Hierarchical Sparse Spectral Clustering for Image Set Classification

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

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

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.

A schematic diagram of the proposed algorithm is shown in Fig. 1. The intrinsic data dimensionality in the image sets is often less than the apparent dimensions. Therefore, we reduce the data dimensionality using PCA basis computed from the training (gallery) sets. For each reduced dimensionality gallery set, we pre-compute sparse dictionaries of varying sizes. 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 a cost function $\frac{1}{n}\sum_{j=1}^{n_i} f(x_j, D_i)$. Each column of D_i represents a basis vector for the image set X_i. We use the convex ℓ_1 formulation of the Lasso as the cost function [3]

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

Sparse dictionaries D_i of various sizes for each of the training set are learned from (1).

We start from the smallest size dictionary with p_i atoms per gallery set. A sparse dictionary with p_q atoms is learned for the query (probe) image set as well. Let D_G be the set of learned dictionaries for the gallery image sets and D_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 be the labels of the gallery image sets and L_q be the labels for the query image set. We append dictionaries in an array $D_s = [D_G | D_q]$ and the labels as well $L_s = [L_G | L_q]$.

As a similarity measure, we compute the sparse coefficients required to represent a particular dictionary atom as a linear combination of the remaining atoms [1] in D_s. We take one atom d_i out of D_s, replace it by zeros, and represent d_i as a sparse linear combination of the remaining atoms. We use a fast implementation of LARS to find the sparse coding coefficients α_i of d_i computed as

$$\min_{\alpha_i} ||d_i - D_s \alpha_i||_2^2 \ \ s.t. \ \ ||\alpha_i||_1 \leq \lambda \ . \tag{2}$$

We append all α_i as columns in a similarity matrix $S = \{\alpha_i\}_{i=1}^{P+p_q}$.

Considering each dictionary atom in D_s as a node in a fully connected graph G, the sparse linear coefficients in S are 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. To form a positive weight undirected graph, the modified adjacency matrix is computed as $A = |S| + |S^t|$, where $|\cdot|$ stands for absolute value. In order to apply spectral clustering [4] on A, we first compute the degree matrix $D(i,j) = \sum_{i=1}^{P+p_q} A(i,j)$ if $i = j$ and $D(i,j) = 0$ if $i \neq j$. Using D and A, we compute un-normalized graph Laplacian matrix $L = D - A$ and then the normalized Laplacian [6]

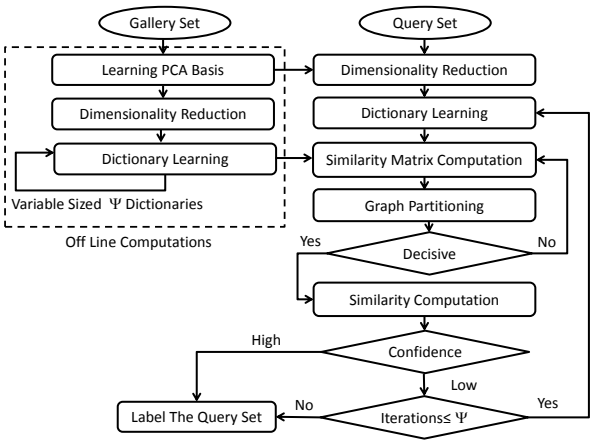

Figure 1: Block diagram of the Hierarchical Sparse Spectral Clustering (HSSC) algorithm.

$L_w = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$. To cut the graph into two balanced partitions, we compute the eigen vectors of L_w. Using the sign of the elements of the second eigenvector we divide the set of all dictionary atoms into two partitions.

We recursively perform binary partitioning of the graph G until each partition is identified as a decisive cluster which may contain atoms from only one gallery set along with query atoms (critical cluster), only query atoms (non-critical cluster) or zero query atom and one or more gallery atoms (non-critical cluster). For each gallery set, we count the number of atoms in all critical clusters and the corresponding number of query atoms in those clusters as well. The product of both counts represents a similarity score of the query set with that particular gallery set. Based on the distribution of query atoms in the critical clusters, a confidence score is also defined. If the confidence is high, the algorithm stops and a label is predicted for the query set based on the maximum similarity score. If the confidence is low, the full process is repeated with an increased dictionary size. If confidence remains low for consecutive dictionary sizes, however the predicted query label remains consistent, that label will soon accumulate high confidence and the algorithm will stop. If confidence remains low over a number of dictionary sizes and the predicted label is inconsistent, the algorithm will stop after executing the maximum number of iterations and the the final label will be predicted as the label with the maximum mode over all iterations. This may occur in the case of difficult matches and the predicted label confidence will remain low.

Experiments are performed on the Honda/UCSD [2] and CMU Mobo data [5] 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 11 imposters selected randomly from a random gallery set and mixed with the probe sets.

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

[2] K. C. Lee, J. Ho, M. H. Yang and D. Kriegman. Video-Based Face Recognition Using Probabilistic Appearance Manifolds. In *CVPR*, pages 313–320, 2003.

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

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

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

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