measure of peak energy is $\eta = \rho^1_{max}/E_{max}$. In many cases, the ρ array may not be very sparse and may contain more than one peaks. We consider the maximum peak ρ^1_{max} as the signal and the second maximum peak ρ^2_{max} as the noise. The signal to noise ratio (SNR) must be high for a good match: $SNR = \rho^1_{max}/\rho^2_{max}$. We combine both η and SNR in one measure

$$\zeta = \frac{(\rho^1_{max} - \rho^2_{max})}{E_{max}}\ , \tag{9}$$

where ζ is the confidence. If ζ is high, we stop the process and accept the label with 100% confidence. Otherwise, we repeat the whole process with the next higher dictionary size. Easy cases are identified at smaller dictionary sizes. However, challenging ones may never get high confidence. In this case, we stop beyond a certain dictionary size and assign the query set the most repeated label.

3 Experimental Validation

We evaluate the proposed HSSC algorithm on the Honda/UCSD [7] and CMU Mobo [14] databases. In both cases, faces are detected using [13], cropped and converted to gray scale.

Figure 3: Example images from (a) Honda/UCSD dataset. (b) CMU Mobo dataset. Each row is selected from one different image set.

For Honda database, we use 20x20 face images (similar to [16]) after histogram equalization. The number of images per set varies from 17 to 645. For CMU Mobo dataset, face images are resized to 40×40 (similar to [5]) and their LBP features [18] are used.

In both cases, we project data on a 200 dimensional PCA space computed from only the gallery image sets. For dictionary learning, we use the open source sparse modeling software (SPAMS) [11] to solve Equation (1) with 200 iterations and $\lambda = 0.15$. The size of dictionary is varied from 6 to 21 atoms for each gallery set as well as for the query set with an increment of 1 atom. Recall that the gallery dictionaries are learned off line and only the probe dictionaries are learned online.

The confidence can either be high or low and its threshold varies with the dictionary size p_i. For a high confidence, one of the following conditions must be met: (1) $p_i \leq 10$ and $\zeta = 1$, (2) $10 < p_i \leq 15$ and $\zeta \geq 0.8$, (3) $15 < p_i \leq 20$ and $\zeta \geq 0.7$, (4) $p_i \geq 21$ and $\zeta \geq 0.6$. If none of these conditions is satisfied, but a particular test set gets the same label for over half the number of iterations, the confidence of that label is also assumed to be high.

We compare the proposed HSSC algorithm with existing image set classification methods including Discriminant Canonical Correlation (DCC) [20], Manifold to Manifold Distance (MMD) [16], Manifold Discriminant Analysis (MDA) [15], linear Affine Hull based Image Set Distance (AHISD) [5] and linear Convex Hull based Image Set Distance (CHISD) [5]. The parameters of all methods are carefully optimized. For DCC, the embedding space dimension is set to 100, the subspace dimensionality is set to 10 and set similarity is computed from the 10 maximum correlations. For MMD and MDA, the parameters are selected as suggested in [16] and [15].

3.1 Results on the Honda/UCSD Dataset

This dataset contains 59 videos of 20 subjects with varying poses and expressions. Sample images are given in Fig.3. Our experiments are based on the configuration proposed by [7]. We randomly selected one image set per subject as gallery and the remaining 39 sets were used as query sets. Identification rates of different techniques are summarized in Table 1. HSSC outperforms all other techniques and achieves 100% accuracy. Although correctly classified, only 15.38% matches were with high confidence and 84.61% with low confidence. Note that our results for CHISD in Table 1 are lower compared to [5] because we use smaller image sizes i.e. 20x20 compared to 40x40 in [5].

3.2 Results on the CMU Mobo Dataset

CMU Mobo dataset contains 96 video sequences of 24 subjects walking on a treadmill. For each subject there are four sequences for four different walking patterns, slow, fast, inclined, and ball carrying. Each sequence was captured by a different camera. Image features con-

Table 1: Identification Rate comparison of various techniques on Honda/UCSD dataset

Techniques	Performance
DCC [20]	94.87%
MMD [16]	94.87%
MDA [15]	97.44%
AHISD [5]	89.74%
CHISD [5]	92.31%
HSSC	100.00%

Table 2: Average identification rates and standard deviations on the CMU Mobo dataset

Technique	Average Performance $\pm$ STD
DCC [20]	$91.53 \pm 1.66\%$
MMD [16]	$89.72 \pm 3.48\%$
MDA [15]	$95.97 \pm 1.90\%$
AHISD [5]	$94.58 \pm 2.57\%$
CHISD [5]	$96.52 \pm 1.18\%$
HSSC	$96.67 \pm 1.87\%$

sist of uniform LBP histograms using circular (8, 1) neighborhoods extracted from 8×8 gray scale patches. We randomly select one image set for each subject as training and the remaining three as testing. We perform a 10-fold experiment by repeating the random selection 10 times and average the recognition rates. Results are reported in Table 2. The proposed HSSC algorithm outperforms all methods. In this case, the HSSC algorithm identifies 95.14% matches with high confidence and only 4.86% matches with low confidence. The identification rate for only high confidence matches is $98.34\% \pm 1.59$.

Note that the performance of CHISD [5] is very close to HSSC. However, CHISD is a sample based classification technique and is not robust to outliers while, HSSC is a structure based technique and more robust to outliers as demonstrated in our next experiment.

3.3 Robustness to Outliers

We have performed two different robustness experiments on the Honda data set. In the first experiment, in each of the probe set we added randomly selected n_r images from a randomly selected gallery set and the value of n_r is varied from 1 to 12. The accuracy of HSSC remained 100% for $n_r \leq 11$ and 97.44% for $n_r = 12$. In the second experiment, in each of the probe set we added $n_r \times g$ images where $g = 20$ is the gallery size and n_r images are taken from each gallery and added to probe sets. n_r is varied from 1, 2 and 3 and the actual number of outliers added to each probe set are 19, 38 and 57. The accuracy of HSSC algorithm has remained 100%, 97.44% and 92.30% respectively for the three cases. These experiments demonstrate that HSSC is a significantly robust algorithm.

4 Conclusion

We presented a hierarchical sparse spectral clustering algorithm for image set classification. The proposed algorithm performs unsupervised hierarchical clustering guided by the probe set which is more accurate and robust to outliers. Experiments on two benchmark datasets show that the proposed algorithm outperforms existing techniques. We also demonstrated the robustness of our algorithm to outliers in the query set.

5 Acknowledgements

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

References

[1] A. W. Fitzgibbon and A. Zisserman. Joint Manifold Distance: A New Approach to Appearance based Clustering. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 26–33, 2003.

[2] M. Aharon, M. Elad, and A.M. Bruckstein. The k-svd: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[3] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[4] Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *CVPR*, pages 2790–2797. IEEE, 2009.

[5] H. Cevikalp and B. Triggs. Face Recognition Based on Image Sets. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 2567–2573, 2010.

[6] J. Weng, C. H. Evans and W.-S. Hwang. An Incremental Learning Method for Face Recognition under Continuous Video Stream. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pages 251–256, 2000.

[7] K.-C. Lee, J. Ho, M.-H. Yang and D. Kriegman. Video-Based Face Recognition Using Probabilistic Appearance Manifolds. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 313–320, 2003.

[8] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient sparse coding algorithms. In *In NIPS*, pages 801–808. NIPS, 2007.

[9] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, December 2007. ISSN 0960-3174.

[10] M. Nishiyama, M. Yuasa, T. Shibata, T. Wakasugi, T. Kawahara and O. Yamaguchi. Recognizing Faces of Moving People by Hierarchical Image-Set Matching. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[11] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *ICML*, ICML '09, pages 689–696, New York, NY, USA, 2009. ACM.

[12] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla and T. Darrell. Face Recognition with Image Sets Using Manifold Density Divergence. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 581–588, 2005.

[13] P. Viola and M. J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

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

[15] R. Wang and X. Chen. Manifold Discriminant Analysis. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 429–436, 2009.

[16] R. Wang, S. Shan, X. Chen and W. Gao. Manifold-Manifold Distance with Application to Face Recognition based on Image Set. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2008.

[17] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8):888–905, 2000.

[18] T. Ahonen, A. Hadid and M. Pietikainen. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[19] T.-K. Kim, J. Kittler and R. Cipolla. Incremental Learning of Locally Orthogonal Subspaces for Set-based Object Recognition. In *Proc. British Machine Vision Conf.*, pages 559–568, 2006.

[20] T.-K. Kim, O. Arandjelovic and R. Cipolla. Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(6):1005–1018, 2007.

[21] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[22] Yiqun Hu and Ajmal S. Mian and Robyn Owens. Sparse Approximated Nearest Points for Image Set Classification. In *CVPR*, pages 121–128, June 2011.