Joint Optimization of Manifold Learning and Sparse Representations

Raymond Ptucha
Department of Computer and Information Systems
Rochester Institute of Technology
Rochester, New York 14623, USA
rwpeec@rit.edu

Andreas Savakis
Department of Computer Engineering
Rochester Institute of Technology
Rochester, New York 14623, USA
andreas.savakis@rit.edu

Abstract—**Dimensionality reduction via manifold learning offers an elegant representation of data whereby the high dimensional feature space is parameterized by a lower dimensional space where the data resides. Sparse representations efficiently represent test patterns by sparse linear coefficients from a dictionary of training exemplars. Sparse representations have been adopted for classification purposes, but the resulting classifiers may have to deal with data in high dimensions and large dictionaries. This paper analyzes the interaction between dimensionality reduction and sparse representations. The proposed technique, called K-LGE, presents a unified framework which utilizes a semi-supervised variant of Linear extension of Graph Embedding with K-SVD dictionary learning. An iterative procedure optimizes the dimensionality reduction matrix, sparse representation dictionary, sparse coefficients, and linear classifier. Results are demonstrated in a wide variety of facial and activity recognition problems to demonstrate the robustness of our proposed method.**

Keywords- dimensionality reduction; manifold learning; sparse representation; facial analysis; activity recognition.

I. INTRODUCTION

Given n data samples, x_1, x_2, …x_n, each sample $x_i \in \mathbf{R}^D$, stored in matrix X, $X \in \mathbf{R}^{Dxn}$ and $D < n$, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are effective techniques for obtaining a lower dimensional representation of X. During PCA or LDA, the top d eigenvectors are used in projection matrix U such that the low dimensional representation of X is $Y = X^T U$, $Y \in \mathbf{R}^{nxd}$. Although these linear dimensionality reduction techniques produce meaningful results, we wish to find an alternate low dimension d such that $d << D$. Further, the underlying linearity assumption of PCA and LDA may be limiting when modeling the behavior of complex imagery such as face representations.

Manifold learning techniques reduce the dimensionality of input data by identifying a non-linear lower dimensional space where the data resides [1, 2]. Methods include Isomap [3] and Locally Linear Embedding (LLE) [4]. In order to support the extension of the manifold model to new examples, linearized techniques called Linear extension of Graph Embedding (LGE) [5], solve a linear approximation of the non-linear object. The dimensionality reduction offered by the LGE techniques generally affords $d << D$.

The notion of Sparse Representations (SRs), or finding sparse solutions to underdetermined systems, has found applications in a variety of scientific fields. The resulting sparse models are similar in nature to the network of neurons in V1, the first layer of the visual cortex in the human, and more generally, the mammalian brain [6, 7].

In SR systems, images x_i are efficiently represented by sparse linear coefficients from a dictionary Φ of overcomplete basis functions, where $\Phi \in \mathbf{R}^{Dxn}$. SR solves for coefficients $a \in \mathbf{R}^n$ that satisfy the ℓ^1 minimization problem $x' = \Phi a$. It has been shown that under typical conditions, the minimal solution is the sparsest one [8, 9]. There have been several studies optimizing both the ℓ^1 minimization [10, 11] as well as the selection of dictionary elements [12, 13]. In our work, we construct Φ from low dimensional samples Y, $\Phi \in \mathbf{R}^{dxn}$.

Although the SR framework is designed for reconstruction purposes, it has been adapted successfully for classification problems. Wright *et al.* [14, 15] achieved state-of-the-art facial recognition results by feeding the a coefficients directly to a classifier. In this framework, the dominant signal always prevails, but it could produce some unintended effects. For example, when trying to extract facial identity, pose variation may contaminate or even dominate the sparse coefficients. This coefficient contamination is unfortunate yet important, as it has been shown that images of a single person under multiple poses exhibit greater variation than images of different people at a single pose [16].

Tzimiropoulos *et al.* [17] demonstrated computational and accuracy improvements by doing the ℓ^1 optimization in a dimensionally reduced space. Zafeiriou and Petrou [18] used Principal Component Analysis (PCA) and SR techniques based on [15] for facial expression recognition. The work in [18] struggled with coefficient contamination, noting that applying Wright's framework is not a straightforward process. Ptucha *et al.* [19] addressed the coefficient contamination problem by preprocessing the data with supervised manifold learning. Similarly to subspace clustering [20], supervision in manifold learning encourages clustering of sample images in accordance with their classification labels.

By preprocessing SR classification with manifold learning, we are able to achieve superior classification results as well as faster runtimes. Research in manifold learning has influenced the SR community and vice-versa. Sparsity Preserving Projections [21] replaces the adjacency matrix used in LGE techniques with SR sparse coefficients. Discriminative Sparse Coding [22] uses sparse coefficients in an LDA framework. Graph Regularized Sparse Coding [23] adds the LGE objective function on sparse coefficients to the traditional ℓ^1 sparse objective function as it jointly learns the sparse coefficients and dictionary terms.

In this paper, we employ the SR concept in a dimensionality reduced space, obtained by a semi-supervised variant of Linear extension of Graph Embedding with K-SVD dictionary learning, and iterate over this space to minimize reconstruction errors. To the best of our knowledge, this is the first paper which jointly optimizes dimensionality reduction matrix U, with ℓ^1 dictionary Φ, and ℓ^1 sparse coefficients a. We contrast our technique, which we call K-LGE, to other recently introduced techniques across a wide variety of facial and activity classification problems.

The rest of this paper is organized as follows. Sections 2 and 3 introduce the necessary principles of manifold learning and SR concepts. Section 4 describes how to jointly optimize manifold learning and SRs. Section 5 presents experimental results. Section 6 summarizes with conclusions.

II. Manifold learning

A. Dimensionality Reduction

Complex objects often necessitate representations of high dimensionality. The features used in facial understanding problems vary from processed image pixels to SIFT descriptor points, Gabor jets, and concatenated block histograms. This high dimensional feature space is not only inefficient and computationally intensive, but the sheer number of dimensions often masks the discriminative signal embedded in the data.

For samples $x_i \in \mathbf{R}^D$ we seek a lower dimensional representation yielding $y_i \in \mathbf{R}^d$, where $d << D$. For linear models, e.g. PCA or LDA, $y_i = x_i^T U$, where U is a $D \times d$ projection matrix. Alternatively, the high dimensional feature space can be parameterized by a lower dimensional embedded manifold discovered using manifold learning [1, 2]. In addition to being more compact, the resulting lower dimensional manifold representation is more discriminative and thus more appropriate for subsequent classification.

B. Linear extension of Graph Embedding (LGE)

During manifold learning a fully connected graph of the input space is constructed, where each of the n input samples or nodes is connected to all other (n-1) input samples with a weight, $0 \leq w_{ij} \leq 1$, $i,j = 1\ldots n$. The resulting connection matrix W is called the adjacency matrix and the connections or weights w_{ij} can be solved several ways. For example, w_{ij} is set to 1 if x_i is amongst the z nearest neighbors of x_j, 0 otherwise. Alternatively, w_{ij} is set to 1 if $\|x_i - x_j\| < \varepsilon$, and 0 otherwise. Allowing w_{ij} to take on continuous values between 0 and 1 offers more control in describing sample connections.

The goal of graph embedding is to preserve the similarities amongst neighbors in both high and low dimensional space. The optimal Y is found by minimizing:

$$\sum_{i,j} (y_i - y_j)^2 w_{ij} \tag{1}$$

As such, if neighbors y_i and y_j have a strong connection w_{ij}, their Euclidean distance should be minimal. W is defined similarly for X and Y, such that if neighbors x_i and x_j are close, y_i and y_j are also close. LGE seeks a linear approximation to this nonlinear concept of the form $y_i = x^T_i U$ or $Y = X^T U$. We define D as a diagonal matrix of the column sums of W, $D_{ii} = \Sigma_j w_{ij}$; and L is the Laplacian matrix, $L = D - W$. After simplification, this problem reduces to:

$$\hat{U} = \min \frac{U^T X L X^T U}{U^T X D X^T U} \tag{2}$$

The optimal U is given by the minimum eigenvalue of the generalized eigenvector problem:

$$X L X^T U = \lambda X D X^T U \tag{3}$$

where U is the resulting projection matrix.

Different choices of W yield a multitude of dimensionality reduction techniques such as LDA, Locality Preserving Projections (LPP) [24], and Neighborhood Preserving Embedding (NPE) [25]. For each approach, W is initialized to all zeros, and then connected w_{ij} entries are set as follows.

For LDA, nodes i and j are connected if they are from the same class. Connected w_{ij} entries are set to $1/k_n$, where k_n is the number of samples per their shared class:

$$w_{ij} = 1/k_n \tag{4}$$

For LPP, if nodes i and j are connected, then:

$$w_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}} \tag{5}$$

For NPE, if nodes i and j are connected, we solve the following objective function for element i in local reconstruction matrix $M \in \mathbf{R}^{nxn}$ as a function of z nearest neighbors of x_i, $N_z(x_i)$:

$$min \left\| x_i - \sum_{j \in N_{z(x_i)}} M_{ij} x_j \right\|^2, \sum_{j \in N_{z(x_i)}} M_{ij} = 1 \tag{6}$$

Then let:

$$W = M + M^T - M^T M \tag{7}$$

Both LPP and NPE can be used in supervised mode by defining connected neighbors as those which share similar class labels.

III. Sparse signalrepresentation

A. Sparse Representations

A natural way to represent a low dimensional sample $y \in \mathbf{R}^d$ from a training dictionary $\Phi \in \mathbf{R}^{dxn}$ is by solving $\hat{y} = \Phi a$, where $a \in \mathbf{R}^n$ is the weight of each training exemplar in the dictionary Φ. However, in most practical cases, the system has either no solution or multiple solutions. For sparse signals, the objective of SRs is to identify the smallest number of nonzero coefficients $a \in \mathbf{R}^n$ such that $\hat{y} = \Phi a$.

A convex relaxation approach was introduced by Donoho *et al.* [8] and Candes *et al.* [9], where it was shown that under certain constraints, such as the sparsity of the representation, the minimal solution is equivalent to the solution of the following Lasso regression problem in statistics:

$$\hat{a} = \arg\min \|a\|_1 \quad \text{s.t. } \hat{y} = \Phi a \tag{8}$$

where $\|a\|_1 = \Sigma\ |a|$. The benefit of using the ℓ^1 minimization is that the problem can be efficiently solved using convex optimization algorithms. When noise is present in the signal, a perfect reconstruction is typically not feasible. Therefore, we require that the reconstruction be within an error tolerance. This optimization, called Basis Pursuit Denoising (BPDN), reformulates (8) as:

$$\hat{a} = \min\|a\|_1 \quad s.t. \|\hat{y} - \Phi a\|_2 \leq \epsilon \tag{9}$$

Often (9) is approximated by loosening the error constraints and reconfigured to specifically include a regularization term, λ which encourages sparseness by incurring a penalty on the resulting coefficients:

$$\hat{a} = \min\{\|\hat{y} - \Phi a\|_2^2 + \lambda\|a\|_1\} \tag{10}$$

Perhaps the most widely used method to solve the ℓ^1 minimization of (9) or (10) is Orthogonal Matching Pursuit (OMP) [26]. OMP iteratively selects one dictionary element at a time in a greedy fashion, minimizing a residual reconstruction error at each step. Given the SR coefficients $\hat{a}$ of a test image using the dictionary Φ, a reconstruction error method estimates the class k^* of a query sample y. Given k classes, the reconstructed sample using sparse coefficients a from all classes is compared to the reconstructed sample using coefficients a^i from each respective class:

$$k^* = \min_{i=1:k}\left\|y - \Phi a^{\mathrm{i}}\right\|_2 \tag{11}$$

When constructing Φ the goal is to generate an over-complete dictionary with more samples than dimensions per sample. This allows the necessary degrees of freedom to choose the sparsest solution and produces smooth and graceful coefficient activity across diverse test samples [27]. For efficiency, it is desirable to have a dictionary of small size, necessitating linearly independent or decorrelated samples.

K-SVD [28] was introduced as a means to learn an over-complete but small dictionary. K-SVD is an iterative technique, where at each iteration, training samples are first sparsely coded using the current dictionary estimate, and then dictionary elements are updated one at a time while keeping others fixed. Each new dictionary element is a linear combination of training samples. Rubinstein [29] implemented an efficient implementation of K-SVD using Batch Orthogonal Matching Pursuit.

The works of [30, 31] jointly optimize dictionary learning and classifier training to select exemplars that minimize both reconstructive and discriminative errors. Jiang *et al.* [12] devised efficient methods for choosing Φ from a set of training exemplars by minimizing both reconstruction and classification errors in an optimal fashion. The work in [12] encourages input samples from the same class to have similar sparse codes.

B. Dimensionality Reduction and Sparse Representations

Although methods for populating the adjacency matrix W vary, sparseness is one common characteristic across all techniques. Sparsity Preserving Projections (SPP) [21] replaces the neighbor coefficients in row i of matrix M from (6) with sparse coefficients $\hat{a}$ corresponding to sample x_i. Global Sparse Representation Projections [32] modifies the dimensionality reduction function in SPP to simultaneously maximize supervised class separability and minimize sparse representation error. [22] uses the sparse coefficients to populate matrix W, then adds supervised similarity and dissimilarity matrices akin to LDA. [23] replaces the y terms in (1) with coefficients $\hat{a}$, claiming that nearby samples should have similar coefficients.

Each of the above methods introduces a new dimensionality reduction technique or a new SR technique. What lacks is a single, unified method that optimizes dimensionality reduction projection matrix U with dictionary Φ, and coefficients $\hat{a}$. In the next section we present such a method, which we call K-LGE for K-SVD with Linear extension of Graph Embedding.

IV. The K-LGE method

We wish to combine the dimensionality reduction matrix U from (2) with a method to learn a dictionary Φ and sparse coefficients a. K-SVD solves:

$$< \hat{\Phi}, \hat{a} >= \min\|x - \Phi a\|_2^2 \quad s.t. \|a\|_0 \leq T \tag{12}$$

Combining (2) with (12), we get:

$$< \hat{U}, \widehat{\Phi}, \hat{a} \geq \min\|X^T U - \Phi a\|_2^2 + \frac{U^T X L X^T U}{U^T X D X^T U} \quad s.t. \|a\|_0 \leq T \tag{13}$$

The first term performs K-SVD optimization in low dimensional space, and the second term is the LGE dimensionality reduction objective function. Since we are solving a classification problem, we need to convert test sample coefficients a into a classification label estimate. Because dictionary elements from K-SVD are a linear combination of input samples, we cannot use the minimum reconstruction error in (11). Instead, we shall perform classification with coefficient transformation matrix C, $C \in \mathbf{R}^{mxk}$, where k is the number of classes and m is the number of dictionary elements.

We define H as a sparse ground truth matrix, $H \in \mathbf{R}^{kxn}$. Each column of H corresponds to a training sample, where the k^{th} element is set to 1 if y_i belongs to class k, 0 otherwise. Coefficients a from each training example are stored into matrix A, $A \in \mathbf{R}^{mxn}$. This problem is formulated as:

$$\hat{C} = \min\|H - C^T A\|_2^2 \tag{14}$$

Which can be solved directly via ridge regression:

$$C = (AA^T)^{-1} AH^T \tag{15}$$

A. Training Procedure for K-LGE

Equation (13) is neither directly solvable nor convex. Using K-SVD with n training samples, we learn a dictionary Φ of m atoms, where $m \leq n$ via an implicit transformation T, $T \in \mathbf{R}^{mxn}$, resulting in $\Phi = TY = TX^T U$. As such, the dictionary transformation function T and the dimensionality reduction transformation function U will oscillate if we indiscriminately iterate one after the other.

For smooth and reliable results, we desire an overcomplete dictionary in which the number of samples, m is greater than the number of dimensions d. T has rank $\leq m$, U has rank $\leq d$. Since $m \geq d$, T in general has more degrees of freedom and it is preferable to iterate on T more often than U. As we minimize reconstruction errors in (13), coefficients a offer a more accurate representation of X and lower classification errors.

In [19] it was sown that supervised dimensionality reduction minimizes SR coefficient contamination by enforcing class separation. A discriminative dictionary was utilized in [12, 21, 22, 32], but we find similar results at much faster runtimes by doing class discrimination and dimensionality reduction prior to dictionary learning. The initial value of U is solved via supervised LGE. Subsequent updates of U need to be done in context of the current dictionary Φ and training sample coefficients A. This update problem is formulated as:

$$\widehat{U} = \min \|X^T U - A^T \Phi^T\|_2^2 \qquad (16)$$

Which can be solved directly:

$$U = (XX^T)^{-1} X A^T \Phi^T \qquad (17)$$

Fig. 1 summarizes the training procedure.

```
WHILE $\varepsilon$ has not converged or $\varepsilon > \tau$
    IF firstIteration
        1a. Calculate $U$ using LGE.
    ELSE
        1b. Calculate $U$ using (17).
    ENDIF
    2. Calculate low dimensional samples $Y= X^T U$.
    3. Initialize the $m$ samples of $\Phi$ randomly from the $n$
       low dimensional training samples.
    4. Calculate $\{A, \Phi\}$ using K-SVD, substituting $Y$ for $X$.
    5. Calculate $C$ using (15).
    6. Calculate verification set error, $\varepsilon = \|H - C^T A\|_2^2$.
ENDWHILE
```

Figure 1. Training procedure for K-LGE.

The choice of LGE in Step 1a of the training procedure should be a discriminative embedding which maintains input topology. The best approach we have found uses a convex combination of supervised and unsupervised adjacency matrices W_{LDA} and $W_{Gaussian}$ corresponding to (4) and (5) respectively. The two are combined into a single W:

$$W = \alpha W_{LDA} + (1 - \alpha) W_{Gaussian} \qquad (18)$$

For posed datasets which are linearly separated, W_{LDA} should be weighted higher. For natural datasets or classification problems in which the number of classes is small, we emphasize the addition of $W_{Gaussian}$. Classification problems with small number of classes reduce the rank of W_{LDA} to (k-1). For example, to determine gender from facial images, W_{LDA} would restrict U to a D×1 projection matrix, where D is the number of image pixels in each face exemplar.

B. Testing Procedure for K-LGE

With training complete, given a test sample x, along with U, Φ, and C, Fig. 2 summarizes the testing procedure:

```
1. Calculate low dimensional sample $y= x^T U$.
2. Calculate sparse coefficients $a$ using (10).
3. Use $C$ along with $a$ to estimate class label vector
   $l \in \mathbf{R}^{kx1}$ where the maximum value of $l$ is used as a
   class predictor, and other $l$ values provide confidence
   values as to which class $x$ belongs.
            $\hat{l} = \max_{i=1:k}(l = C^T a_i)$        (19)
```

Figure 2. Testing procedure for K-LGE.

V. Experiments

We evaluate our K-LGE approach on four public databases: the extended Cohn-Kanade (CK+) facial expression dataset [33], the extended Yale B facial recognition database [34], the Facial Expression Recognition and Analysis Challenge (FERA2011) GEMEP-FERA [35] dataset, and the i3DPost multi-view activity recognition dataset [36]. We test each dataset across three categories of 1) dimensionality reduction; 2) sparse representation; and 3) combined techniques. The dimensionality reduction techniques include PCA, LDA, LPP [24], NPE [25], and Sparsity Preserving Projections (SPP) [21]. The sparse representation methods include K-SVD [28], LC-KSVD1 and LC-KSVD2 [12]. The combined methods include Sparse Representation-based Classification (SRC) [15], Manifold based Sparse Representation (MSR) [19], and our proposed K-LGE method.

A. Testing Datasets

The CK+ [33] expression dataset contains 118 subjects in 327 sequences exhibiting the expressions of anger, disgust, fear, happiness, sadness, surprise, and contempt. An Active Appearance Model (AAM) automatically localizes 68 points on the face. To contrast a low dimensional representation of the face vs. a higher dimensional representation, the AAM eye and mouth corner points are used to define an affine warp to a canonical face of 60x51 pixels. As such, from this dataset we compare two variants: D=68x2=136 (AAM point based), and D=60x51=3060 (pixel based). Each has 164 training and 163 testing faces (chosen randomly), and the K-SVD methods use a dictionary size of 63 elements.

The Extended YaleB facial recognition dataset contains 2,414 frontal images of 38 people under varying illumination and facial expression. Each face is 192x168 pixels which are reduced to D=504 via random projections following [15]. The test set contains 1216 training faces and 1198 testing faces. The K-SVD methods use a dictionary size of 570 elements.

The GEMEP-FERA temporal expression dataset contains 155 training and 134 testing videos. Each video sequence varies from 20-150 frames of 10 actors exhibiting the five emotions of anger, fear, joy, relief, and sadness. Automatically localized eye and mouth corner points define a affine warp to a canonical face of 60x51 pixels per each frame. A sequence of 16 frames at the 1/3rd and 2/3rd mark of each video is fed into Motion History Image (MHI) [37] analysis yielding a 24x20 dense optical flow per sequence. The X and Y coordinates at each 24x20 grid point for each of the two sequences formed the *D*=1920 input dimensions per sample. The K-SVD methods use a dictionary size of 75 elements.

The i3DPost multi-view [36] activity recognition dataset contains 768 videos of 8 people performing 12 actions from 8 views. The 12 activities were walk, run, jump, bend, hand-wave, jump in place, sit-stand, run-fall, walk-sit, run-jump-walk, handshake, and pull. Each video is MHI processed, giving 125 MHI sequences, each sequence containing 1500 motion vector points. PCA yielded 767 dimensions per video. The dataset contains 512 training videos and 256 testing videos. The K-SVD methods use a dictionary size of 450 elements.

B. Testing Methodologies

The dimensionality reduction techniques capture 99.9% of the data variance, and all use multi-class linear SVM as a classifier. LDA uses equation (4), LPP uses (5), and NPE uses (6) and (7). SPP is similar to NPE, but modifies equation (6) to use sparse coefficients.

The sparse representation techniques all use K-SVD to define an optimal training dictionary of size *m*, where $m<n$. Coefficient transformation matrix *C* is generated from the training set as per (15). Test samples use the *m* element dictionary to generate sparse coefficients using (10). These sparse coefficients are converted to a class estimate using (19). LC-KSVD1 modifies the K-SVD objective function to favor clustering of coefficients by class and LC-KSVD2 further modifies the K-SVD objective function to include the solution of coefficient transformation matrix *C*.

The SRC method uses random projection matrices for dimensionality reduction. The low dimensional projection of training samples forms the training dictionary. The corresponding sparse coefficients of test samples use (11) to make a final classification estimate. The MSR method is identical to SRC, except the random projection dimensionality reduction is replaced with LPP.

C. Experimental Results

Table I demonstrates the performance of the 5 dimensionality reduction methods, the 3 sparsity based methods, and the two combined methods against K-LGE on the 7-class CK+ dataset using the 68 AAM points. Because the data is only 136 dimensions, no dimensionality reduction is used for K-SVD, LC-KSVD1, LC-KSVD2, or SRC. This is a posed dataset, and as such LDA performs the best of the dimensionality reduction techniques.

Table II uses the same CK+ dataset from Table 1, but uses 60x51 images as input. This higher dimensional space is not as discriminative as the 68 AAM points, but all methods do well because of the large separation of facial expression in each class.

Table III uses the 38-class YaleB facial recognition dataset. The 504 random projection input for all methods was further reduced in dimensionality as indicated by the *d* column, where *d* is the dimension where classification is performed. The SR methods are advantaged over the dimensionality reduction methods, while the combined methods of MSR and K-LGE perform the best.

TABLE I. 7-CLASS CK+ EXPRESSION DATASET, 68 AAM POINTS. 164 TRAINING AND 163 TESTING SAMPLES.

Method	*d*	*m*	*% Accuracy*
PCA	62	-	82.2
LDA	6	-	89.6
LPP	62	-	83.4
NPE	24	-	80.4
SPP	48	-	87.7
K-SVD	136	63	79.1
LC-KSVD1	136	63	79.1
LC-KSVD2	136	63	75.5
SRC	136	164	43.6
MSR	62	164	75.5
K-LGE (this paper)	62	63	**92.0**

TABLE II. 7-CLASS CK+ EXPRESSION DATASET, 60X51 IMAGES. 164 TRAINING AND 163 TESTING SAMPLES.

Method	*d*	*m*	*% Accuracy*
PCA	162	-	82.8
LDA	6	-	**86.5**
LPP	163	-	84.7
NPE	71	-	84.0
SPP	80	-	77.9
K-SVD	3060	63	84.0
LC-KSVD1	3060	63	85.9
LC-KSVD2	3060	63	84.7
SRC	500	164	71.8
MSR	163	164	79.1
K-LGE (this paper)	163	63	**86.5**

TABLE III. 38-CLASS YALEB RECOGNITION DATASET. 192X168 PIXEL IMAGES REDUCED TO 504 DIMENSIONS VIA RANDOM PROJECTIONS. 1216 TRAINING IMAEGS, 1198 TESTING IMAGES.

Method	*d*	*m*	*% Accuracy*
PCA	477	-	89.1
LDA	37	-	90.3
LPP	477	-	89.3
NPE	271	-	91.2
SPP	288	-	88.7
K-SVD	504	570	93.2
LC-KSVD1	504	570	93.7
LC-KSVD2	504	570	93.4
SRC	504	1216	86.1
MSR	477	1216	**96.5**
K-LGE (this paper)	477	570	95.3

Table IV uses the 5-class GEMEP-FERA emotion dataset. Two MHI optical flow sequences per video were used as input. The dimensionality reduction methods are advantaged over the SR methods, and the combined methods perform better than the dimensionality reduction methods.

TABLE IV. 5-CLASS GEMEP-FERA EMOTION DATASET. MHI MOTION VECTORS. 155 TRAINING VIDEOS, 134 TESTING VIDEOS.

Method	*d*	*m*	*% Accuracy*
PCA	154	-	55.2
LDA	4	-	55.2
LPP	154	-	55.2
NPE	66	-	56.7
SPP	75	-	52.2
K-SVD	1920	75	51.5
LC-KSVD1	1920	75	53.7
LC-KSVD2	1920	75	51.5
SRC	500	155	57.5
MSR	154	155	56.0
K-LGE (this paper)	154	75	**60.5**

Table V uses the 12-class i3DPost multi-view activity recognition dataset. The 767 PCA projection input for all methods was further reduced in dimensionality as indicated by the *d* column. While there is no clear winner on this dataset, the LPP and K-LGE methods, both based on semi-supervised LPP dimensionality reduction, perform the best.

TABLE V. 12-CLASS I3DPOST MULTI-VIEW ACTIVITY RECOGNITION DATASET. 512 TRAINING VIDEOS, 256 TESTING VIDEOS.

Method	*d*	*m*	*% Accuracy*
PCA	510	-	94.9
LDA	510	-	94.5
LPP	510	-	**96.1**
NPE	224	-	94.9
SPP	241	-	91.0
K-SVD	767	450	94.1
LC-KSVD1	767	450	95.3
LC-KSVD2	767	450	93.8
SRC	767	512	88.7
MSR	510	512	95.3
K-LGE (this paper)	510	450	**96.1**

Fig. 3 (left) shows the effect of the α blend parameter used in (18). Our K-LGE is robust to 0.1≤α≤0.9. Fig. 3 (center) shows the percent improvement from one iteration of K-LGE to stopping condition for each of the five datasets. One iteration of K-LGE is often preferred to most other methods. The number of K-LGE iterations, shown above each point, is often small as the K-LGE method converges quickly. Fig 3 (right) shows the effect of the dictionary size *m* on the i3Dpost multi-view dataset. While the performance of other techniques decreases noticeably with smaller dictionary sizes, K-LGE remains robust to dictionaries as small as *m*=50.

Close inspection of the data in Tables I–V show that no single technique works best in all conditions. However, the K-LGE method consistently ranked first or near the top. We attribute this to the discriminative strengths of dimensionality reduction, the classification power of SR methods, along with the K-LGE graceful unification of the two methods.

When SR methods have insufficient training exemplars in Φ, their performance lags behind SVM classification methods. When datasets are posed, LDA dimensionality reduction is preferred; when datasets are natural, supervised LPP or NPE methods are preferred. K-LGE offers discriminative properties of LDA while maintaining the local topology of complex data representations in low dimensional manifold spaces.

VI. CONCLUSIONS

This paper presents K-LGE, a new method that optimizes both manifold-based dimensionality reduction and sparse representations within a single framework. We believe this is the first attempt to co-optimize dimensionality reduction matrix *U* with dictionary Φ, and training coefficients *a*. We leverage LGE dimensionality reduction techniques and K-SVD dictionary learning techniques to formulate K-LGE. By utilizing semi-supervised LGE dimensionality reduction before SR classification, we not only achieve faster compute times, but are able to minimize coefficient contamination. Successive optimizations of *U,* Φ, and *a*, minimize reconstruction errors, which results in lower classification accuracy. Our results show that our proposed K-LGE framework provides significant advantages over other techniques across a wide variety of facial and activity classification problems.

VII. ACKNOWLEDGMENTS

This paper was supported in part by Cisco and a National Science Foundation Graduate Research Fellowship.

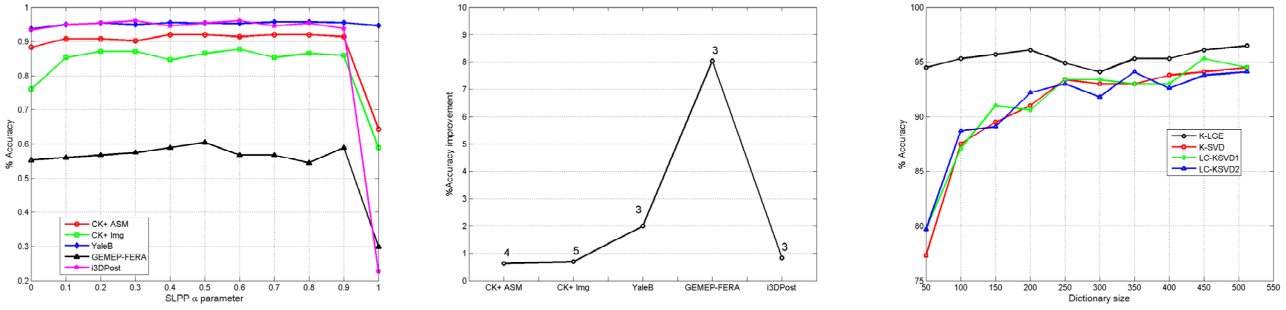

Figure 3. (left) Accuracy of K-LGE as a function of adjacency matrix *W* parameter α in (18). (center) Accuracy improvement by K-LGE iterations. Number on top of each point is the number of iterations to convergence. (right) Performance of the four K-SVD methods as a function of dictionary size on i3DPost dataset.

VIII. REFERENCES

[1] A. Ghodsi., "Dimensionality Reduction A Short Tutorial," University of Waterloo, Ontario,m CA, 2006.

[2] L. Cayton., "Algorithms for manifold learning," University of California, San Diego, Tech Rep. CS2008-0923, 2005.

[3] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science,* vol. 290, pp. 2319-23, 2000.

[4] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science,* vol. 290, pp. 2323-6, 2000.

[5] C. Deng, H. Xiaofei, H. Yuxiao, H. Jiawei, and T. Huang, "Learning a spatially smooth subspace for face recognition," in *CVPR,* 2007.

[6] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?," *Vision Research,* vol. 37, pp. 3311-25, 1997.

[7] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature,* vol. 381, pp. 607-9, 1996.

[8] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Information Theory,* vol. 52, pp. 6-18, 2006.

[9] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory,* vol. 52, pp. 489-509, 2006.

[10] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient Sparse Coding Algorithms," presented at the Advances in Neural Information Processing Systems, 2006.

[11] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least Angle Regression," *Ann. Statist.,* vol. 32, 2004.

[12] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *CVPR,* , 2011.

[13] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher Discrimination Dictionary Learning for Sparse Representation," in *ICCV*, 2011.

[14] A. Yang, J. Wright, Y. Ma, and S. Sastry, "Feature Selection in Face Recognition: A Sparse Representation Perspective," University of California at Berkely, 2007.

[15] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and M. Yi, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 31, pp. 210-27, 2009.

[16] J. Sherrah, S. Gong, and E. J. Ong, "Face distributions in similarity space under varying head pose," *Image and Vision Computing,* vol. 19, pp. 807-819, 2001.

[17] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Sparse representations of image gradient orientations for visual recognition and tracking," in *CVPR,* 2011.

[18] S. Zafeiriou and M. Petrou, "Sparse representations for facial expressions recognition via l1 optimization," in *IEEE CVPR,* 2010.

[19] R. Ptucha, G. Tsagkatakis, and A. Savakis, "Manifold Based Sparse Representation for Robust Expression Recognition without Neutral Subtraction," in *ICCV*, 2011.

[20] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *CVPR*, 2009.

[21] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognition,* vol. 43, pp. 331-341, 2010.

[22] F. Zang and J. Zhang, "Discriminative learning by sparse representation for classification," *Neurocomputing,* vol. 74, pp. 2176-2183, 2011.

[23] M. Zheng*, et al.*, "Graph regularized sparse coding for image representation," *IEEE Transactions on Image Processing,* vol. 20, pp. 1327-1336, 2011.

[24] X. He and P. Niyogi, "Locality Preserving Projections," in *Advances in Neural Information Processing Systems*, Vancouver, Canada, 2003.

[25] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *ICCV*, 2005.

[26] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Asilomar Conference on Signals, Systems and Computers,* 1993.

[27] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *Transactions on Information Theory,* vol. 38, pp. 587-607, 1992.

[28] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing,* vol. 54, pp. 4311-22, 2006.

[29] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient Implementationof the K-SVD Algorithm using Batch Orthogonal Matching Pursuit," Technion, Computer Science Dept., Haifa, Israel, 2008.

[30] Z. Qiang and L. Baoxin, "Discriminative K-SVD for Dictionary Learning in Face Recognition," in *CVPR*, 2010.

[31] W. Jinjun, Y. Jianchao, Y. Kai, L. Fengjun, T. Huang, and G. Yihong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010.

[32] L. Zhihui, J. Zhong, and Y. Jian, "Global sparse representation projections for feature extraction and classification," in *Chinese Conference on Pattern Recognition*, 2009.

[33] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[34] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, pp. 643-60, 2001.

[35] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. R. Scherer, "The First Facial Expression Recognition and Analysis Challenge," in *Face and Gesture Recognition*, 2011.

[36] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3DPost multi-view and 3D human action/interaction," in *CVMP*, 2009.

[37] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, pp. 257-67, 2001.