Joint Optimization of Manifold Learning and Sparse Representations for Face and Gesture Analysis

Raymond Ptucha
rwpee@rit.edu

Artificial Intelligence Seminar
Cornell University
March 1, 2013

RIT Acknowledgements

Dissertation Advisor:
Dr. Andreas Savakis, Professor, Computer Engineering, RIT

Dissertation Committee:
Dr. Nathan Cahill, Associate Professor, School of Mathematical Sciences, RIT
Dr. Joe Geigel, Associate Professor, Computer Science, RIT
Dr. Andreas Savakis, Professor, Computer Engineering, RIT
Dr. Linwei Wang, Assistant Professor, GCCIS PhD, RIT

PhD Program Director:
Dr. Pengcheng Shi, Professor and Department Head of GCCIS PhD Program, RIT

Motivation

• Facial understanding and gesture recognition are powerful enablers in intelligent vision systems.
• Potential applications include surveillance, security, entertainment, smart spaces, and human computer interfaces (HCI).
• Tomorrow's devices will need to embrace human subtleties while interacting with them in their natural conditions.

Interactive Digital Signage
A Few Milestones

- Yang [PAMI '07] used dimensionality reduction with SRs for classification purposes.
- Wright [PAMI '09] used SRs for best in class facial recognition.
- Zafeiriou [CVPR '10] used PCA and SR methods based on Wright for facial expression, but reported significant coefficient contamination.
- Ptucha [ICCV '11] used supervised manifold learning to minimize coefficient contamination.
- Jiang [CVPR '11,'12] used K-SVD to jointly optimize classification accuracy and more efficient dictionaries.

Agenda

- Introduction to Dimensionality Reduction
- Introduction to Sparse Representations
- Merging the two concepts into Manifold based Sparse Representations
- Optimizing the two concepts with LGE-KSVD
- Sample Results

Hypothesis

- Methods based on manifold learning and sparse representations can achieve accurate, robust, and efficient classifiers for scene understanding.

Feature Extraction ➔ Feature Normalization ➔ Manifold Learning
Sparse Representation ➔ Classification Model ➔ Temporal Filtering
Dimensionality Reduction

- For the purpose of facial understanding, the dimensionality of a 26x20 (∈ R^{520}) pixel face image or a 82x2 (∈ R^{164}) set of ASM coordinates are artificially high.
- The high dimensionality space makes the facial understanding algorithms more complex than necessary.
- The set of 520 pixels (or 164 coordinates) actually are samples from a lower dimensional manifold that is embedded in a higher dimensional space.
- We would like to discover this lower dimensional manifold representation (to simplify our facial modeling)- a technique formally called manifold learning. [Cayton '05, Ghodsi '06]
- Given a set of inputs x_1...x_n ∈ R^D, find a mapping y_i = f(x_i), y_1...y_n ∈ R^d; where d < D.

Locality Preserving Projections* (LPP) [He '03]

- Given a set of input points x_1...x_n ∈ R^D, find a mapping y_i = A^T x_i, where the resulting y_1...y_n ∈ R^d, where d << D.
  - Same algebra as PCA, if we kept the top d eigenvectors!
  - Create a fully connected adjacency graph W. Assign high weights to close/similar nodes, and low weights to far/dissimilar nodes.
  - Mimic local neighborhood structure from input to projected space.
- LPP is a linear approximation to the nonlinear Laplacian Eigenmap and is solved via the generalized eigenvector problem:
  \[ X \lambda a = \lambda X D X^T a \]
  - Where:
    - D is a diagonal matrix whose values are the column sums of W,
    - L is the Laplacian matrix: L = D-W,
    - a is the resulting projection matrix (== "eigenvectors") , and
    - \lambda is the resulting vector importance (== "eigenvalues")..

PCA vs. Supervised LPP

- 1072 samples, 26x20 pixel faces (R^{520} → R^3)

Applying Dimensionality Reduction to Pose Training Set

- 21 subjects, each at 21 poses
- Each 164 Dim ASM Face is Mapped Down to 1 point in plot this plot.
- Yaw: R=45 G=30 B=15 C=0 M=15 Y=30 K=45
- Pitch: V = down C = center U = up
Apply Dimensionality Reduction to Pose Training Set
21 subjects, each at 21 poses

Model Manifold Surface

Sparse Representations

- Inspired by studies of neurons in the visual cortex, the notion of Sparse Representations (SRs) has been proven applicable to a variety of scientific fields.
- For many input signals, such as natural images, only a small number of exemplars are needed to represent new test samples.
- SR gives state-of-the-art results for pattern recognition, noise reduction, super-resolution, tracking, …
- At the The First Facial Expression Recognition and Analysis Challenge (FERA2011) at FG’11:
  - 13/15 entrants used SVM, but 0/15 entrants used SR

\[ \hat{x} \approx \sum_{j=1}^{k} a_j \phi_j \]

s.t. \( a_j \)'s are mostly zero (“sparse”)
Sparse Representations

- Given $y$ and $\Phi$, the objective of SRs is to identify the smallest number of nonzero coefficients $a \in \mathbb{R}^n$ such that:
  
  $y \approx \hat{y} = \Phi a$.

- The solution is equivalent to the Lasso regression:
  
  $$\hat{a} = \min \left\{ \|y - \Phi a\|^2 + \lambda \|a\| \right\}$$

  where $\|a\| = \sum |a|$.

- Although not differentiable like a ridge regression, the $\ell^1$ minimization problem can be efficiently solved using convex optimization algorithms. [Donoho ’06, Candes ’06]

- Some of the fastest approaches include several variants of Least Angle Regression with lasso (LARS). [Efron ’04]

Putting it Together

- Manifold based Sparse Representations (MSR) exploit the discriminative behavior of manifold learning, and combines it with the parsimonious power of sparse signal representation.

Sparse Coefficients

Top non-negative ‘$u$’ sparse coefficients for Test “sad” Face.

Interesting... but how do we turn this into a classifier?

- Max peak?
- Max non-zero coefficients?
- Max Energy?
Reconstruction Error

• A reconstruction error classifier generally outperforms other methods. [Yang '07, Wright '09]

• Estimate the class, $c^*$ of a query sample $y$ by comparing the reconstruction error incurred when only the reconstruction coefficients $a_c$ corresponding to a specific class $c$ are selected.

$$c^* = \arg\min_{c=1..z} ||y - \Phi a_c||_2$$

Use non-zero coefficients from all classes to estimate, $y = \Phi \cdot a$

Coefficient Contamination

• Applying the reconstruction error is not a straightforward process for natural images.

• For example, facial identity of the person is often confused with facial expression.

• The usage of semi-supervised manifold learning encourages clustering of sample images in accordance with classification labels.

[Images of PCA and LPP coefficient magnitudes and dimensions for happy and angry faces]
Supervision & Regularization

Region and Pixel Processing

- It is quite conceivable that different regions of the face [Kumar '08] may benefit from different types of pixel processing.
  - Each pixel processing↔facial region combination is a valid feature input to the statistical inference model.

<table>
<thead>
<tr>
<th>MSR accuracies on CK expression dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
</tr>
<tr>
<td>LBP</td>
</tr>
<tr>
<td>Gabor</td>
</tr>
</tbody>
</table>

MSR Used On Other Facial Attributes

<table>
<thead>
<tr>
<th>LFW Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. class</td>
</tr>
<tr>
<td>Gender 2</td>
</tr>
<tr>
<td>Glasses 4</td>
</tr>
<tr>
<td>Hair 7</td>
</tr>
<tr>
<td>Race 5</td>
</tr>
<tr>
<td>Mixed 10</td>
</tr>
<tr>
<td>AVG</td>
</tr>
</tbody>
</table>

Posed vs. Natural Datasets

- MSR enables evaluation of any region of the face

MSR correlates well with [Pflister ICCV2011]
Optimization of Dimensionality Reduction and Sparse Representations

- Global SR Projections [Lai ’09], Discriminative Sparse Coding [Zang ’11], and Graph Regularized Sparse Coding [Zheng ’11] create variations of joint objective function (DR and SR).
- Supervised LPP [Cai ’11] modifies LPP to have (unsupervised) Laplacian and (supervised) LDA properties.
- LC-KSVD [Jiang ’11] forces (unsupervised) sparse terms to be (supervised) discriminative and jointly learns a (supervised) classifier.

LGE-KSVD

- Each of the previous methods introduce a new dimensionality reduction technique or a new SR technique.
- What lacks is a unified approach that optimizes dimensionality reduction projection matrix $U$ with dictionary $\Phi$, and sparse coefficients $\hat{u}$.
- The next few slides will present such a method called LGE-KSVD, for the optimization and infusion of Linear extension of Graph Embedding with K-SVD dictionary learning.  
  - Note: LGE is a broader category of linear dimensionality reduction methods which use adjacency matrix $W$ to describe neighbor to neighbor topology (includes LDA, LPP, and NPE).

LGE-KSVD

- Classification frameworks based on SR concepts have been found to suffer from:
  1. Coefficient contamination that compromises classification accuracy; and
  2. Computational inefficiencies due to high dimensional features and large dictionaries.
- LGE-KSVD uses:
  - Semi-supervised dimensionality reduction to address both limitations.
  - K-SVD dictionary learning to not only make the dictionaries more efficient, but yield higher classification accuracies.

K-SVD

- K-SVD [Aharon ’06] was introduced as a means to learn an over-complete but small dictionary.
  $$\{\hat{\phi}, \hat{u}\} = \min \|x - \Phi \hat{u}\|_2 \ s.t. \ \|u\|_2 \leq \delta$$
- 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 ’08] implemented an efficient implementation of K-SVD using Batch Orthogonal Matching Pursuit (http://www.cs.technion.ac.il/~ronrubin/software.html)
Classification of K-SVD Sparse Coefficients

- Because dictionary elements from K-SVD are a linear combination of input samples, we cannot use the minimum reconstruction error.
- Alternatively we can pass SR coefficients into any regression or machine learning classifier.

Define $H$ as ground truth (GT) matrix, $H \in \mathbb{R}^{k \times n}$.
- Each column of $H$ corresponds to a GT sample. The $k$th position is 1 if $y_i$ belongs to class $k_j$, otherwise 0.
- Coefficients $a$ from each training sample are stored in matrix $A$, $A \in \mathbb{R}^{m \times n}$.
- Then solve for coefficient transformation matrix $C$.

\[
\hat{C} = \min \left\| H - C^T A^T \right\|_2 \quad \Rightarrow \quad C = (A^T A)^{-1} A^T H^T
\]

LGE-KSVD Objective Function

- Combining LGE dimensionality reduction with K-SVD minimization functions, we get:

\[
\{ \hat{U}, \hat{\Phi}, \hat{a} \} = \min \left\{ \left\| X^T U - \Phi a \right\|_F^2 + \frac{U^T X L X^T U}{U^T X D X^T U} \right\} \quad s.t. \left\| \phi \right\|_0 \leq \delta
\]

- The above equation is neither directly solvable nor convex.

Training Procedure for LGE-KSVD

WHILE $\varepsilon$ has not converged or $\varepsilon > \tau$

1. Calculate $U$ using LGE
2. Calculate low dimensional samples $Y = X^T U$
3. Initializes the $m$ samples of $\Phi$ randomly from the $n$ low dimensional training samples
4. Calculate $\{A, \Phi\}$ using modified K-SVD, substituting $Y$ for $X$.
5. Calculate $C$ using $C = (AA^T)^{-1} AH^T$
6. Calculate verification set error, $\varepsilon = \|H - C^T A\|_2^2$

ENDWHILE
Testing Procedure for LGE-KSVD

- Given a test sample \( x \), along with \( U, \Phi, \) and \( C \):

1. Calculate low dimensional sample \( y = x^T U \).
2. Use \( \Phi \) and \( y \) to calculate sparse coefficients \( a \), using a pursuit* algorithm.
3. Use \( C \) along with \( a \) to estimate class label vector \( l \in \mathbb{R}^{k \times 1} \) where the maximum value of \( i \) is used as a class predictor.

\[ i = \max_{i \in 1:k} (C^T a) \]

*such as SLEP, http://www.public.asu.edu/~jye02/Software/SLEP

Modified K-SVD

- K-SVD enforces sparsity by fixing the support of each atom in the iteration process to a subset of training samples.
- The addition of supervision injects classification smarts into K-SVD, but still fixes atom support.
- We propose to use semi-supervised LGE adjacency matrix \( W \) to regulate the support of each dictionary element.

Modified K-SVD

- The support of each dictionary element \( j \) may:
  - Expand: Modify the support of element \( j \) by adding (union) all training entries similar to element \( j \).
  - Contract: Modify the support of element \( j \) by removing (intersection) training entries not similar to element \( j \).
  - Redefine: Set the support of element \( j \) to be only training samples similar to element \( j \).
  - Fixed: Maintain the support of element \( j \), as in the K-SVD algorithm.
- Similar is defined in terms of the LGE adjacency matrix.

Results: CK+ Expression Dataset

7 static facial expressions, 68 AAM points, 164 training and 163 testing samples

<table>
<thead>
<tr>
<th>Method</th>
<th>( d )</th>
<th>( m )</th>
<th>% Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>62</td>
<td>-</td>
<td>82.2</td>
</tr>
<tr>
<td>LOA</td>
<td>6</td>
<td>-</td>
<td>89.6</td>
</tr>
<tr>
<td>LPP</td>
<td>62</td>
<td>-</td>
<td>83.4</td>
</tr>
<tr>
<td>NPE</td>
<td>24</td>
<td>-</td>
<td>80.4</td>
</tr>
<tr>
<td>SPP</td>
<td>48</td>
<td>-</td>
<td>81.7</td>
</tr>
<tr>
<td>K-SVD</td>
<td>136</td>
<td>63</td>
<td>79.1</td>
</tr>
<tr>
<td>LC-KSVD1</td>
<td>136</td>
<td>63</td>
<td>79.1</td>
</tr>
<tr>
<td>LC-KSVD2</td>
<td>136</td>
<td>63</td>
<td>75.5</td>
</tr>
<tr>
<td>SRC</td>
<td>136</td>
<td>164</td>
<td>43.8</td>
</tr>
<tr>
<td>MSR</td>
<td>62</td>
<td>164</td>
<td>75.5</td>
</tr>
<tr>
<td>LGE based K-SVD</td>
<td>62</td>
<td>63</td>
<td>92.0</td>
</tr>
</tbody>
</table>

Ptucha '13 62

Ptucha '13 63

Ptucha '13 64

Ptucha '13 65
### Results: CK+ Expression Dataset
7 static facial expressions, 60x51 images, 164 training and 163 testing samples

<table>
<thead>
<tr>
<th>Method</th>
<th>d (dimension)</th>
<th>m (# dictionary atoms)</th>
<th>% Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>162</td>
<td>-</td>
<td>82.8</td>
</tr>
<tr>
<td>LDA</td>
<td>6</td>
<td>-</td>
<td>86.5</td>
</tr>
<tr>
<td>LPP</td>
<td>163</td>
<td>-</td>
<td>84.7</td>
</tr>
<tr>
<td>NPE</td>
<td>71</td>
<td>-</td>
<td>84.0</td>
</tr>
<tr>
<td>SPP</td>
<td>80</td>
<td>-</td>
<td>77.9</td>
</tr>
<tr>
<td>K-SVD</td>
<td>3060</td>
<td>65</td>
<td>84.0</td>
</tr>
<tr>
<td>LC-KSVD1</td>
<td>3060</td>
<td>65</td>
<td>85.9</td>
</tr>
<tr>
<td>LC-KSVD2</td>
<td>3060</td>
<td>65</td>
<td>84.7</td>
</tr>
<tr>
<td>SPC</td>
<td>155</td>
<td>104</td>
<td>73.8</td>
</tr>
<tr>
<td>MSR</td>
<td>163</td>
<td>164</td>
<td>79.1</td>
</tr>
<tr>
<td>LGE based K-SVD</td>
<td>163</td>
<td>63</td>
<td>86.5</td>
</tr>
</tbody>
</table>

### Results: YaleB Recognition Dataset
38 subjects, 192x168 static images reduced to 504 dimensions via random projections, 1216 training and 1198 testing samples

<table>
<thead>
<tr>
<th>Method</th>
<th>d (dimension)</th>
<th>m (# dictionary atoms)</th>
<th>% Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>477</td>
<td>-</td>
<td>89.1</td>
</tr>
<tr>
<td>LDA</td>
<td>37</td>
<td>-</td>
<td>90.3</td>
</tr>
<tr>
<td>LPP</td>
<td>477</td>
<td>-</td>
<td>89.3</td>
</tr>
<tr>
<td>NPE</td>
<td>271</td>
<td>-</td>
<td>91.2</td>
</tr>
<tr>
<td>SPP</td>
<td>288</td>
<td>-</td>
<td>88.7</td>
</tr>
<tr>
<td>K-SVD</td>
<td>504</td>
<td>570</td>
<td>93.2</td>
</tr>
<tr>
<td>LC-KSVD1</td>
<td>504</td>
<td>570</td>
<td>93.7</td>
</tr>
<tr>
<td>LC-KSVD2</td>
<td>504</td>
<td>570</td>
<td>93.4</td>
</tr>
<tr>
<td>SRC</td>
<td>504</td>
<td>1216</td>
<td>96.1</td>
</tr>
<tr>
<td>MSR</td>
<td>477</td>
<td>1216</td>
<td>96.5</td>
</tr>
<tr>
<td>LGE based K-SVD</td>
<td>477</td>
<td>570</td>
<td>95.3</td>
</tr>
</tbody>
</table>

### Results: GEMEP-FERA Emotion
5 class, two 24x20 MHI static images per video, 155 training and 134 testing samples

<table>
<thead>
<tr>
<th>Method</th>
<th>d (dimension)</th>
<th>m (# dictionary atoms)</th>
<th>% Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>154</td>
<td>-</td>
<td>55.2</td>
</tr>
<tr>
<td>LDA</td>
<td>4</td>
<td>-</td>
<td>55.2</td>
</tr>
<tr>
<td>LPP</td>
<td>154</td>
<td>-</td>
<td>55.2</td>
</tr>
<tr>
<td>NPE</td>
<td>66</td>
<td>-</td>
<td>56.7</td>
</tr>
<tr>
<td>SPP</td>
<td>75</td>
<td>-</td>
<td>52.2</td>
</tr>
<tr>
<td>K-SVD</td>
<td>1920</td>
<td>75</td>
<td>51.5</td>
</tr>
<tr>
<td>LC-KSVD1</td>
<td>1920</td>
<td>75</td>
<td>53.7</td>
</tr>
<tr>
<td>LC-KSVD2</td>
<td>1920</td>
<td>75</td>
<td>51.5</td>
</tr>
<tr>
<td>SRC</td>
<td>500</td>
<td>155</td>
<td>57.5</td>
</tr>
<tr>
<td>MSR</td>
<td>154</td>
<td>155</td>
<td>56.0</td>
</tr>
<tr>
<td>LGE based K-SVD</td>
<td>154</td>
<td>75</td>
<td>68.5</td>
</tr>
</tbody>
</table>

### Results: i3DPost Multi-View Activity
12 class, 125 MHI sequences per video, PCA reduced 767 dimensions per video, 512 training and 256 testing samples

<table>
<thead>
<tr>
<th>Method</th>
<th>d (dimension)</th>
<th>m (# dictionary atoms)</th>
<th>% Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>510</td>
<td>-</td>
<td>94.9</td>
</tr>
<tr>
<td>LDA</td>
<td>510</td>
<td>-</td>
<td>94.5</td>
</tr>
<tr>
<td>LPP</td>
<td>510</td>
<td>-</td>
<td>96.1</td>
</tr>
<tr>
<td>NPE</td>
<td>224</td>
<td>-</td>
<td>94.9</td>
</tr>
<tr>
<td>SPP</td>
<td>241</td>
<td>-</td>
<td>91.0</td>
</tr>
<tr>
<td>K-SVD</td>
<td>767</td>
<td>450</td>
<td>94.4</td>
</tr>
<tr>
<td>LC-KSVD1</td>
<td>767</td>
<td>450</td>
<td>95.3</td>
</tr>
<tr>
<td>LC-KSVD2</td>
<td>767</td>
<td>450</td>
<td>93.8</td>
</tr>
<tr>
<td>SRC</td>
<td>767</td>
<td>512</td>
<td>88.7</td>
</tr>
<tr>
<td>MSR</td>
<td>767</td>
<td>512</td>
<td>95.3</td>
</tr>
<tr>
<td>LGE based K-SVD</td>
<td>510</td>
<td>450</td>
<td>96.1</td>
</tr>
</tbody>
</table>
Temporal Processing

- Communication between humans naturally contains temporal signature.
  - Rolling of eyes, waving of hand, wink, etc.
- Previous studies adopted both sparse and dense optical flow techniques and contrast to static methods.
- Facial expressions and gestures can occur at any point in time and are variable in length.
- We define sliding temporal windows, $W^l$, each of duration $\theta$ frames, $l=1..m$ sliding windows.

Examine Video In Variable Size Rolling Frame Buffers

$n$ Video Frames

All Buffers of size 2
Examine Video In Variable Size Rolling Frame Buffers

Examine Video In Variable Size Rolling Frame Buffers

Analysis Example

- Let's say, we are looking at window widths of 8.
- Our first position center is frame 12.
- We then look at 7 motion trajectories:

Facial Feature Point Tracking

Similarly, can compute point tracking from current frame the mean frame.
Summary

- Face and gesture understanding problems can be reliably solved in unconstrained scenes using SRs.
- The usage of semi-supervised LPP before SR clusters by classification task, avoiding coefficient contamination.
- The usage of K-SVD dictionary learning makes the dictionaries more compact and results in higher classification accuracies.
- If the training dictionary is not over complete, SR methods have trouble generalizing test samples from training dictionary exemplars.

References (1 of 2)

References (2 of 2)


Thank you!!

Ray Ptucha
rwpec@rit.edu
raymond.ptucha@kodak.com