Image and Vision Computing 31 (2013) 365–378

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Manifold based sparse representation for facial understanding in natural images ☆

Raymond Ptucha *, Andreas Savakis

Rochester Institute of Technology, Rochester, NY, USA

ARTICLE INFO

Article history:
Received 10 April 2012
Received in revised form 31 January 2013
Accepted 24 March 2013

Keywords:
Classification of facial expression
Race
Gender
Glasses and facial hair

ABSTRACT

Sparse representations, motivated by strong evidence of sparsity in the primate visual cortex, are gaining popularity in the computer vision and pattern recognition fields, yet sparse methods have not gained widespread acceptance in the facial understanding communities. A main criticism brought forward by recent publications is that sparse reconstruction models work well with controlled datasets, but exhibit coefficient contamination in natural datasets. To better handle facial understanding problems, specifically the broad category of facial classification problems, an improved sparse paradigm is introduced in this paper. Our paradigm combines manifold learning for dimensionality reduction, based on a newly introduced variant of semi-supervised Locality Preserving Projections, with a ℓ^1 reconstruction error, and a regional based statistical inference model. We demonstrate state-of-the-art classification accuracy for the facial understanding problems of expression, gender, race, glasses, and facial hair classification. Our method minimizes coefficient contamination and offers a unique advantage over other facial classification methods when dealing with occlusions. Experimental results are presented on multi-class as well as binary facial classification problems using the Labeled Faces in the Wild, Cohn–Kanade, Extended Cohn–Kanade, and GEMEP-FERA datasets demonstrating how and under what conditions sparse representations can further the field of facial understanding.

1. Introduction

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 [1,2]. Patterns of light are represented by a series of innate or learned basis functions whereby sparse linear combinations form a surrogate input stimuli to the brain. Similarly, for many input signals of interest, such as natural images, a small number of exemplars can form a surrogate representation for a new test image.

In SR systems, new test images are efficiently represented by sparse linear coefficients on a dictionary Φ of overcomplete basis functions. Specifically, SR systems are comprised of an input sample $y \in \mathbf{R}^m$ along with a dictionary Φ of n samples, $\Phi \in \mathbf{R}^{mxn}$. SR solves for coefficients $a \in \mathbf{R}^n$ that satisfy the ℓ^1 minimization problem $\hat{y} = \Phi a$.

It has been shown that under typical conditions, the minimal solution is the sparsest one [3,4]. There have been several studies optimizing both the ℓ^1 minimization [5,6] as well as the selection of dictionary elements [7,8].

Distinct advantages afforded by SR architectures include:

1. The class and numeric value of each non-zero coefficient $a_i \in \mathbf{R}^n$ can be used as salient input to a classifier.
2. The usage of sparse coefficients provides a robust method for occlusion detection.
3. Since the number of non-zero coefficients α_i is small, computational overhead is minimized.
4. There are no restrictions on the upper bound of the size of the training dictionary Φ.

Researchers have chosen to tackle the problem of reproducing the sparse nature of simple-cells in the primary visual cortex by independently proposing similar architectures [9,10]. Sparse features are rectified, pooled, and then dimensionality reduced in preparation for classification. The rectification step, or taking the absolute value of sparse features, compensates for negative coefficients. The pooling step introduces invariance—e.g. when dealing with facial classification the invariance is with respect to small amounts of pose and appearance changes. Simple-cells in the visual cortex appear to exhibit non-negative properties leading to the Nonnegative Matrix Factorization (NMF) algorithm [18]. NMF forces all coefficients to be positive, which negates the need for rectification. The common solution for

☆ This paper has been recommended for acceptance by Stefanos Zafeiriou.
* Corresponding author. Tel.: +1 585 797 5561.
E-mail address: rwpeec@rit.edu (R. Ptucha).

0262-8856/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2013.03.003

pooling is to lowpass filter and down sample, or use probabilistic max pooling of the input data. An alternate approach is to learn the underlying embedded manifold of the input data and then perform appropriate dimensionality reduction.

Perhaps the most influential usage of SRs in the facial understanding community was the work of Wright et al. [11,12]. Rather than using SR as a pre-processor to a classifier, as in [9,10], Wright et al. applied SR on random projections of pixel data, and by feeding the α coefficients directly to a classifier demonstrated state-of-the-art facial recognition results. 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 the sparse coefficients. This coefficient contamination is unfortunate, 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 [13].

Huang et al. [14] improved Wright's framework by introducing invariance to modest in-plane transformations. Tzimiropoulos et al. [15] demonstrated computational and accuracy improvements by doing the ℓ^1 optimization in a dimensionally reduced space. Zafeiriou and Petrou [16] used Principal Component Analysis (PCA) and SR techniques based on [12] for facial expression recognition. The work in [16] struggled with coefficient contamination, noting that applying Wright's framework is not a straightforward process because the facial identity of the person is often confused with facial expression. To deal with this problem, [16] proposed to subtract a neutral expressive image from the test image before processing, which can be impractical when such an image is not available. We aim to introduce a framework that can be applied across all facial understanding problems.

Facial expression [17,18] and facial understanding problems in general, require accurate feature extraction. Corners of the eyes and mouth, as localized by Active Shape Models (ASM) or Active Appearance Models (AAM) [19,20], can define a mapping to a canonical face representation. Kumar et al. [21] used these points to define a set of facial regions for facial verification. Unfortunately, these facial regions constitute high dimensional representations of facial features. Linear dimensionality reduction techniques such as PCA or Linear Discriminant Analysis (LDA) produce more meaningful lower dimensional representations. However, the underlying linearity assumption may not be suitable for efficiently modeling the behavior of the high dimensional imagery such as face representations.

Manifold learning [22,23] techniques reduce the dimensionality of input data by identifying a linear or nonlinear lower dimensional space where the data resides. Methods include Isomap [24] and Locally Linear Embedding (LLE) [25]. In order to support the extension of the manifold model to new examples, linearized techniques, such as the Locality Preserving Projections (LPP) [26], solve a linear approximation of a non-linear object.

Similar to subspace clustering [27], supervision in manifold learning encourages clustering of sample images by class, minimizing ℓ^1 coefficient contamination. For example, in [28] we eliminated the need for neutral frame subtraction by preprocessing with supervised LPP, which encourages clustering by facial expression rather than by facial identity. Unsupervised manifold learning preserves the local topology from input to output space, more effectively modeling the inherent manifold structure and yielding more accurate classification on natural datasets. Our newly introduced semi-supervised LPP utilizes advantages from both supervised and unsupervised manifold learning methodologies to simultaneously minimize coefficient contamination and classification errors when dealing with natural datasets.

Despite their successes in facial recognition, there has been slow adoption of SRs for other applications of facial understanding, such as expression, gender, age, and race estimation. As evidence of such, none of the fifteen entrants into the 2011 Facial Expression Recognition and Analysis Challenge utilized SRs. Coefficient contamination remains the biggest drawback and is most problematic in natural imagery due to unconstrained facial pose, expression, and imaging conditions.

Our approach, inspired by improvements to Wright's framework offers two key advantages:

1. We have virtually eliminated the coefficient contamination problem without the need for neutral reference frames.
2. Our solution is robust in the presence of occlusions.

The classification results reported in this paper use the Labeled Faces in the Wild (LFW) [29] dataset, the Cohn–Kanade (CK) [30] dataset, the extended Cohn–Kanade (CK+) [31] dataset, as well as the dataset used for the Facial Expression Recognition and Analysis Challenge (FERA2011), referred to as the GEMEP-FERA [32] dataset. The CK and CK+ datasets are representative of posed datasets—faces are frontal and expressions are generally exaggerated. The LFW and GEMEP-FERA datasets are representative of natural datasets—faces are of varying pose, expression, gender, race, facial hair, glasses, and occlusion. In addition, the LFW dataset includes a large range of image fidelity.

This paper introduces a new SR architecture geared towards understanding and improving SR methods for facial classification problems. As opposed to trying to mimic biology [33], we seek to improve classification accuracy by combining the best of biologically inspired approaches with the best of computer vision techniques. Our Manifold-based Sparse Representation (MSR) method:

a) Combines manifold learning within the SR architecture. The introduction of manifolds allow for more aggressive dimensionality reduction, and offer elegant representations while minimizing computational resources. We explore, through numerous experiments, the performance of various manifold methods with various SR methods.
b) Minimizes coefficient contamination. We introduce a new adjacency matrix calculation which blends the benefits of LDA with local manifold methods. This calculation simultaneously encourages clustering of sample images by class while respecting the local topology from input to output space. We use several variants of posed and natural datasets to demonstrate the robustness of our proposed framework.
c) Improves facial classification performance in the presence of occlusions. We introduce a multi-part facial statistical inference model, which, when used in combination with a ℓ^1 reconstruction model [11], demonstrates state-of-the-art results with or without occlusions.

The rest of this paper is organized as follows. In Section 2 the method proposed by this paper is overviewed. In Sections 3 and 4 principles of manifold learning and sparse signal representation are introduced. In Section 5 experimental results are presented. Section 6 summarizes with conclusions.

2. Sparse representations in manifold space

The facial classification approach used in this paper is outlined in Fig. 1. After face detection, ASM placement automatically localizes eye and mouth corner points. Faces are affine mapped to a canonical face representation. Targeted facial regions are extracted via a masking process, normalized, and then mapped onto a low dimensional manifold surface learned by a newly introduced variant of semi-supervised LPP more suitable for natural imagery. Sparse representations encode the test face regions projected onto the manifold as a nonnegative linear combination of dictionary elements.

A reconstruction model compares the reconstructed facial region, using sparse coefficients from all classes, to the reconstructed facial region using only coefficients from each respective class. The class with the smallest reconstruction error to the fully reconstructed facial region is indicative of the facial classification. We term this architecture

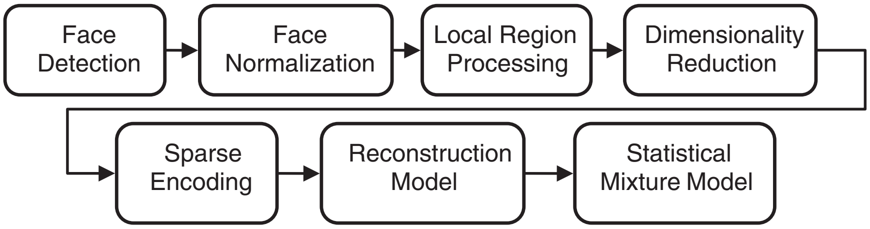


Fig. 1. Flowchart of manifold-based sparse representation architecture.

Manifold-based Sparse Representation or MSR. By examining several facial regions simultaneously, statistical inference models boost final classification performance and introduce robustness to occlusions.

3. Manifold learning

3.1. 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, from Gabor jets, to concatenated block histograms. For example, Lucey et al. [31] represent faces using 68 landmark points and 87 × 93 pixels. As a result, each face resides in $\mathbf{R}^D$ where $D = (68 \times 2 + 87 \times 93) = 8227$. 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.

Formally, the input feature space contains n samples, x_1, x_2, …x_n, each sample $x_i \in \mathbf{R}^D$. These n samples are projected onto a lower dimensional representation, yielding y_1, y_2, …y_n, each output sample $y_i \in \mathbf{R}^d$. As $d \le D$ in all cases, we are interested in the case where $d \ll D$. In the general sense, we are interested in finding a function f, such that $Y \approx f(X)$. For linear models, $y_i^T = Ux_i^T$, where U is a $d \times D$ projection matrix.

To enable dimensionality reduction using the commonly used approaches of PCA or LDA, the top d eigenvectors that encode most of the data variance are used for the projection matrix. To model the top T % variance of the training samples, d is solved such that $\Sigma\lambda_j \,/\, \Sigma\lambda_k \ge T \,/\, 100$; where $j = 1,2,…d$; $k = 1,2,…D$.

Alternatively, the high dimensional feature space can be parameterized by a lower dimensional embedded manifold discovered using manifold learning [29,30]. In addition to being more compact, the resulting lower dimensional manifold representation is often more discriminative and thus more appropriate for subsequent classification. One such technique uses a fully connected graph representation of the input space whereby each of the n input samples or nodes is connected to all other $(n - 1)$ input samples with a weight, $0 \le w_{ij} \le 1$, $i,j = 1…n$. A weight of 0 signifies no connection between samples i and j, and a weight of 1 signifies the strongest connection. This graph is commonly called an adjacency graph. 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 K nearest neighbors of x_j, 0 otherwise. Alternatively, w_{ij} is set to 1 if $||x_i - x_j|| < \varepsilon$, and 0 otherwise. Setting w_{ij} to continuous values between 0 and 1 offers more control in describing sample connections.

3.2. Locality Preserving Projections

Manifold learning using Locality Preserving Projections (LPP) [26] is based on the geometric structure of the input space, found by solving a linear approximation to the nonlinear Laplacian Eigenmap [34]. LPP is obtained from a graph model of our input data and shares many of the data topology properties of nonlinear techniques, but is linear and defined throughout space. LPP creates an adjacency map of the top z neighbors for each feature point x_i, weighting each neighbor by distance to form $n \times n$ adjacency matrix W with entries w_{ij}. Close or similar neighbors have high values, far neighbors set $w_{ij} = 0$. As such, LPP represents the local manifold structure of the z nearest neighbors. We seek to minimize the objective function:

$$\sum_{i,j} \left(y_i - y_j\right)^2 w_{ij}. \tag{3.1}$$

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. LPP seeks a linear approximation to this nonlinear problem of the form $y_i = ux_i$ or $y^T = u^T x$. 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, $L = D - W$. The matrix U that minimizes our objective function is represented by a linear extension of graph embedding framework whose solution is given by the minimum eigenvalue of the generalized eigenvector problem:

$$XLX^T U = \lambda XDX^T U \tag{3.2}$$

where U is the resulting projection matrix.

One of the more effective strategies to set the connection weights w_{ij} of adjacency matrix W is Gaussian kernel weighting (also referred to as heat kernel weighting). If nodes i and j are connected, set:

$$w_{ij} = e - \frac{||x_i - x_j||^2}{t}. \tag{3.3}$$

When supervised labels are available, a discriminative embedding can be achieved. Formally, W is initialized to all zeros, and then w_{ij} entries corresponding to the same classes are set to $1 \,/\, k_n$, where k_n is the number of samples per class:

$$w_{ij} = 1/k_n. \tag{3.4}$$

We shall refer to Eq. (3.4) as an LDA kernel [35]. Although the LDA kernel can produce excellent results, the rank of the matrix is $k - 1$, where k is the number of classes, and thus the maximum number of eigenvectors is $k - 1$, restricting $d \le (k - 1)$. When k is small, say for binary gender classification, such extreme dimensionality reduction is not practical.

To maintain input topology and increase the rank of W, [36] proposed adding the LDA between-class and LDA within-class matrix to the LPP objective function in Eq. (3.2). Recognizing that some problems are linearly separable, while others are not, we propose utilizing a convex combination of the Gaussian and LDA kernels:

$$W = \alpha W_{LDA} + (1-\alpha) W_{Gaussian} \tag{3.5}$$

where $0 \le \alpha \le 1$. The kernels W_{LDA} and $W_{Gaussian}$ are $n \times n$ matrices, where n is the number of input samples. For posed imagery, or datasets with exaggerated differences between classes, the choice of α is forgiving $\forall \alpha > 0$. For natural imagery, any value of α (other than 0 or 1) produces improved results over the Gaussian or LDA kernel. Conceptually, by adding a percentage of the Gaussian kernel to the LDA kernel, we mimic the local non-linear input topology while simultaneously increase the matrix rank. For datasets that are difficult to separate by class, α should be decreased to learn the local topology. Additionally, for LDA kernels of low rank, the performance improvement can be dramatic. The greater k is, the higher the LDA kernel rank becomes, and the lower the improvement obtained. In this work we employ this form of semi-supervised LPP, termed SLPP.

To demonstrate the power of our SLPP method, 1072 faces of 5 expressions were taken from the CK [31] dataset. Each face was cropped and resampled down to 26 × 20 pixels, for an input $D = 520$. These faces were trained via PCA and SLPP, where only the top three eigenvectors from each were utilized for our projection matrix, U, a 3 × 520 projection matrix. Fig. 2 demonstrates the advantage supervised dimensionality reduction methods such as SLPP can have over unsupervised methods when performing subspace clustering. We will see shortly that good subspace clustering is critical to minimizing coefficient contamination of sparse coefficients.

4. Sparse signal representation

4.1. Sparse representation

Many computer vision problems can be reformulated as finding a linear representation of an input signal from a dictionary, Φ, of training examples. Let the input signal be $y \in \mathbf{R}^d$ and the dictionary of examples $\Phi \in \mathbf{R}^{dxn}$. A natural way to represent y from Φ is by solving $\hat{y} = \Phi a$, where $a \in \mathbf{R}^n$ is the weight of each training exemplar in our dictionary Φ. However, in most practical cases, the system has either no solution or multiple solutions. This problem is usually tackled by applying a least squares regression:

$$\hat{a} = \arg\min||a||_2 \quad s.t. \hat{y} = \Phi a \tag{4.1}$$

where $||a||_2$ is the ℓ^2 norm and the reconstruction coefficients a are given by $a = \Phi^\dagger y$, where $\Phi^\dagger = (\Phi^T\Phi)^{-1}\Phi^T$ is the Moore–Penrose pseudo-inverse of Φ. The ℓ^2 solution offers two significant benefits, a closed form solution and the generation of a unique solution. Although regularizing the energy of the solution has been successfully applied in many problems, minimizing the ℓ^2 norm may not lead to the optimum solution for specific types of signals such as images.

The notion of sparse representation (SRs), i.e. exploiting the sparsity in modeling the signal, has been instrumental for various computer vision problems. For sparse signals, the objective of SRs is to identify the smallest number of nonzero coefficients $a \in \mathbf{R}^n$ such that $\hat{y} = \Phi a$. It has been shown that under typical conditions, the minimal solution is the sparsest one [3,4]. The solution to this problem can be obtained by solving the following optimization problem:

$$\hat{a} = \arg\min||a||_0 \quad s.t. \hat{y} = \Phi a \tag{4.2}$$

where the sparsity constraint is given by the zero-norm $||\cdot||_0$ which counts the non-zero elements—perhaps the best measure of sparsity. Unfortunately, the problem in Eq. (4.2) is non-convex and therefore difficult to solve for practical problems. Two approaches have been presented for reformulating the intractable problem in Eq. (4.2) into an efficient optimization: convex relaxation and greedy algorithms.

The convex relaxation approach was introduced by Donoho et al. [3] and Candes et al. [4], where it was shown that if the solution satisfies certain constraints, such as the sparsity of the representation, the solution of Eq. (4.2) is equivalent to the solution of the following Lasso regression problem in statistics:

$$\hat{a} = \arg\min||a||_1 \quad s.t. \hat{y} = \Phi a \tag{4.3}$$

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 Eq. (4.3) as:

$$\hat{a} = \min||a||_1 \quad s.t.||\hat{y} - \Phi a||_2 \leq \varepsilon. \tag{4.4}$$

Often Eq. (4.4) is approximated by loosening the error constraints and reconfigured to specifically include a regularization term, λ

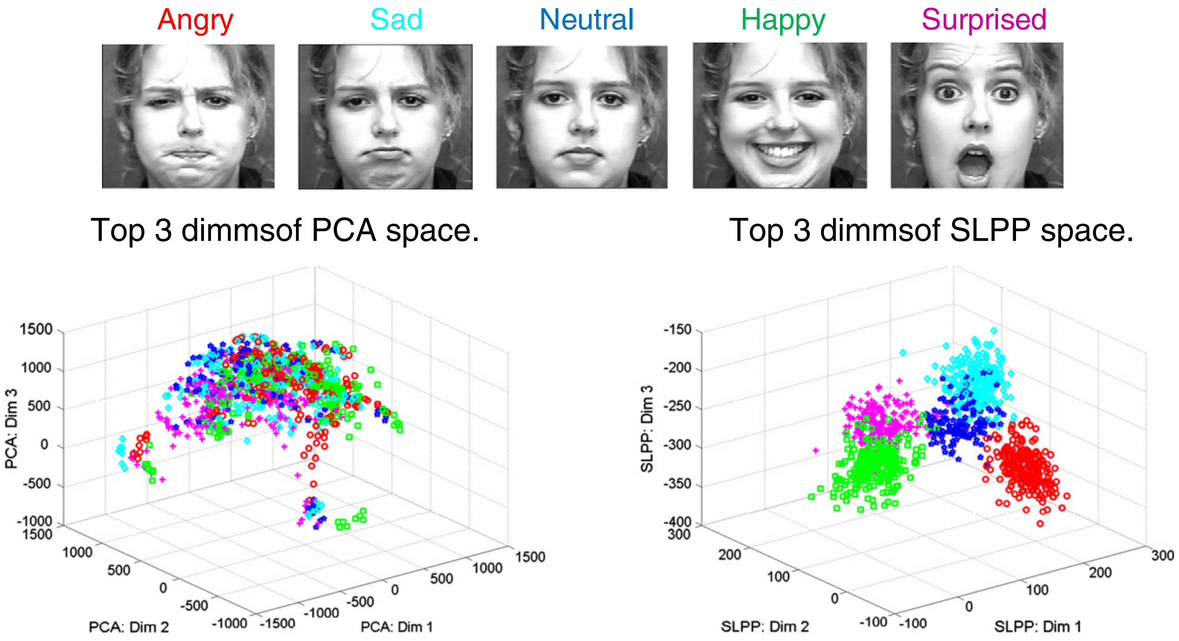


Fig. 2. Top row: A sample subject at the five training expressions. Bottom row: Each point represents one of 29 subjects exhibiting one of the 5 expressions after reduction to 3 dimensions using PCA (bottom left) and SLPP (bottom right). Expressions are color coded as shown above the training samples.

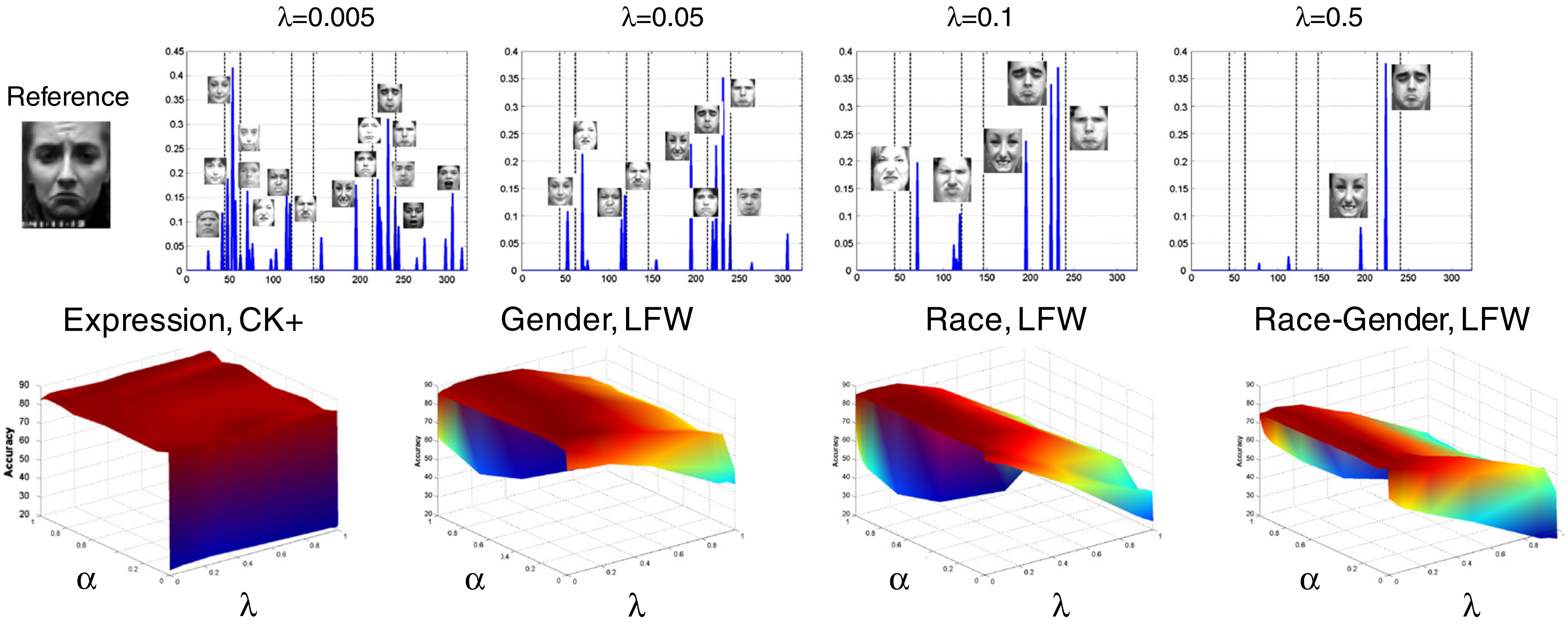


Fig. 3. Top row: Varying the regularization parameter λ affects the sparseness of the ℓ^1 coefficients. More regularization creates fewer non-zero $\hat{a}$ coefficients. Bottom row: Classification accuracy (vertical axis) obtained by varying α, the adjacency matrix blend parameter along with ℓ^1 regularization parameter λ on varying datasets and classification problems. The left-most plot is a 7-class posed dataset. The three remaining plots are 2-class gender, 5-class race, and 10-class race–gender natural datasets. As λ goes to the right, we increase regulation. As α goes to the left, we increase LDA discriminative embedding.

which encourages sparseness by incurring a penalty on the resulting coefficients:

$$\hat{a} = \min\left\{||\hat{y} - \Phi a||_2^2 + \lambda||a||_1\right\}. \tag{4.5}$$

Perhaps the most widely used method to solve the ℓ^1 minimization of Eqs. (4.4) and (4.5) is Orthogonal Matching Pursuit (OMP) [37]. OMP selects one dictionary element at a time in a greedy fashion and can quickly converge to a solution. Forward stagewise is a cautious version of forward selection method. Instead of taking a few large steps, perhaps a thousand smaller steps are taken. Least Angle Regression with laSso (LARS) [6] attempts to strike a balance between forward selection and forward stagewise methods.

The greater the value of λ in Eq. (4.5), the sparser the solution and the fewer atoms from Φ that are used to estimate $\hat{y}$. The top row of Fig. 3 illustrates the effect λ has on the solution of a sample face. Starting with a dictionary size of 330 faces, the regularization parameter λ determines how many atoms are used to represent the reference face. The parsimonious notion of sparsity would indicate that generous values of λ are tolerable. To prevent overfitting to the training data, one can either increase the size of the training dictionary or increase λ On sufficiently over-complete dictionaries, empirical testing has shown $0.05 \leq \lambda \leq 0.25$ yields a good tradeoff between classification accuracy and processing complexity.

To gain further intuition for the SLPP adjacency matrix weight α and the ℓ^1 regularization parameter λ, the bottom row of Fig. 3 shows the effect of varying α and λ to different dataset and classification problems. The CK+ dataset is a posed dataset with rather large class to class separation in SLPP space. The three LFW datasets are natural datasets with considerably more overlap between class distributions in SLPP space.

With respect to the Expression CK+ plot in the bottom of Fig. 3, the inherent data is linearly separable. As long as the adjacency matrix includes some amount of supervision, the accuracy is rather flat. With respect to the LFW plots, it is evident that neither the Gaussian or LDA adjacency matrix performs as well as the convex combinations of the two. Furthermore, while the 2-class gender and 5-class race plots have large drops when $\alpha = 1$, the 10-class race–gender plot is not as pronounced because the rank of W is higher. Additionally, less regularization on the LFW datasets gives preferred results due to considerable sample to sample variability in this dataset, which makes it difficult to have an over-complete dictionary.

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 for choosing the sparsest solution and produces smooth and graceful coefficient activity across diverse test samples [38]. For efficiency, we would like our dictionary size to be small, necessitating the need for linearly independent or decorrelated samples. To minimize reconstruction error, we often need to increase the size of our dictionary, however too many samples in our dictionary result in unstable estimates for $\hat{a}$ as well as greater computational burden. Wright [12] used the entire training set, while Aharan et al. introduced K-SVD [39] to learn an over-complete but small dictionary.

4.2. Sparse representations classification

Given the sparse representation coefficients $\hat{a}$ of a test image using the dictionary Φ, various techniques can be used for identifying the class of the test image. To motivate the selection of a classification scheme based on SRs, Fig. 4 shows a sample expressive image along with its six non-zero $\hat{a}$ coefficients from ℓ^1 sparse representation. The test face belongs to the *sad* class, but it is readily apparent that dictionary elements come from several classes.

Using the set of sparse representation coefficients $\hat{a}$ as input, there are a variety of ways to perform classification. Referring again to Fig. 4, we can assign the test sample to the class with the most significant $\hat{a}$ coefficient, the class with the most non-zero $\hat{a}$ coefficients, or to the class with the greatest sum of all $\hat{a}$ coefficients. The first strategy would have incorrectly classified the exemplar *sad* face as *angry*, while the latter two would have resulted in a correct classification. Our experiments have shown that using a reconstruction error approach outperforms the aforementioned methods. The reconstruction error method estimates the class c^* of a query sample y by comparing the reconstructed facial region using sparse coefficients α from all classes to the reconstructed facial region using coefficients a^c from each respective class:

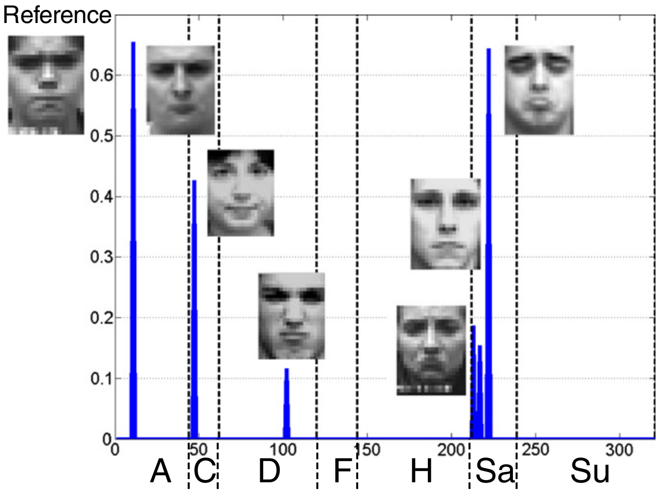


Fig. 4. Sample sad face and its top 6 $\hat{a}$ coefficients. Coefficients are grouped by CK+ categories of Anger (A), Contempt (C), Disgust (D), Fear (F), Happy (H), Sadness (Sa), and Surprise (Su).

$$c^* = \arg\min_{c=1\ldots z}\left|\left|y - \Phi a^c\right|\right|_2. \tag{4.6}$$

Eq. (4.6) assigns the test sample classification to the class most similar to the fully reconstructed facial signal. The coefficients in Fig. 4 have been constrained to be non-negative and y is estimated by using all coefficients, $y \approx \hat{y} = \Phi a$. Empirical studies have shown improved performance of the nonnegative SR method over ℓ^1 SR. This is consistent with previous findings and also mimics the behavior of simple-cells in the visual cortex.

The robustness of Eq. (4.6) is directly tied to the discriminative nature of the $\hat{a}$ coefficients. The α parameter of SLPP is a driving factor. The closer α is to 1.0, the stronger the discriminant nature between classes. To illustrate this effect, Fig. 5 shows two sample faces from the CK+ dataset. Each row shows the $\hat{a}$ coefficients for PCA, LPP, SLPP, and LDA LPP. As α is increased, we observe the tightening of the $\hat{a}$ coefficients to their respective class categories. This tightening minimizes coefficient contamination. As such, SR classification methods benefit from a previous subspace clustering step. Too much tightening (e.g. $\alpha = 1.0$) is likely to cause failure on natural datasets with limited dictionary Φ size.

4.3. Multiple features and the statistical mixture model

Processing individual facial regions is motivated by the need to improve classification accuracy and enable classification in the presence of occlusions.

Fig. 6 shows the eleven facial regions studied in this paper, many inspired by Kumar et al. [21]. Each masked region was independently trained to obtain eleven different SLPP manifold models and eleven

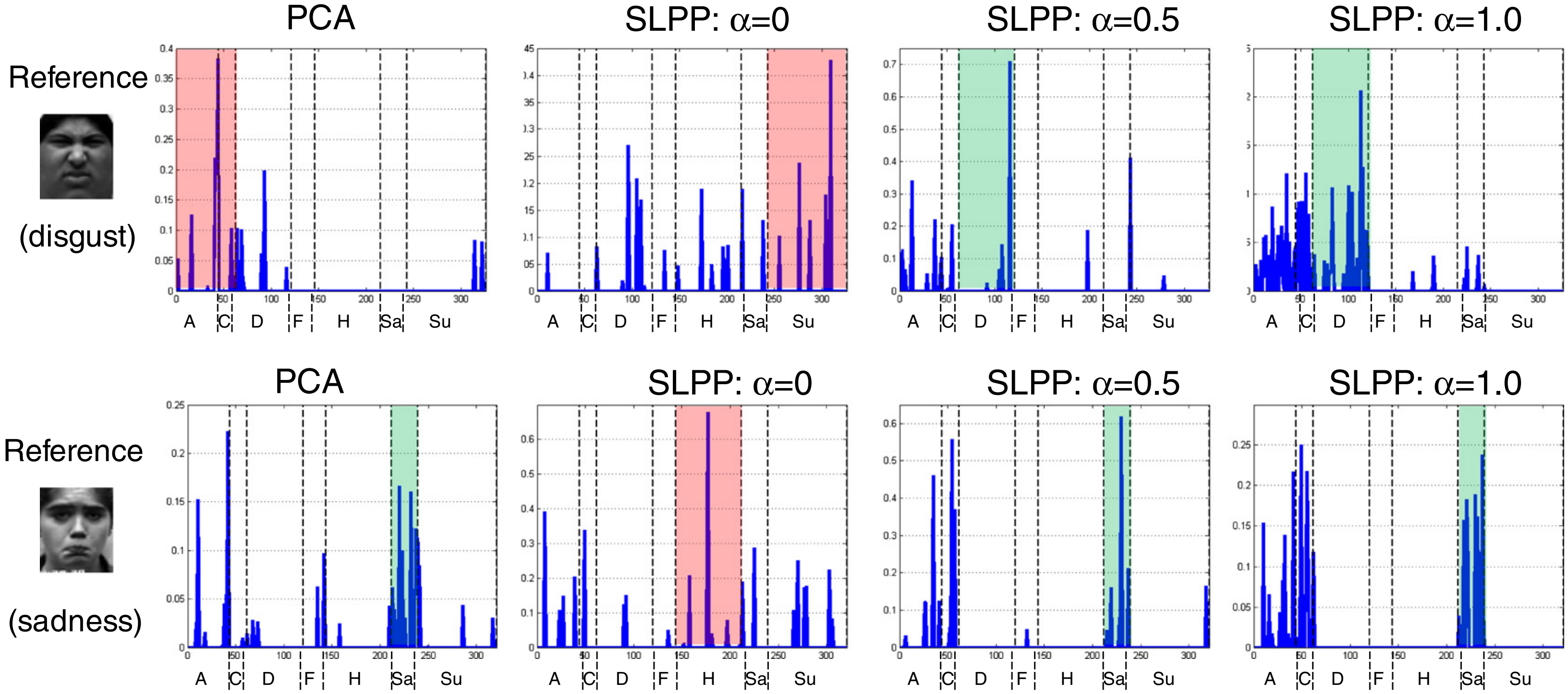


Fig. 5. Sample faces (disgust in top row and sadness in bottom row) along with their top $\hat{a}$ coefficients using PCA dimensionality reduction, LPP, SLPP, and LDA LPP. Coefficients are grouped by CK+ categories of Anger (A), Contempt (C), Disgust (D), Fear (F), Happy (H), Sadness (Sa), and Surprise (Su). Shaded rectangle shows classification made via Eq. (4.6)—red is incorrect, while green is correct.

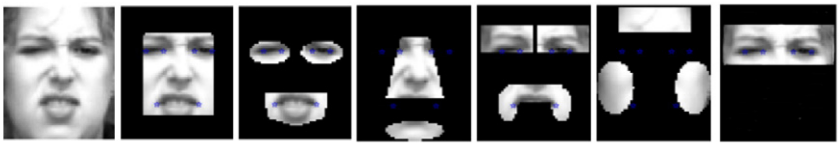


Fig. 6. Eleven facial areas studied: extended face, face, eyes, mouth, nose, chin, eyebrows, mustache, forehead, cheeks, eyereg.

sets of dictionary elements. The final classification includes simultaneous predictions from multiple regions, where the following inference model determines the predicted class:

$$\hat{c} = \arg\max_{c=1\ldots z} \sum P_f c_f * I[c = c^*] \quad (4.7)$$

where the summation is done over all included face regions, each region predicting one of z discrete classifications. P_f is the prior probability of accuracy for the particular face region on the training set (see Tables 4 and 5 for example P_f values), c_f^* is the predicted class from the particular face region, and $I[c = c^*]$ is the indicator function that returns 1 when $c = c^*$, otherwise 0. As such, Eq. (4.7) allows each facial region a weighted vote for a particular class, $c_1 \ldots c_z$, and the test face is assigned to the class with the most weighted votes.

We employ a feature harvesting process for local feature selection, as it is entirely conceivable that different regions of the face may be better described by different types of pixel processing. Six types of pixel processing are demonstrated: luminance, edge intensity based upon the magnitude of the high pass image, edge direction based upon the phase of the high pass image, Local Binary Patterns (LBP) [40,41], Gabor filters [42,43], and Local Phase Quantization (LPQ) [44] pixels. For each region, mean normalization, $x' = x / \mu$, consistently gave superior results over other normalization techniques.

Because none of the training faces have occlusions, the SR framework with reconstruction error is ideally suited for detecting occlusions over portions of a face. If a test face is partially occluded, the reconstruction error will be large in the occluded regions. For example, it would be difficult to represent an occluded test sample mouth region (e.g. with a hand in front of it) using the training mouth regions, since none of the training samples have occluded mouth regions. The MSR method searches the eye and mouth regions first. If the reconstruction error is greater than κ, the algorithm omits those regions from the statistical merging in Eq. (4.7). Thus, if the mouth is occluded, the nose and eye regions can still be used for final classification. This instructs the algorithm to ignore occluded regions of the face and concentrate on areas with small reconstruction errors. If no occlusions are detected, all face regions are eligible for usage in Eq. (4.7).

5. Experimental results

5.1. Face datasets

The classification results reported in this paper use the Cohn–Kanade (CK) [30] dataset, the extended Cohn–Kanade (CK+) [31] dataset, a selection of portrayals from the Geneva Multimodal Emotion Portrayals (GEMEP) corpus [45] used for the Facial Expression Recognition and Analysis Challenge (FERA2011), referred to as the GEMEP-FERA [32] dataset, and the Labeled Faces in the Wild (LFW) [29] dataset.

The CK [30] dataset contains 92 subjects in 229 expression sequences. Each expression sequence is a short video clip from neutral to fully articulated expression. The dataset contains the six universal expressions of anger, disgust, fear, happiness, sadness, and surprise.

The CK+ [31] dataset is an expanded version of the Cohn–Kanade (CK) [30] dataset along with specific testing protocols in an effort to standardize facial expression benchmarking efforts. The CK+ dataset contains 118 subjects in 327 expression sequences. In addition to the six universal expressions, the CK+ dataset includes the contempt expression. Fig. 7 shows samples of the seven expressions from the CK+ dataset. One unique aspect of the CK+ dataset is that any expression sequence that appeared to be fake or unidentifiable by a panel of judges was excluded. The CK+ benchmarking protocol stipulates a leave-one-subject-out cross-validation yielding 118 different training and testing sets. Only the last frame of each sequence is used for each expression.

The GEMEP-FERA dataset was introduced at the 9th IEEE International Conference on Automatic Face and Gesture Recognition in 2011. Also referred to as FERA2011, it was organized to help researchers compare and benchmark facial action unit (AU) and emotion recognition classification algorithms. The emotion portion of the dataset consists of 10 actors exhibiting the five emotions of anger, fear, joy, relief, and sadness. The training set contains 7 actors over 155 videos. The test set contains 6 actors (half of which were not present in the test set) over 134 videos. Training video sequences varied from 20 to 128 frames, with a median of 56 frames. Testing video sequences varied from 26 to 150 frames, with a median of 51 frames. During most video sequences, subjects uttered one of two pseudo-linguistic phoneme sequences. For the remaining sequences, subjects uttered the sustained vowel 'aaa'.

The LFW dataset contains 13,233 images of 5749 unique faces. Each face has been downloaded from the web, and as such is representative of an unconstrained natural pose. Although all faces were found by the Viola–Jones face detector, considerable variability in pose, expression, fidelity, and occlusions exist. The LFW website only releases ground truth for identity, but we have collected ground truth for race {Asian, Black, Caucasian, Hispanic, Indian, other}, gender {female, male}, spectacles {none, reading, eye, sun}, and facial hair {none, beard and mustache, beard only, mustache only, goatee, other}. In the race and facial hair classification results, faces classified as other (<5% of samples), were removed from the analysis. Although gender and glasses are generally agreeable by most judges, there is considerable disagreement over race and facial hair on many samples.

For each face used in this paper, ASM localized eye and mouth corner points define an affine projection matrix to a reference canonical size and position. Projected faces are cropped to 60 × 51 pixels, masked,

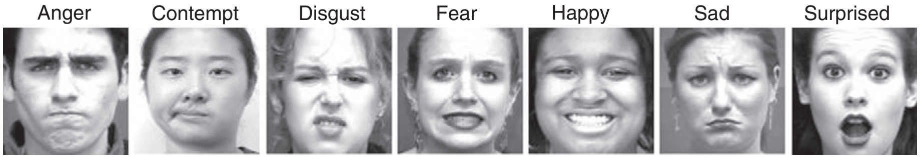


Fig. 7. Example of 7 facial expressions from the CK+ dataset.

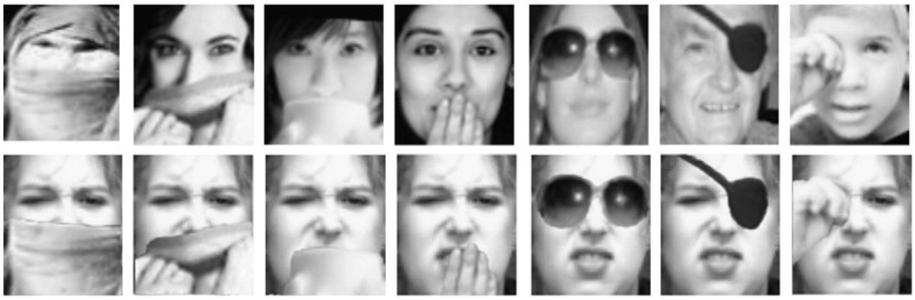


Fig. 8. Examples of eye and mouth occlusions: Original images on top, modified CK+ images on bottom.

and then written in lexicographic ordering in preparation for dimensionality reduction.

5.2. Occlusion dataset

To mimic real-world occlusion problems, a set of over 250 eye or mouth occluded images were downloaded from the internet. Each image was affine mapped to the canonical 60 × 51 pixel image, where the occlusion was manually trimmed in Photoshop to create an occlusion mask. In this fashion, any occlusion mask can be overlaid on top of any test image. Fig. 8 shows a sample CK+ image modified with various eye and mouth occlusions.

5.3. Support of dimensionality reduction and sparse representation methods

To demonstrate the capabilities and limitations of the MSR technique and understand its parameter selection, the 7-class CK+ expression dataset, 2-class LFW gender dataset, 5-class LFW race dataset, and 10-class LFW gender–race datasets are contrasted. The 10-class race–gender LFW dataset assigns each face as one of the 5 races × 2 gender types described in the dataset section. The CK+ results are leave-one-subject-out and the LFW results use the training set for training and (the non-overlapping) test set for testing. Tables 1a, 1b, 1c and 1d shows the classification accuracy of the four ℓ^1 classification strategies discussed in Section 4.2 under various dimensionality reduction techniques. Dimensionality reduction techniques included no dimensionality reduction (No DR), PCA, Gaussian kernel LPP (LPP), SLPP, $\alpha = 0.5$, and LDA LPP. Non-negative ℓ^1 SR coefficients were used as input to the classification method. Extended faces, luminance processing, with mean normalization are used. The $\mathbf{R}^d$ column is the number of dimensions after dimensionality reduction, and time column is the total time for processing (lower numbers are better).

Tables 1a, 1b, 1c and 1d illustrates the beneficial performance of the minimum reconstruction error [12] for facial classification. With regards to the dimensionality reduction methods, the No DR case skips the dimensionality reduction step, passing all 60 × 51 pixels into each of the four SR models. For the PCA case, the eigenvectors corresponding to 99% of the variance were kept. For the LPP cases, the top 30 dimensions were used. The addition of supervision to the LPP framework via Eq. (3.5) minimizes coefficient contamination by forcing samples to be reconstructed from dictionary elements of identical class. This minimizes the reconstruction error for that class, enabling dramatic classification accuracy improvements. The lower dimensions resulting from LPP speed up the resulting classification dramatically. The time column in Tables 1a, 1b, 1c and 1d is the total time in seconds to process the entire dataset, using appropriate cross-validation, on a quad-core I-7 computer inside the Matlab programming environment. The MP %Acc, nonZ %Acc, Energy %Acc, and RE %Acc columns represent the SR classification accuracy using the methods of Max Peak, Most nonzero, Energy ($\boldsymbol{\Sigma}\hat{a}$), and reconstruction error as defined in Section 4.2.

Continuing with the same four datasets, Table 2 compares multiclass linear SVM to four SR methods (LARS ℓ^1, NMF LARS ℓ^1, regularized ℓ^1, NMF regularized ℓ^1). The LARS methods are based upon SparseLab 2.1 (http://sparselab.stanford.edu/), and the regularized methods are based upon SLEP 4.0 (http://www.public.asu.edu/~jye02/Software/SLEP/) toolkits. All tests were done with SLPP ($\alpha = 0.95$) dimensionality reduction. The SR methods used the minimum reconstruction error model to assign the final class. Regularized ℓ^1, although

Table 1a
Seven class expression accuracy on the CK+ dataset. Comparison of dimensionality reduction method vs. SR coefficient classification method.

Method	$\mathbf{R}^d$	Time	MP %Acc	nonZ %Acc	Energy %Acc	RE %Acc
No DR	3060	643	54.9	49.1	62.2	62.0
PCA	199	1180	58.5	70.1	73.7	69.2
LPP	6	192	21.7	21.1	22.9	23.7
SLPP, $\alpha = 0.95$	6	211	62.5	72.5	79.4	84.1
LDA LPP	6	192	65.2	82.6	81.2	**85.2**

Table 1b
Two class gender accuracy on the LFW dataset. Comparison of dimensionality reduction method vs. SR coefficient classification method.

Method	$\mathbf{R}^d$	Time	MP %Acc	nonZ %Acc	Energy %Acc	RE %Acc
No DR	3060	760	76.7	77.3	79.9	80.2
PCA	411	92	75.6	77.5	79.8	82.8
LPP	30	29	62.2	76.4	76.0	74.9
SLPP, $\alpha = 0.95$	30	31	80.8	84.9	85.9	**88.9**
LDA LPP	1	29	51.9	51.9	51.9	51.9

Table 1c
Five class race accuracy on the LFW dataset. Comparison of dimensionality reduction method vs. SR coefficient classification method.

Method	$\mathbf{R}^d$	Time	MP %Acc	nonZ %Acc	Energy %Acc	RE %Acc
No DR	3060	700	53.9	81.0	81.8	81.8
PCA	405	80	67.1	81.8	83.4	83.7
LPP	30	33	50.8	80.7	78.7	77.9
SLPP, $\alpha = 0.95$	30	34	73.0	82.9	83.7	**86.0**
LDA LPP	4	29	32.2	47.6	43.7	40.7

Table 1d
Ten class race–gender accuracy on the LFW dataset. Comparison of dimensionality reduction method vs. SR coefficient classification method.

Method	$\mathbf{R}^d$	Time	MP %Acc	nonZ %Acc	Energy %Acc	RE %Acc
No DR	3060	724	40.9	62.6	64.9	65.1
PCA	405	80	50.4	63.4	66.7	69.1
LPP	30	33	29.7	59.2	55.9	53.4
SLPP, $\alpha = 0.95$	30	33	56.4	71.1	72.9	**76.1**
LDA LPP	9	30	44.5	64.1	60.9	55.4

Table 2
Seven class expression accuracy on the CK+ dataset using linear SVM vs. SR methods.

	Linear SVM	LARS ℓ^1 SR	NMF LARS ℓ^1 NMF SR	regularized ℓ^1 SR	NMF regularized ℓ^1 SR
7 class CK+	82.5	79.4	82.8	79.5	**84.1**
2 class gender LFW	**89.5**	86.7	85.2	88.0	88.9
5 class race LFW	84.6	81.0	64.6	84.9	**86.2**
10 class race–gender LFW	75.7	65.7	43.4	75.2	**76.3**

approximately 25% slower than LARS, is 13% more accurate. The NMF regularized ℓ^1 techniques are convincingly better for all categories except for the 2 class gender LFW dataset.

The accuracy of the SR methods is strongly influenced by the regularization parameter λ, and the adjacency matrix W blend parameter, α. Tables 3a, 3b, 3c and 3d continues with the same four datasets, varying λ from 0 (no regularization) to 1 (full regularization), and varying α from 0 (full heat kernel or only local topology) to 1 (full supervised LDA adjacency matrix). The blended adjacency matrix method is superior to the LDA or Gaussian adjacency matrix for the three natural LFW datasets, but not the CK+ posed expression dataset. For the CK+ dataset, large differences between exaggerated expressions are linearly separable using a linear LDA method and the local topology offered by the Gaussian kernel offers little value. As long as some amount of the LDA kernel is utilized, results are surprisingly flat. Further, because the CK+ dataset has exaggerated sample to sample differences, the regularization parameter λ has little effect. For the LFW datasets, too much regularization generally prevents important $\hat{a}$ coefficients from being passed into the reconstruction model classifier, while too little regularization overfits to the training set.

Because posed datasets with large separations between classes are tolerant to large changes in λ and α, we can concentrate on natural datasets such as LFW. Experimentation by the authors has found that constraining $0.05 \leq \lambda \leq 0.25$, and $0.25 \leq \alpha \leq 0.95$ is generally sufficient to meet a cross spectrum of classification needs. Tougher classification problems require less regularization (which essentially over smoothes the model), and more local topology (lower α). Further, when k, the number of classes is high, the rank of W is high, and the benefit of using the Gaussian heat kernel diminishes.

In our research we have consistently found:

- The improved performance of the reconstruction error classifier over the max peak, most non-zero, and energy classifiers was found in just about every test case we have done.
- The improved performance of the nonnegative (ℓ^1 NMF) SR method over ℓ^1 SR is consistent with our previous classification studies that ultimately use the sparse coefficients for classification purposes.

Table 3a
Seven class expression accuracy on the CK+ dataset. Comparison of SLPP blend parameter α to regularization parameter λ.

		CK + Expression, NMF = 1													
		← Less regularization						λ			More regularization →				
	α \ λ	0.0000	0.0001	0.0010	0.0050	0.0250	0.0500	0.0750	0.1000	0.2500	0.5000	0.9000	0.9500	0.9900	1.0000
← (LDA) α (Heat) →	0.000	23.3	23.1	22.8	22.6	23.1	23.2	23.6	23.7	24.3	23.4	22.3	22.3	21.7	22.9
	0.010	86.1	86.1	86.1	86.1	86.1	86.1	86.1	85.9	85.1	84.9	84.3	84.3	84.3	84.3
	0.100	83.8	83.8	83.8	84.6	84.8	85.6	84.8	84.6	84.1	85.2	82.2	82.2	82.2	81.4
	0.250	84.1	84.1	84.1	84.1	84.1	84.1	84.1	84.1	84.1	84.1	82.9	83.2	83.2	83.2
	0.500	84.8	84.8	84.8	84.8	84.5	84.8	84.8	84.8	83.6	82.9	82.3	82.3	81.8	81.2
	0.750	83.5	83.5	82.9	82.9	83.3	84.1	84.3	84.6	84.6	84.6	85.1	85.1	85.1	85.1
	0.850	84.4	85.0	84.4	85.0	85.0	85.3	85.0	85.0	85.0	83.4	83.2	83.2	83.2	83.2
	0.900	84.6	85.4	84.6	85.4	84.9	85.4	85.4	84.9	85.4	84.3	81.9	81.9	81.9	81.9
	0.950	84.6	84.6	84.6	84.6	84.6	84.6	84.6	84.1	84.6	85.7	84.4	84.4	84.4	84.4
	0.990	83.7	83.7	83.7	83.7	83.7	83.7	83.7	84.0	83.6	85.3	84.5	85.0	85.0	85.0
	0.999	82.3	82.1	82.9	82.9	82.9	82.9	82.9	82.4	83.4	84.5	84.4	85.0	85.0	85.0
	1.000	83.5	83.5	83.5	83.5	84.3	85.2	85.2	85.2	84.4	83.3	81.3	79.7	79.7	79.7

Table 3b
Two class gender accuracy on the LFW dataset. Comparison of SLPP blend parameter α to regularization parameter λ.

		LFW, gender, NMF = 1													
		← Less regularization						λ			More regularization →				
	α \ λ	0.0000	0.0001	0.0010	0.0050	0.0250	0.0500	0.0750	0.1000	0.2500	0.5000	0.9000	0.9500	0.9900	1.0000
← (LDA) α (Heat) →	0.000	76.5	76.5	76.5	76.5	76.4	76.0	75.5	74.9	73.8	66.5	47.4	47.0	45.0	44.8
	0.010	86.6	86.7	86.8	86.6	85.9	84.9	84.2	83.4	77.2	68.6	56.1	54.5	53.1	52.9
	0.100	88.2	88.1	88.2	88.1	88.5	89.1	89.9	90.1	88.1	82.4	72.0	70.4	68.8	68.7
	0.250	88.3	88.4	88.4	88.5	88.8	89.4	89.7	89.8	88.9	84.6	69.4	67.2	65.0	64.3
	0.500	87.2	87.1	87.2	87.4	87.9	88.2	88.5	89.0	90.0	87.7	73.1	70.1	68.3	67.6
	0.750	86.5	86.5	86.6	86.5	87.3	87.7	87.9	88.5	89.9	86.6	71.8	69.2	67.0	66.9
	0.850	86.4	86.3	86.5	86.4	87.2	87.6	87.6	88.4	90.1	86.7	72.1	68.8	66.8	66.3
	0.900	86.0	86.1	86.0	86.0	86.9	88.1	88.6	88.8	89.1	86.7	71.4	69.8	66.9	65.9
	0.950	85.6	85.6	85.6	85.9	87.6	88.0	88.2	88.9	89.6	86.9	72.8	70.0	67.7	67.0
	0.990	86.2	86.2	86.4	86.2	87.2	87.7	88.2	88.9	89.5	87.2	73.0	69.2	66.4	65.8
	0.999	86.3	86.4	86.3	86.4	87.4	87.7	88.2	89.1	89.5	87.4	72.9	68.8	66.7	66.1
	1.000	62.5	62.5	62.5	61.8	59.3	56.9	54.6	51.9	36.3	27.5	24.2	23.8	23.8	23.8

Table 3c
Five class race accuracy on the LFW dataset. Comparison of SLPP blend parameter α to regularization parameter λ.

		LFW, race NM F= 1													
		← Less regularization						λ		More regularization →					
	$\alpha\backslash\lambda$	0.0000	0.0001	0.0010	0.0050	0.0250	0.0500	0.0750	0.1000	0.2500	0.5000	0.9000	0.9500	0.9900	1.0000
↓ (LDA) α (Heat) ↑	0.000	81.3	81.3	81.3	81.3	80.7	79.3	78.5	77.9	69.6	56.4	29.9	27.0	25.1	24.5
	0.010	85.4	85.3	85.7	85.3	84.8	83.9	83.2	82.2	74.6	61.0	44.3	42.7	41.0	40.6
	0.100	86.5	86.6	86.4	86.4	86.0	85.1	84.8	83.7	77.2	64.6	41.6	39.3	36.9	36.8
	0.250	85.9	85.8	86.1	85.9	86.2	86.1	85.9	85.4	84.0	77.7	55.0	52.6	50.5	49.8
	0.500	86.1	85.8	86.0	86.1	86.7	87.3	87.4	87.5	84.9	77.1	54.9	51.6	49.6	49.1
	0.750	86.0	86.0	86.0	86.0	86.5	86.9	86.9	86.9	85.3	76.0	53.1	48.6	45.5	44.9
	0.850	85.7	85.7	85.5	85.4	85.9	86.4	86.5	86.4	85.1	77.3	51.9	48.7	46.4	45.5
	0.900	85.7	85.7	85.7	85.7	85.8	86.1	86.2	86.0	86.0	76.9	54.4	50.7	47.3	46.4
	0.950	85.5	85.5	85.5	85.7	85.9	86.3	86.0	86.0	84.9	76.8	53.1	50.0	47.4	46.8
	0.990	85.8	85.8	85.8	85.8	86.0	86.2	86.3	86.2	85.4	77.0	53.4	49.1	46.0	45.3
	0.999	85.8	85.8	85.8	85.8	86.0	86.1	86.4	86.4	85.5	76.9	52.7	48.9	46.1	45.5
	1.000	77.2	76.7	69.2	60.4	52.1	46.8	43.3	40.7	25.5	15.1	11.6	11.3	11.2	11.2

- The single most important factor towards reducing coefficient contamination is to use some sort of subspace clustering. We have found that supervised methods such as LDA and SLPP outperform unsupervised methods only because unsupervised methods indiscriminately cluster by similarities which often differ from the attribute being classified upon. We have tried many supervised and unsupervised dimensionality reduction techniques, and SLPP is consistently one of the best.
- The improved performance of the nonnegative SR method over the linear SVM classification justifies the inclusion of SR for facial classification problems.

In summary, the combination of Manifold learning with Sparse Representations (MSR) has been found to yield the best facial understanding classification accuracy at a lower computational cost. Specifically, the MSR technique proposed takes normalized images, passed through SLPP, followed by nonnegative SR, and finally classified using the minimum reconstruction error.

5.4. Choice of pixel processing and facial parts selection

Table 4 displays various types of pixel processing over eleven facial regions for the CK+ dataset with no occlusions. The six types of pixel processing include luminance image, edge magnitude, edge phase, $\text{LBP}_{8,1}^{u2}$ pixels, Gabor filters, and LPQ. For each region and pixel processing type, normalized pixels are passed into SLPP dimensionality reduction, where ℓ^1 coefficients are classified according to the reconstruction model using Eq. (4.6). The classification accuracies in Table 4 are used as the P_f values for Eq. (4.7).

The formation of facial expressions involves multiple areas of the face simultaneously. As such, the top features for our Manifold based Sparse Representation (MSR) method are the extended face (extface) and face regions. Surprisingly, for the CK dataset, using only the mouth and mustache regions sacrifice only 10–15 percentage points of accuracy. This is followed by the nose and eyereg which sacrifice 20 and 30 percentage points respectively. The forehead, eyes, eyebrows, chin, and cheek regions were found to be less valuable.

With regards to pixel processing, Gabor processing slightly outperformed luminance and edge magnitude. The edge phase and LBP was 8 and 13 percentage points behind Gabor processing on average. Different regions and pixel processing can be combined for increased accuracy. For example, if the top five feature regions are selected from Table 4, Eq. (4.7) improves the overall classification accuracy to 91.4%.

Table 5 contrasts the results from the CK+ posed dataset in Table 4 with the GEMEP-FERA natural dataset. Table 5 was created by applying the MSR methods to the GEMEP-FERA training set, as no ground truth was available for the test set. A comparison of Tables 5 to 4 (along with the CK+ and GEMEP-FERA images in Fig. 9) show evidence

Table 3d
Ten class race–gender accuracy on the LFW dataset. Comparison of SLPP blend parameter α to regularization parameter λ.

		LFW, race-gender, NMF = 1													
		<-- Less regularization						λ		More regularization -->					
	$\alpha\backslash\lambda$	0.0000	0.0001	0.0010	0.0050	0.0250	0.0500	0.0750	0.1000	0.2500	0.5000	0.9000	0.9500	0.9900	1.0000
<-- (LDA) α (Heat) -->	0.000	61.0	61.0	61.0	60.9	59.1	56.6	54.8	53.4	46.7	36.4	17.7	16.8	15.4	15.0
	0.010	77.5	77.3	77.1	77.3	77.0	76.0	75.5	75.1	70.4	61.6	43.9	42.5	41.1	40.8
	0.100	74.7	74.8	74.8	74.6	75.3	75.7	75.5	76.2	75.1	68.7	54.1	51.6	50.3	50.1
	0.250	75.9	76.0	75.9	76.0	76.2	77.0	78.1	77.5	72.9	62.5	45.2	42.2	40.6	39.9
	0.500	75.8	75.6	75.6	75.9	76.6	76.6	76.8	76.6	73.2	64.2	45.6	42.8	41.8	41.0
	0.750	75.1	75.4	75.2	75.3	75.4	76.0	76.1	76.8	73.9	64.0	43.9	40.8	38.3	38.0
	0.850	75.4	75.2	75.3	75.6	75.5	75.8	76.0	76.4	73.8	64.5	43.7	40.7	39.1	38.9
	0.900	75.2	75.3	75.5	75.6	75.5	75.6	76.3	76.2	73.6	64.3	43.7	40.8	39.4	38.8
	0.950	74.8	74.8	74.9	75.3	75.6	76.1	76.4	76.1	73.8	65.3	43.8	41.1	39.1	38.9
	0.990	75.2	75.1	75.3	75.3	75.5	75.8	76.3	76.3	73.6	63.3	45.1	42.5	40.8	39.6
	0.999	75.3	75.1	75.2	75.6	75.4	75.7	76.4	76.4	74.1	63.7	45.2	42.3	40.3	40.1
	1.000	73.9	73.9	72.8	71.0	64.3	60.4	57.5	55.4	43.3	27.9	17.3	16.0	15.0	15.0

Table 4
CK+ dataset classification accuracy for facial regions across six different pixel processing techniques.

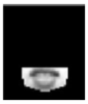
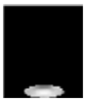
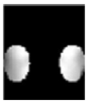
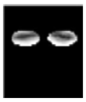
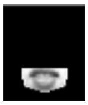
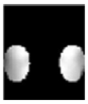
	 extface	 face	 nose	 must.	 mouth	 chin	 cheeks	 eyes	 eyebrow	 eye reg	 forehead
Lum	83.8	85.8	63.3	**76.6**	**81.0**	43.2	53.2	58.0	54.3	56.8	30.8
Mag	86.6	84.3	72.6	70.3	69.0	30.6	53.2	54.3	45.6	56.8	35.9
Phase	**88.0**	81.2	64.4	66.0	64.3	38.0	47.7	43.2	45.1	53.2	23.7
LBP	85.4	80.2	58.1	55.5	60.0	26.1	45.2	44.1	47.9	48.4	22.1
Gabor	83.6	**88.0**	**70.7**	**76.6**	79.7	**44.1**	**56.8**	**62.9**	**57.6**	**64.8**	**33.9**
LPQ	84.0	76.7	61.3	61.8	67.6	38.5	42.7	43.4	48.5	57.7	28.4

Table 5
MSR model applied to all 8865 frames of the GEMEP-FERA training dataset. Numbers shown are classification accuracy shown for sample facial regions across five different pixel processing techniques.

	 extface	 face	 nose	 must.	 mouth	 chin	 cheeks	 eyes	 eyebrow	 eye reg	 forehead
Lum	91.9	86.5	72.8	62.2	62.6	40.3	65.8	69.1	67.9	77.1	52.2
Mag	92.9	**89.0**	76.6	**68.7**	**65.5**	40.4	**70.7**	**81.5**	**85.4**	**90.4**	71.2
Phase	90.5	80.4	67.2	62.1	54.9	**45.3**	64.4	64.9	75.3	76.2	61.0
LBP	88.3	77.0	66.1	58.2	52.6	36.0	63.2	67.5	76.0	78.8	57.8
Gabor	**93.5**	87.9	**74.2**	61.7	61.0	44.2	67.4	66.9	74.2	80.0	**51.9**

towards the difference in facial region effectiveness for posed vs. natural datasets. Posed datasets emphasize the mouth region, while natural datasets emphasize the eye regions.

5.5. Benchmarking MSR performance for facial expression

We now compare the performance of the MSR technique to other works for the 6-class expression CK dataset, the 7-class expression CK+ dataset, and the 5-class emotion GEMEP-FERA dataset.

Table 6 compares our MSR approach to two other recently published methods that use SRs for facial expression. The first two entries of Table 6 show MSR with the top 1 and 5 features respectively. Zafeiriou et al. [16] used neutral subtracted frames, PCA dimensionality reduction, nonnegative matrix factorization, and most nonzero classification methods. Zhi et al. [46] used a nonnegative based SR without neutral frame subtraction using k-fold cross validation, which is less challenging and less realistic than leave-one-subject-out testing.

Using Eq. (4.7), MSR can automatically detect occluded regions. Table 7 shows the expression classification performance of the MSR technique on the CK dataset in the presence of eye and mouth occlusions. In place of the real-world occlusion dataset used in this paper, [16] used black box rectangles for occlusions to the eye region, and as before used neutral frame subtraction.

Table 8 compares leave-one-subject-out expression classification accuracy of our MSR method on the 7-class expression CK+ dataset to other recently published results. Lucey et al. [31] used neutral frame subtraction, AAM points, and face pixels along with a linear SVM classification model. Chew et al. [47] used normalized pixels from constrained local models along with linear SVM. Jain et al. [48] used AAM points, PCA, and linear SVM. The last two columns of Table 8 show the expression classification performance of the MSR technique on the CK+ dataset in the presence of simulated eye and mouth occlusions.

The MSR model was then applied to the GEMEP-FERA dataset. As this dataset is temporal in nature, the MSR model was applied to each frame of each test video sequence, and then majority voting determined final emotion classification for each video. Results returned from the FERA2011 organizers are done three ways: 1) Person dependent: subjects in the test set are in the training set; 2) Person independent: subjects in the test set are not in the training set; and 3) Overall.

As part of the FERA2011 challenge, 14 papers were submitted, including a baseline. The MSR method scored 98.5% on the person dependent, or 2nd place; 56.6% on the person independent test, or 10th place;

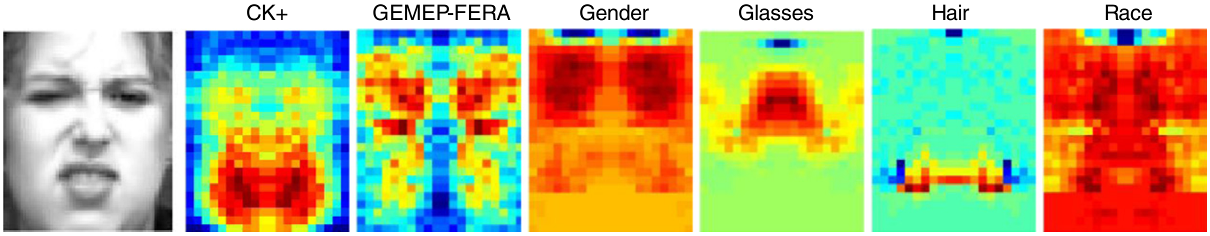


Fig. 9. Sample face and face importance demonstrating discriminative regions of the face for (left to right): expression on CK+ dataset; expression on the GEMEP-FERA dataset; gender, glasses, facial hair, and race classification on the LFW dataset.

Table 6
Comparison of expression classification performance of the MSR methods introduced in this paper to other SR methods using the CK dataset.

Method	Accuracy	Neutral subtract	Dim	Cross valid
MSR, 1 feature	92.0	N	6	Subj
MSR, 5 features	94.6	N	6	Subj
Zafeiriou [16]	81.0	Y	40	Subj
Zhi-35dim [46]	90.2	N	35	k-fold

Table 7
Expression classification accuracy on CK dataset without and with eye and mouth occlusions.

	MSR no occl.	MSR mouth	MSR eye	Zafeiriou [16] eye
Accuracy	94.6	64.5	84.4	78.7

and 73.5%, or 5th place overall. The poor performance on the person independent results is a common failure mode of sparse representations when there are insufficient training exemplars in Φ. With only 7 subjects in the training set, this is not surprising. The excellent performance on the person dependent results shows the power of SR methods with sufficiently populated dictionaries.

5.6. MSR performance for other facial attributes

To further test the robustness of the MSR technique to other types of facial understanding and classification problems, the facial attributes of gender, glasses, facial hair, and race were tested on the LFW dataset. Table 9 compares multi-class linear SVM and MSR classification accuracies with and without occlusions. Both the SVM and MSR results in Table 9 use SLPP dimensionality reduction. To ensure that both SVM and MSR were utilized to the best of their abilities, optimal values for α and λ values were independently solved and used for each. Radial basis function SVM's were tested, but offered no statistical improvement above linear SVM. In general, the MSR technique compares favorably to SVM with or without occlusions. When occlusions are not present, MSR outperforms SVM consistently. When occlusions have a strong impact on a classification, such as mouth occlusions with facial hair detection, the MSR technique performs significantly better. This demonstrates the power of the minimum reconstruction error working in conjunction with the statistical inference model. When occlusions don't have a strong impact on classification, such as facial hair with eye occlusion, SVM and MSR are comparable. While one can introduce occlusion detection into the SVM model, it is not as simple or elegant as the sparse model. For example, none of the training subjects had their hands over their mouth. If we now take a test sample with a hand over their mouth, the minimum reconstruction error is very high for all classes, a clear indication of occlusion. In the SVM paradigm, we will get a class prediction and can get confidence levels for each class. Unfortunately, there is no guarantee that confidence levels will be low for all classes in the presence of occlusions—in fact, we often map occluded samples far away from decision boundaries, which are falsely reported as high levels of confidence.

Table 8
Expression classification accuracy on CK+ dataset without and with eye and mouth occlusions.

	MSR no occl.	Lucey [31] no occl.	Chew [47] noocl.	Jain [48] no occl.	MSR mouth occl.	MSR eye occl.
Accuracy	**91.4**	83.3	74.4	84.1	60.5	70.9

Table 9
Gender, glasses, facial hair, race, and mixed race–gender classification accuracy on LFW dataset using SVM and the MSR technique without and with eye and mouth occlusions.

	SVM no occl.	MSR no occl.	SVM mouth occl.	MSR mouth occl.	SVM eye occl.	MSR eye occl.
Gender	89.6	**90.8**	89.8	**90.3**	80.5	**80.8**
Glasses	85.0	**87.9**	84.3	**85.0**	71.8	**79.6**
Hair	86.9	**87.7**	80.8	**85.6**	87.3	**87.4**
Race	85.1	**87.5**	**85.0**	84.3	78.7	**82.0**
Mixed	75.9	**78.5**	76.2	**76.6**	64.6	**66.5**
Avg.	84.5	**86.5**	83.2	**84.4**	76.6	**79.3**

5.7. Facial expression recognition on posed vs. natural datasets

The unique region based pixel processing methods used by MSR allow for any size facial region of any location to be evaluated for its effectiveness in facial classification. Referring to Fig. 9, the CK+ and GEMEP-FERA images show each area of the face color coded by its ability to classify facial expression. The mouth area, specifically the mouth corners, offers the most salient information for the CK+ dataset. The upper cheek and eye region offers the most salient information for the GEMEP-FERA dataset. The subjects in the GEMEP-FERA dataset were talking as they were exhibiting expressions. Talking not only varied the mouth position, but constrained the mouth from exhibiting exaggerated positions. As such, the mouth importance dropped significantly. Equally intriguing is the increase in upper cheek and eye classification capability. Concerning facial expression, by comparing Tables 4 and 5 along with Fig. 9, we can unequivocally state that posed datasets favor the mouth region, while natural datasets favor the eye, eyebrow, and upper cheek regions. Similar findings have been noted in the facial expression literature [49].

The gender, glasses, hair, and race plots in Fig. 9 show the importance of individual regions of the face on natural datasets with regards to their respective classification problems. Gender classification favors the eye and eyebrow regions. Glasses detection favors the nose, but not eye regions. Facial hair classification favors the mustache region, but not the chin region. The sides of the nose, the corners of the mouth, and the inner eye sockets play a prominent role in race classification.

6. Conclusions

This paper presents a novel method for robust facial understanding based on manifold learning and sparse representations. With the MSR paradigm introduced in this paper, sparse methods may prove to be beneficial to the facial understanding community. The introduced semi-supervised SLPP dimensionality reduction mitigates the coefficient contamination problem without the need for reference neutral frames. The passing of $\hat{a}$ coefficients from ℓ^1 NMF sparse representations into a reconstruction model enables highly accurate classifications. The introduced statistical inference model with occlusion detection further increases classification accuracy and offers robustness to eye and mouth occlusions. MSR facial expression results show that posed datasets over-emphasize the importance of the mouth region as compared to natural datasets which are more dependent on the upper cheek and eye regions. With the introduction of the MSR technique, sparse methods have been shown to be beneficial to the facial understanding community for tackling a variety of interesting facial classification problems.

Acknowledgements

This research was supported in part by the National Science Foundation through a Graduate Research Fellowship, Cisco, and the Center for Emerging and Innovative Sciences (CEIS), a NYSTAR-designated Center for Advanced Technology in New York State.

References

[1] B.A. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: a strategy employed by V1? Vis. Res. 37 (1997) 3311–3325.
[2] B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (1996) 607–609.
[3] D.L. Donoho, M. Elad, V.N. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise, IEEE Trans. Inf. Theory 52 (2006) 6–18.
[4] E.J. Candes, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inf. Theory 52 (2006) 489–509.
[5] H. Lee, A. Battle, R. Raina, A. Ng, “Efficient sparse coding algorithms,” presented at the Advances in Neural Information Processing Systems, 2006.
[6] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Statist. 32 (2004) 407–499.
[7] Z. Jiang, Z. Lin, L.S. Davis, Learning a discriminative dictionary for sparse coding via label consistent K-SVD, IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[8] M. Yang, L. Zhang, X. Feng, D. Zhang, Fisher discrimination dictionary learning for sparse representation, International Conference on Computer Vision, 2011.
[9] K. Jarrett, K. Kavukcuoglu, M.A. Ranzato, Y. LeCun, What is the best multi-stage architecture for object recognition? IEEE International Conference on Computer Vision, 2009.
[10] R. Rigamonti, M. Brown, V. Lepetit, Are sparse representations really relevant for image classification? IEEE Computer Vision and Pattern Recognition, 2011.
[11] A. Yang, J. Wright, Y. Ma, S. Sastry, Feature Selection in Face Recognition: A Sparse Representation Perspective, University of California at Berkeley, 2007.
[12] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, M. Yi, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 210–227.
[13] J. Sherrah, S. Gong, E.J. Ong, Face distributions in similarity space under varying head pose, Image Vision Comput. 19 (2001) 807–819.
[14] H. Junzhou, H. Xiaolei, D. Metaxas, Simultaneous image transformation and sparse representation recovery, IEEE Computer Vision and Pattern Recognition, 2008.
[15] G. Tzimiropoulos, S. Zafeiriou, M. Pantic, Sparse representations of image gradient orientations for visual recognition and tracking, IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2011.
[16] S. Zafeiriou, M. Petrou, Sparse representations for facial expressions recognition via l1 optimization, IEEE Conference on Computer Vision and Pattern Recognition-Workshops, 2010.
[17] L. Shuai-Shi, T. Yan-Tao, L. Dong, New research advances of facial expression recognition, Eighth International Conference on Machine Learning and Cybernetics (ICMLC), 2009.
[18] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognit. 36 (2003) 259–275.
[19] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 681–685.
[20] J. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vis. 60 (2004) 135–164.
[21] N. Kumar, P. Belhumeur, S. Nayar, FaceTracer: a search engine for large collections of images with faces, 10th European Conference on Computer Vision, 2008.
[22] A. Ghodsi, Dimensionality Reduction. A Short Tutorial, University of Waterloo, Ontario, Canada, 2006.
[23] L. Cayton, Algorithms for manifold learning, Tech Rep. CS2008-0923, University of California, San Diego, 2005.
[24] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[25] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[26] X. He, P. Niyogi, Locality preserving projections, Advances in Neural Information Processing Systems, 16, 2003.
[27] E. Elhamifar, R. Vidal, Sparse subspace clustering, IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2009.
[28] R. Ptucha, G. Tsagkatakis, A. Savakis, Manifold based sparse representation for robust expression recognition without neutral subtraction, IEEE International Conference on Computer Vision-Workshops, 2011.
[29] G. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, University of Massachusetts, Amherst, 2007.
[30] T. Kanade, J.F. Cohn, T. Yingli, Comprehensive database for facial expression analysis, International Conference on Automatic Face and Gesture Recognition, 2000.
[31] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The Extended Cohn–Kanade Dataset (CK+): a complete dataset for action unit and emotion-specified expression, IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2010.
[32] M. Valstar, B. Jiang, M. Mehu, M. Pantic, K.R. Scherer, The first facial expression recognition and analysis challenge, Face and Gesture Recognition, 2011.
[33] Y. Bengio, Learn. Deep Archit. AI 2 (1) (2009).
[34] M. Belkin, P. Niyogi, Laplacian Eigenmaps and spectral techniques for embedding and clustering, Adv. Neural Inf. Process. Syst. 14 (2001).
[35] D. Cai, X. He, J. Han, Document clustering using locality preserving indexing, IEEE Trans. Knowl. Data Eng. 17 (2005) 1624–1637.
[36] C. Xian-Fa, W. Gui-Hua, W. Jia, L. Jie, Enhanced supervised locality preserving projections for face recognition, International Conference on Machine Learning and Cybernetics, 2011.
[37] Y.C. Pati, R. Rezaiifar, P.S. Krishnaprasad, Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, Proceedings of 27th Asilomar Conference on Signals, Systems and Computers, 1993.
[38] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, D.J. Heeger, Shiftable multiscale transforms, IEEE Trans. Inf. Theory 38 (1992) 587–607.
[39] M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (2006) 4311–4322.
[40] X. Feng, M. Pietikainen, A. Hadid, Facial expression recognition based on local binary patterns, Pattern Recognit Image Anal. 17 (2007) 592–598.
[41] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on Local Binary Patterns: a comprehensive study, Image Vision Comput. 27 (2009) 803–816.
[42] T. Serre, L. Wolf, T. Poggio, Object recognition with features inspired by visual cortex, IEEE Conference on Computer Vision and Pattern Recognition, 2005.
[43] I. Buciu, C. kotropoulos, I. Pitas, ICA and Gabor representation for facial expression recognition, Proceedings of International Conference on Image Processing, 2003.
[44] B. Jiang, M.F. Valstar, M. Pantic, Action unit detection using sparse appearance descriptors in space–time video volumes, Face and Gesture Recognition, 2011.
[45] T. Banziger and K.R. Scherer, “Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) Corpus,” in Blueprint for Affective Computing: A Sourcebook, ed Oxford: Oxford University Press, 2010, pp. 271–294.
[46] Z. Ruicong, R. Qiuqi, Discriminant sparse nonnegative matrix factorization, IEEE International Conference on Multimedia and Expo, 2009.
[47] S. Chew, P. Lucey, S. Lucey, J. Saragih, J. Cohn, S. Sridharan, Person-independent facial expression detection using constrained local models, Automatic Face and Gesture Recognition, 2011.
[48] S. Jain, C. Hu, J.K. Aggarwal, Facial expression recognition with temporal modeling of shapes, IEEE International Conference on Computer Vision Workshops, 2011.
[49] T. Pfister, L. Xiaobai, Z. Guoying, M. Pietikainen, Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework, IEEE International Conference on Computer Vision Workshops, 2011.