A Compressive Sensing Approach for Expression-Invariant Face Recognition

Pradeep Nagesh and Baoxin Li

Dept. of Computer Science & Engineering
Arizona State University, Tempe, AZ 85287, USA
{pnagesh, baoxin.li}@asu.edu

Abstract

We propose a novel technique based on compressive sensing for expression-invariant face recognition. We view the different images of the same subject as an ensemble of intercorrelated signals and assume that changes due to variation in expressions are sparse with respect to the whole image. We exploit this sparsity using distributed compressive sensing theory, which enables us to grossly represent the training images of a given subject by only two feature images: one that captures the holistic (common) features of the face, and the other that captures the different expressions in all training samples. We show that a new test image of a subject can be fairly well approximated using only the two feature images from the same subject. Hence we can drastically reduce the storage space and operational dimensionality by keeping only these two feature images or their random measurements. Based on this, we design an efficient expression-invariant classifier. Furthermore, we show that substantially low dimensional versions of the training features, such as (i) ones extracted from critically-downsampled training images, or (ii) low-dimensional random projection of original feature images, still have sufficient information for good classification. Extensive experiments with publically-available databases show that, on average, our approach performs better than the state-of-the-art despite using only such super-compact feature representation.

1. Introduction

Face recognition (FR) has been a highly active research area for many years. A typical approach involves two tasks: feature extraction and classification. Commonly-used feature extraction methods include subspace techniques such as *principle component analysis* (PCA or eigenface), *independent component analysis* (ICA), *linear discriminant analysis* (LDA or fisherface) and so on [1, 2]. With features extracted, classifiers based on techniques such as *nearest neighbor* and *support vector machines* can then be used to perform recognition. The above feature extraction methods are well-understood and in a sense have reached their maturity. Researchers are now looking for different methods and theories to address the persisting challenges in FR like expression, illumination and pose variation, and dimensionality reduction, etc. Reducing the space complexity and in particular the operational dimensionality of the classifier is vital for practical applications involving large databases.

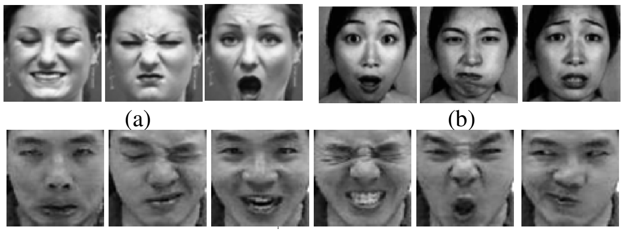
The recently-emerged Compressive Sensing (CS) theory [6,10,12-16], while originally intended to address signal sensing and coding problems, has shown tremendous potential for other problems like pattern representation and recognition [3,4], often beating the conventional techniques. In this paper we propose a new technique for face feature extraction and classification, based on the CS theory. We focus on addressing expression variation in FR. Expression-invariant FR is a challenging task owing to complex and varied nature of facial expressions. Some sample face images are shown in Fig. 1 to illustrate the complexity of the problem. Our method relies on distributed CS and joint sparsity models (JSM) [5, 10]. The JSM was originally proposed for efficient coding of multiple inter-correlated signals. In our work, we formulate the JSM from a "representation" perspective so that it can be readily applied to computer vision problems requiring compact representation of multiple correlated images such as instances of the same face in the context of FR, which is our focus of discussion in this paper. Further, we design feature extraction and classification algorithms based on the formulation. Unlike existing FR work based on sparse representation (e.g., [3]), the proposed approach has a natural and close knit with the CS theory and thus many potential benefits of CS apply (e.g., projecting the input image into ultra-low dimensions, as discussed in Section 4.2).

Specifically, we consider the training face images of a single subject as an ensemble of inter-correlated signals and propose a technique to represent each subject class with two feature images: (i) one that captures holistic or gross face features (the common component) and (ii) the other that captures mostly the unique features (like expressions) of all images in a single image (the gross innovation component). Then, we design a CS based reconstruction algorithm that can produce a close approximation of a new face image of the subject, using only the two training features. In particular, the algorithm

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

first produces an approximation of expressions in the new face image using the gross innovation feature and then uses this with the common component to reconstruct the given face image. A face classifier is designed based on the same principle, where the class of the test image is decided based on how well it can be approximated using the training features of labeled classes. Since we store only two feature images per subject (or their low dimensional measurements), we drastically reduce the training set storage space and the operational dimensionality of the classifier, compared with the sparse-representation-based algorithm of [3], while being able to achieve better performance than the state-of-art results reported therein. Further, our method is more robust in scenarios where only a few samples are available for training.

Section 2 reviews the background and related work. Section 3 presents our method for feature extraction based on JSM. A new classifier is designed and discussed in Section 4, followed by experimental results in Section 5. We conclude with discussion on future work in Section 6.



(a) (b)

(c)

Figure 1: Sample face images with expressions from (a) [9], (b) [8], and (c) [7].

2. Background and Related Work

In this section, we first briefly review basics of the CS theory, and then discuss one most recent work on FR based on sparse representation and CS.

According to the CS theory, if a signal $\boldsymbol{x} \in \mathbb{R}^N$ is *K-sparse,* with respect to a basis $\boldsymbol{\Psi} \in \mathbb{R}^{N\times N}$ (i.e in the expansion $\boldsymbol{\theta} = \boldsymbol{\Psi}^T\boldsymbol{x}$, there are only $K<N$ non-zero or significant coefficients), then $\boldsymbol{x}$ can be recovered by its measurement $\boldsymbol{y} \in \mathbb{R}^M$, $M<N$, obtained by projecting $\boldsymbol{x}$ onto a second basis $\boldsymbol{\Phi} \in \mathbb{R}^{M\times N}$, as long as (*i*) $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are incoherent and (*ii*) M is of the order $\geq K\log(N/K)$ [6,10,12-16]. Mathematically, if we write the measurement as $\boldsymbol{y}= \boldsymbol{\Phi x}$, $\boldsymbol{y} \in \mathbb{R}^M$, then the signal recovery can be done by convex l^1 optimization:

$$\hat{\boldsymbol{\theta}} = \arg\min\|\boldsymbol{\theta}\|_1 \quad s.t. \boldsymbol{y} = \boldsymbol{\Phi\Psi\theta} \qquad (1)$$

or $$\hat{\boldsymbol{\theta}} = \arg\min\|\boldsymbol{\theta}\|_1 \quad s.t. \|\boldsymbol{y} - \boldsymbol{\Phi\Psi\theta}\|_2 < \epsilon \qquad (2)$$

Eqn. (1) is the Basis Pursuit problem and Eqn. (2) is the Basis Pursuit Denoising problem, which is well suited in cases where the measurements are noisy. A popular approximation equivalent to (2) is the unconstrained version given by

$$\hat{\boldsymbol{\theta}} = \arg\min\{\, \tau\|\boldsymbol{\theta}\|_1 \;+\; 0.5 * \|\boldsymbol{y} - \boldsymbol{\Phi\Psi\theta}\|_2^2 \,\} \qquad (3)$$

There are efficient algorithms that use interior-point methods to solve the l^1 minimization of (1) and (2). One of the earlier implementations is l^1-magic [18] which recasts these problems as a second-order cone program and then applies the primal log-barrier approach. More recent interests are in sparse recovery algorithms solving the unconstrained optimization of (3), since it is much faster than directly solving (1) or (2). Gradient Projection for Sparse Reconstruction (GPSR) [11] is one such more recent algorithm, which is reported to outperform prior approaches [17].

Recently, an FR algorithm (called SRC) based on ideas of sparse representation and CS has been proposed [3], which appears to be able to handle changing expression and illumination. The work was enhanced by another paper [4] to handle pose variation. In the SRC algorithm, it is assumed that the whole set of training samples form a dictionary (each image is a base atom), and then the recognition problem is cast as one of discriminatively finding a sparse representation of the test image as a linear combination of training images by solving the optimization problem in (1), (2) or (3). While the SRC model demonstrates the power of harnessing sparsity in face recognition problem via l^1 minimization, it has some disadvantages. First, for accurate recognition, sufficiently large training images for each subject are needed. But in practice, only a few instances might be available for a few or even all of the subjects. Second, all training images (or their low dimensional versions) have to be stored and accessed during testing, and thus for a large training set, both the space complexity and the speed performance may pose as practical challenges.

Nevertheless, the comparison with other existing approaches in [3] suggests that the SRC algorithm is among the best and thus we treat it as the state-of-the-art and will use it as a bench mark in our study in this paper.

3. Face Feature Extraction and Training

The problem of recognition of an unknown object is to correctly identify the class to which it “belongs to”, using some information derived from labeled training samples belonging to K distinct classes. Here we refer to feature extraction as training. In this section, we propose a feature extraction algorithm based on the JSM CS recovery scheme [5, 10]. Our algorithm finds the common (holistic) and innovation components, with the latter corresponding to expressions, of all training images of class k. Since we use a sparsifying basis (like DCT), we term this as B-JSM feature extraction.

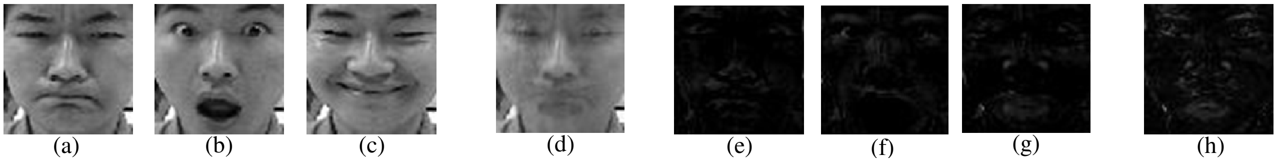


Figure 2: B-JSM feature extraction with DCT basis. (a), (b) and (c) are images of the same subject with different expressions (mean is added back); (d) The common component of (a), (b) and (c) or $\boldsymbol{z}^c$ with mean added. Images (e), (f) and (g) are the innovation components of (a), (b) and (c) respectively ($\boldsymbol{z}_1^i, \boldsymbol{z}_2^i, \boldsymbol{z}_3^i$); Image (h) is the sum of the innovation components (e), (f) and (g) (or $\boldsymbol{z}^A$). It serves as a global representation of the unique features of (a), (b) and (c) together. Note that the eye-brow and mouth regions are blurred in common component $\boldsymbol{z}^c$ in (d), where as these are captured as expression information in $\boldsymbol{z}^A$ shown in (h).

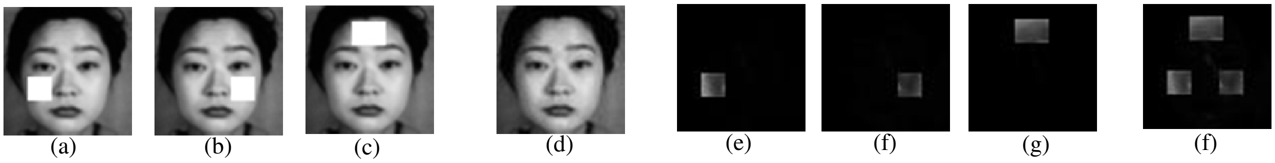


Figure 3: Illustration of common and innovation features using S-JSM. (a), (b) and (c) are the same images with added white patch (innovations). (d) is the obtained common component, in which even the skin texture at the patches is nearly retained; (e), (f) and (g) are the innovation components of (a), (b) and (c) respectively, each retaining an innovation as gray patches (white patch subtracted with the intensity of skin at regions of patches). (h) is the sum of the innovation components. It serves as a global representation of the innovations in all images. (For visual clarity, means are added back to in (a), (b), (c) and (d).)

3.1. B-JSM Feature Extraction

To present the idea, let us first assume a grayscale image represented as 1-D column vector $\boldsymbol{x} \in \mathbb{R}^N$, N=N_1xN_2. The extension of the presented idea to 2-D is straightforward. Since the features of interest lie in the textures but not the intensity of an image, we assume that $\boldsymbol{x}$ has its mean intensity subtracted. We assume that there are K distinct classes (i.e., subjects), with each class having J_k training images, k =1,2,…,K. Let the images of class k be represented as an ensemble $\{\boldsymbol{x}_{k,j}\}, j = 1, ..., J_k$, or simply $\{\boldsymbol{x}_{k,j}\}$. Jointly, such an ensemble can be represented as,

$$\boldsymbol{y_k} = \left[\boldsymbol{x}_{k,1} \ \ \boldsymbol{x}_{k,2} \ \ \ldots\ldots \ \ \boldsymbol{x}_{k,J_k}\right]^T \in \mathbb{R}^{N \times J_k} \qquad (4)$$

Noting that all signals in $\{\boldsymbol{x}_{k,j}\}$ for a given k are highly intercorrelated, we may represent the j-th training image of class k as the sum of a common component and an innovation component as follows,

$$\boldsymbol{x}_{k,j} = \boldsymbol{z}_k^c + \boldsymbol{z}_{k,j}^i \qquad (5)$$

Further, let $\boldsymbol{\Psi} \in \mathbb{R}^{N \times N}$ be the matrix representation of some orthonormal basis (e.g., DCT) that can sparsely represent the training images, so that coefficients $\boldsymbol{X}_{k,j} = \boldsymbol{\Psi x} \in \mathbb{R}^N$ of signal $\boldsymbol{x}$ can be written as,

$$\boldsymbol{X}_{k,j} = \boldsymbol{\theta}_k^c + \boldsymbol{\theta}_{k,j}^i = \boldsymbol{\Psi z}_k^c + \boldsymbol{\Psi z}_{k,j}^i; \ \ \boldsymbol{\theta}_k^c, \boldsymbol{\theta}_{k,j}^i \in \mathbb{R}^N \quad (6)$$

Here $\boldsymbol{\theta}_k^c$ is common to all the J_k training images of class k and $\boldsymbol{\theta}_{k,j}^i \ j = 1, \ldots J_k$, is unique to each image. Under this model, let the common and innovation components of class k be jointly represented by the vector

$$\boldsymbol{W}_k = \left[\boldsymbol{\theta}_k^c \ \ \boldsymbol{\theta}_{k,1}^i \ \ \boldsymbol{\theta}_{k,2}^i \ldots . \boldsymbol{\theta}_{k,J_k}^i\right]^T \in \mathbb{R}^{N \times (J_k+1)} \qquad (7)$$

Note that there might be more than one value of $\boldsymbol{z}_k^c$ or $\boldsymbol{\theta}_k^c$ satisfying (5) or (6), but the one we are interested in is the component $\boldsymbol{\theta}_k^c$ that is strictly derived from the common support in the ensemble $\{\boldsymbol{X}_{k,j}\}$ such that the vector $\boldsymbol{W}_k$ is the sparsest representation of $\{\boldsymbol{x}_{k,j}\}$ (Eqn.(4)) under the basis $\boldsymbol{\Psi}$. For highly correlated signals, naturally $\boldsymbol{\theta}_k^c$ would be strong and relatively denser compared to the very sparse innovations. From a feature extraction point of view, for FR with varying expression, this representation is useful since the common component $\boldsymbol{z}_k^c$ would retain all the gross common face features (holistic), while the innovation components $\boldsymbol{z}_{k,j}^i$ retains the unique features owing to changes in facial expressions. An example of such a representation is shown in Figure 2 and 3 and will be discussed in more detail later in this subsection.

In the distributed CS theory of [5, 10], the additive model of (5) was assumed in the sense of "jointly recovering" correlated signals from measurements, which would help reduce the number of measurements in coding of multi-sensor signals. In our case, essentially we are interested in forming a new representation of $\{\boldsymbol{x}_{k,j}\}$ given in (7) so as to use the common and innovation features for facilitating the FR task. From (4)-(7), we may write,

$$\boldsymbol{y}_k = \widetilde{\boldsymbol{\Psi}} \boldsymbol{W}_k \qquad (8)$$

where $\widetilde{\boldsymbol{\Psi}} \equiv [\![[\boldsymbol{I}_1] \ [\boldsymbol{I}_2]]\!]$ is formed by concatenating two matrices given by $\boldsymbol{I}_1 = [\boldsymbol{\Psi}^T \ \boldsymbol{\Psi}^T \ldots \ \boldsymbol{\Psi}^T]^T \in \mathbb{R}^{(J_k N) \times N}$ and $\boldsymbol{I}_2 = \mathrm{diag}(\boldsymbol{I}_1) \in \mathbb{R}^{(J_k N) \times (J_k N)}$, with $\mathrm{diag}(\boldsymbol{p})$ being a diagonal matrix whose diagonal elements are $\boldsymbol{p}_1, \boldsymbol{p}_2 \ldots \boldsymbol{p}_N$ in $\boldsymbol{p} = [\boldsymbol{p}_1 \ \boldsymbol{p}_2 \ldots \boldsymbol{p}_N]^T$. Note that $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ correspond to the common and innovation components respectively. The $\boldsymbol{W}_k$ vector can be found by solving the following l^1-minimization problem,

$$\boldsymbol{W}_k = \arg\min \|\boldsymbol{W}_k\|_1 \quad s.t.\ \boldsymbol{y}_k = \widetilde{\boldsymbol{\Psi}}\, \boldsymbol{W}_k$$
$$\text{or} \quad \boldsymbol{W}_k = \min\{\ \tau\|\boldsymbol{W}_k\|_1\ +\ 0.5 * \left\|\boldsymbol{y}_k - \widetilde{\boldsymbol{\Psi}}\, \boldsymbol{W}_k\right\|_2^2 \} \quad (9)$$

The spatial domain common and innovation components can be recovered by the inverse transformation as,

$$\boldsymbol{w}_k = \boldsymbol{\Lambda} \boldsymbol{W}_k \quad (10)$$

where $\boldsymbol{\Lambda} = \text{diag}([\boldsymbol{\Psi}^T\ \boldsymbol{\Psi}^T \ldots\ \boldsymbol{\Psi}^T]^T) \in \mathbb{R}^{(J_k N)\times(J_k N)}$ and $\boldsymbol{w}_k = \left[\boldsymbol{z}_k^c\ \ \boldsymbol{z}_{k,1}^i\ \ \ldots .\boldsymbol{z}_{k,J_k}^i\right]^T \in \mathbb{R}^{N\times(J_k+1)}$. For convenience and future reference, we represent the process described by the sequence of equations (8)-(10) for class *k* as

$$\text{B-JSM} \coloneqq (\{\boldsymbol{x}_{k,j}\}, j = 1, \ldots J_k) \rightarrow \left[\boldsymbol{z}_k^c\ \ \ldots .\boldsymbol{z}_{k,J_k}^i\right]^T \quad (11)$$

The last step in feature extraction is to form the gross innovation component denoted by $\boldsymbol{z}_k^A$, (the superscript *A* standing for "all") that can be computed as,

$$\boldsymbol{z}_k^A = \textstyle\sum_{j=1}^{J_k} \boldsymbol{z}_{k,j}^i \quad (12)$$

For each class *k*, we store only two feature images: the common component $\boldsymbol{z}_k^c$ and the gross innovation component $\boldsymbol{z}_k^A$ and discard the training and other innovation images. Hence there is a significant reduction in the total storage space compared with the SRC method of [3]. Further dimensionality reduction of feature space can be achieved by storing just sufficient random measurements of $\boldsymbol{z}_k^c$ and $\boldsymbol{z}_k^A$ instead of the whole feature images (see Section 4.2 for more on this). Since the innovations (changes in expressions) are sparse (and mostly with different support), the gross innovation component $\boldsymbol{z}_k^A$ captures most of the unique features of all images in one single image of the same size. It is worth mentioning that there may be some loss of innovation information in the representation of (12), especially if $\boldsymbol{\theta}^c$ is very sparse with a small support while the $\boldsymbol{\theta}_,^i$'s are relatively dense with significant overlap in their support. However, for aligned face images of the same subject, we can expect $\boldsymbol{\theta}_k^c$ to be dense with a significant support compared to the innovations. We will show with examples that the representation of (12) indeed has sufficient information about the innovations (or expressions) of all training image for the purpose of face recognition.

Refer to Fig. 2, where we have three images of a subject with different expressions. (For visual clarity, we have added back the mean of individual training images and also the overall mean to the common component.) It can be seen that the common component retains all the gross features like the face structure, nose region etc. The innovation components retain unique features in respective images (for example, the raised eye-brows and open mouth of the second image in (b) are clearly captured in $\boldsymbol{z}_2^i$ of (f) and so on). It is to be noted that $\boldsymbol{\theta}_,^{i'}\boldsymbol{s}$ are sparse and corresponding spatial domain version $\boldsymbol{z}_,^{i'}\boldsymbol{s}$ in the figure are not sparse, but have some negative pixel values, due to which they appear visually dark. The gross innovation image $\boldsymbol{z}^A$ captures most of the innovation features of all three images in (a), (b) and (c). We will show later that, given only these two features, sufficient innovation information (or expressions) of any image can be recovered and a good estimation can be done using (5).

3.2. S-JSM: A Special Case of B-JSM

A special case of the B-JSM feature extraction method described above is when the common and innovations are directly extracted from spatial image supports (we may call it S-JSM, with S standing for spatial). However, such an approach is sensitive to image alignment, while B-JSM is more robust if a basis like DCT or Wavelet is used. Nevertheless, we present here this alternative so as to provide better insights about the common and innovation features. For S-JSM, we assume that the basis matrix $\boldsymbol{\Psi}$ in Equations (1)-(10) is an identity matrix of size *N*. With these changes, $\boldsymbol{\theta}_k^c = \boldsymbol{z}_k^c$ and $\boldsymbol{\theta}_k^i = \boldsymbol{z}_k^i$ in (6) and the algorithm is expressed as

$$\text{S-JSM} \coloneqq \left(\{\boldsymbol{x}_{k,j}\}, j = 1, \ldots J_k\right) \rightarrow \left[\boldsymbol{z}_k^c\ \ \ldots .\boldsymbol{z}_{k,J_k}^i\right]^T \quad (13)$$

Fig. 3 shows an example of S-JSM features where white patches were intentionally added to the same image to simulate "innovations". (Again, for visual clarity we add back the mean intensity.) We note that the common component retains almost all the information of the face (even the skin intensity at locations of the patches are closely recovered). The innovation component of an image retains the gray patch (which is the difference of the actual patch and the skin intensity at those pixel locations). Hence these effectively carry the information of the original white patches, given the common component. Fig. 3(f) shows that the gross innovation retains all the three gray patches which are unique features of all images. This intuitively illustrates our argument earlier about why the gross innovation is sufficient as long as the individual innovations are sparse (with the hope that the overlap of the innovations should have been captured by the common component).

4. Face Classification

4.1. Expression Recovery and B-JSM Classifier

With the given training features (the common and gross innovation images), there can be many different ways to design a classifier. Let $\boldsymbol{c} \in \mathbb{R}^N$ be a test image of unknown class. One simple way is to assume that $\boldsymbol{c}$ is highly correlated with the correct training class (say class *k*), and hence it would have the same common component $\boldsymbol{z}_k^c$ if we consider the ensemble $\{\boldsymbol{x}_{k,j}, c\}, j = 1,2, ..J_k+1$. So the test image $\boldsymbol{c}$ can be expressed as

$$\boldsymbol{c} = \boldsymbol{z}_k^c + \boldsymbol{c}_k^i \quad (14)$$

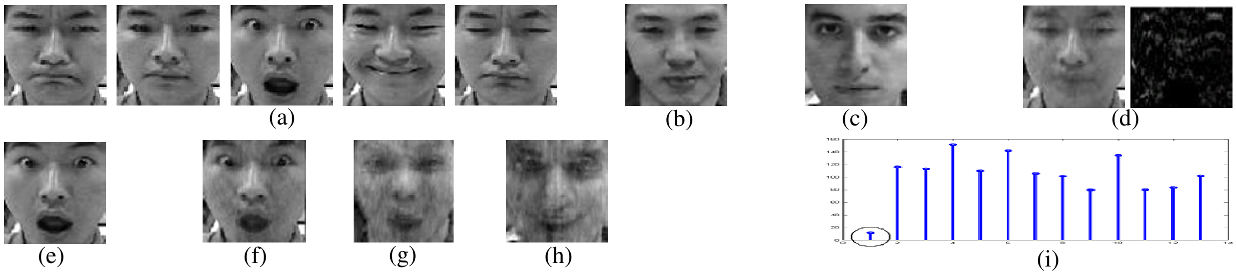


Figure 4: Feature extraction, expression recovery and classification illustrated. (a) Five training images of class 1 (from CMU AMP EXpression database [7] with 13 subjects); (b) & (c) One sample training image of class 2 & 3 respectively; (d) The feature images computed from (a); (e) A sample test image; (f) The estimate of (e) with recovered expressions using class 1 training features of (d); (g) & (h) Estimates with class 2 and class 3 training features respectively; (i) Residuals computed from Eqn. (17), which are used to determine the correct class. (Note: the mean is added back to all images except the gross innovation for better visual clarity).

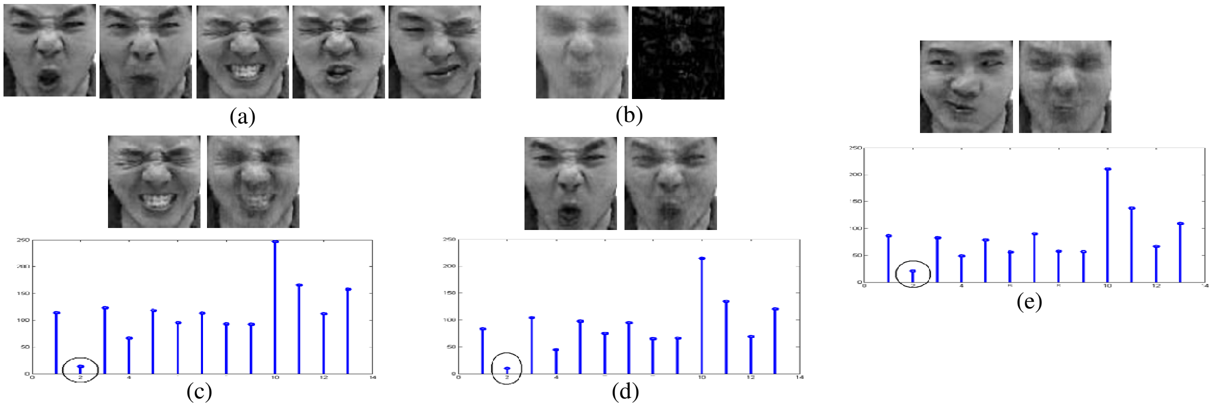


Figure 5: Illustration of image recovery and classification under drastic expressions (13 subjects, five training samples each). (a) Training images of class 2; (b) Training features of (a); (c), (d) and (e) The image on the left is an actual test image of class 2, on the right is the reconstructed image using class 2 features (in(b)), and at the bottom is the residual of Eqn. (17) for all thirteen classes.

where $\boldsymbol{c}_k^i$ is the innovation of $\boldsymbol{c}$. In reality, we need to determine the correct class label k, which may be found as the k for which the energy (or l^2 norm) for $\boldsymbol{c}_k^i$ is minimum. Another approach would be to simply consider sparsity or number of non-zero components of the expansion of $\boldsymbol{c}_k^i$ in basis $\boldsymbol{\Psi}$. However, these methods ignore the information from the gross innovation component $\boldsymbol{z}_k^A$. A better approach would be to first ask the question – “If any at all, what unique feature present in the test innovation $\boldsymbol{c}_k^i$ is also present in $\boldsymbol{z}_k^A$ "? In other words, we want to find the estimate of the innovation component $\boldsymbol{c}_k^i$ of (14) (or expressions) in the test image $\boldsymbol{c}$ using the training features. Assuming B-JSM feature extraction, a good way to estimate $\boldsymbol{c}_k^i$ is to extract a common component $\boldsymbol{F}_k = \boldsymbol{\Psi}\boldsymbol{f}_k$, from the support set common between $\boldsymbol{C}_k^i$ ($\boldsymbol{C}_k^i = \boldsymbol{\Psi}\boldsymbol{c}_k^i$) and $\boldsymbol{Z}_k^A$ ($\boldsymbol{Z}_k^A = \boldsymbol{\Psi}\boldsymbol{z}_k^A$). This can be achieved using the B-JSM recovery model in (11) as follows,

$$\text{B-JSM}(\{\boldsymbol{c}_k^i, \boldsymbol{z}_k^A\}) \rightarrow [\boldsymbol{f}_k, \boldsymbol{f}_k^i, \boldsymbol{z}_k^{iA}]^T \qquad (15)$$

where $\boldsymbol{f}_k^i$ and $\boldsymbol{z}_k^{iA}$ are innovations of $\boldsymbol{c}_k^i$ and $\boldsymbol{z}_k^A$. We may form the estimate of the test image for class k features as,

$$\hat{\boldsymbol{c}}_k = \boldsymbol{z}_k^c + \boldsymbol{f}_k \qquad (16)$$

The correct class label can then be determined as,

$$\boldsymbol{l} = \text{argmin}_{\boldsymbol{k}}\langle\|\hat{\boldsymbol{c}}_k - \boldsymbol{c}\|_2\rangle \qquad (17)$$

Fig. 4 illustrates the results of expression recovery and the classification algorithm explained above for images from the CMU AMP EXpression database [7] (thirteen subjects with five training images chosen per subject). Fig. 4(a) shows all five training images of one subject labeled as class 1 with different expressions. Note that in this case the training common shown in (d) is visually closer to the training images compared to the case in Fig. 2. It is difficult to visually interpret the gross innovation image (Fig. 4(d), right) since it contains a lot information. Nevertheless with the algorithm described above, the innovation information or expressions of a new test image of the correct class (e) can be well recovered, as in (f). On the other hand, for images of the wrong classes (e.g., (b) and (c)), the reconstruction is poor (as in (g) and (h)).

A more challenging case is shown in Fig. 5, illustrating the algorithm performance under drastic variation in expression. Despite the challenge, the expression is fairly well recovered and the classifier residual is very small for the correct class compared to the other classes, leading to correct classification. Note that, in (e), the test image has a totally different expression that is not present in any of the training images. However, the classifier still yields the correct result. This can be attributed to the dominance of the “common component” over the innovations in terms of information for discriminative classification. However, if full or part of expression information is recovered, the discrimination would be more pronounced (compare the residuals of all three test images in Fig. 5). Hence, the B-JSM classifier is robust even in cases where the expression information is missing in the training set. One such practical case is when only a few training images (per subject) are available.

4.2. Low-Dimensional Feature Subspace

We have presented our CS-based algorithm for feature extraction and classification, but have not explicitly considered the underdetermined or ill-poised case involving reduced measurement as in conventional CS coding problems [5, 6, 10, 14-16]. With sparsity prior, (under mild conditions as suggested in CS theory [6, 15, 16]), significant dimensionality reduction in the feature space can be handled by the B-JSM algorithm. This can be explained considering (5), (14) and (15). As discussed in Section 3, the J_k innovations $\boldsymbol{\theta}_k^i$of (5) (for class k) are very sparse with respect to the whole image. Suppose that the test image $\boldsymbol{c}$ belongs to class k, then we may assume that it is sufficiently correlated with the training images (i.e., the training common $\boldsymbol{z}_k^c$ is significant in$\boldsymbol{c}$), which means that $\boldsymbol{c}_k^i$ in (14) is also very sparse with its sparsity of the order comparable to any training innovations $\boldsymbol{z}_{k,j}^i$. Essentially, in the B-JSM expression recovery of (15), we estimate a highly sparse signal and hence the estimate of $\boldsymbol{c}$ via (16) can be done in a lower-dimensional feature space than the original ($\boldsymbol{z}_k^c$ and $\boldsymbol{z}_k^A$). Furthermore, since our emphasis is classification alone and not the fidelity of reconstruction, there is more scope for descending down to extreme low-dimensions.

Let the dimensionality reduction system be $\boldsymbol{\Phi} \in \mathbb{R}^{M\times N}$ ($\boldsymbol{\Phi}$ can be random or any matrix highly incoherent with $\boldsymbol{\Psi}$), a low-dimensional projection of the test image is,

$$\tilde{\boldsymbol{c}} = \boldsymbol{\Phi}\boldsymbol{c} \in \mathbb{R}^M \qquad (18)$$

And the low dimensional versions of the training features are $\tilde{\boldsymbol{z}}_k^c$ and $\tilde{\boldsymbol{z}}_k^{iA}$ given by,

$$\tilde{\boldsymbol{z}}_k^c = \boldsymbol{\Phi}\boldsymbol{z}_k^c,\ \tilde{\boldsymbol{z}}_k^{iA} = \boldsymbol{\Phi}\,\boldsymbol{z}_k^{iA}, \in \mathbb{R}^M \qquad (19)$$

These can be stored right after the training process of Section 3.1. Then the B-JSM algorithm of (15) is computed using the modified version of (9) as below,

$$\boldsymbol{W}_k = \min\left\{ \tau\|\boldsymbol{W}_k\|_1 + 0.5 * \left\|\tilde{\boldsymbol{y}}_k - \boldsymbol{\Phi}\tilde{\boldsymbol{\Psi}}\,\boldsymbol{W}_k\right\|_2^2 \right\} \qquad (20)$$

where $\tilde{\boldsymbol{y}}_k = \left[\,(\tilde{\boldsymbol{c}} - \tilde{\boldsymbol{z}}_k^c) \quad \tilde{\boldsymbol{z}}_k^{iA}\right]^T \in \mathbb{R}^{2M\times 1}$ and $\boldsymbol{W}_k = [\boldsymbol{F}_k\,, \boldsymbol{F}_k^i\,, \boldsymbol{Z}_k^{iA}]^T$ (the transform coefficients of the right hand side of (15). The estimate of the test image can then be determined by (16) as before.

Fig. 6 illustrates the performance of the above algorithm in critically low dimensional feature space for the same setting as in Fig. 5. The original test image is of size 32x32, which is then down-sampled to 16x16. It is obvious that downsampling does change the residuals much. A $\boldsymbol{\Phi}$ operator is applied such that only 10% of linear measurements are retained (102 features for the 32x32 case and merely 25 features for the 16x16 one). Again, the residuals do not alter much. Thus in all cases, correct classification is achieved (more results from entire databases are to be presented and discussed in Section 5).

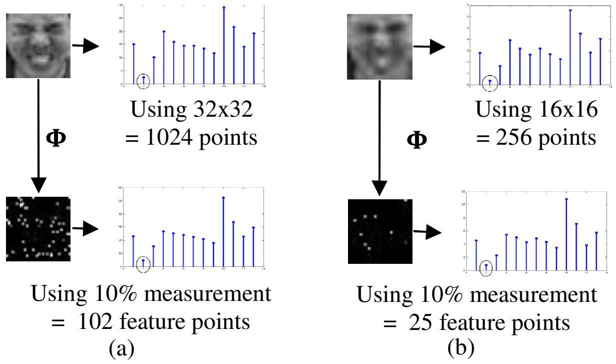


Figure 6: Recognition with critically low-dimensional features. In (a) and (b), the top-left is the input image, and the bottom is its 10% measurement, and on the right are the residuals.

5. Experimental Results and Evaluation

5.1. Experimental Setup

The proposed algorithms described in Sections 3 and 4 were implemented for working with 2-D images instead of vectored images for speed consideration. Further, in all cases unless specified otherwise, the GPSR algorithm [11] was used to solve the unconstrained version if l^1 minimization in (9). (We obtained similar results with other algorithms like TV minimization [12, 16]). For GPSR, we set τ =0.009 and used the continuation approach [11] with first τ factor $\tau f = 0.8(\max(|\tilde{\boldsymbol{\Psi}}^T\boldsymbol{y}_k|/\tau))$. We assume DCT as the sparsifying basis $\boldsymbol{\Psi}$ in our algorithm. Although the sparsifying operation ($\boldsymbol{X} = \boldsymbol{\Psi}\boldsymbol{x}$) is not exactly equivalent to 1-D DCT on vectored image or 2-D DCT on 2-D image (but is actually 1-D DCT on columns of a 2-D image $\boldsymbol{x}$), it yields satisfactory results.

We used three face expression databases: (1) CMU AMP Face EXpression Database [7] (henceforth referred to as CMU), (2) Japanese Female Expression database [8] (henceforth JAFFE), and (3) Cohn-Kanade face expression database [9] (henceforth CK). The CMU database contains 975 images (13 subjects with 75 images per subject) with different facial expressions. The JAFFE database has 213 images with seven different expressions (10 female subjects). The CK database is the most challenging of the three, with 97 subjects and a total of

8795 images. Images of each subject were obtained in 5 to 9 sessions, each session having multiple instances of similar expressions. Since this is a large database, we created three sub-databases of 2317 images by sampling the original frames (uniformly for the first two, randomly for the third) for all subjects. The images were normalized. Our results are compared with the most recent sparse representation based face classification algorithm (SRC) [3], which reported results superior to other methods.

5.2. Experiments, Results, and Comparison

In all the experiments, the training set was formed by randomly selecting *J* images per subject, leaving the rest for testing. The experiments are summarized below.

(i) ***Validation***: We performed validation for various values of *J* with multiple repetitions for each ***J*****:** *J* = 4 to 10 for CMU**,** with 10 trials each; 2 to 5 for JAFFE, with 40 trials each; and 5 to 10 for CK with 9 trials, 3 from each sub-database. The statistics of the recognition rates (High, Low and Average) are given in Tables 1, 2, and 3 with comparison with the SRC algorithm. For CMU and JAFFE, we used 32x32 image size. The results with these two databases show that at lower number of training images, our algorithm invariably outperforms the SRC algorithm and shows better stability. As the number of training images increase, the performance for both methods are on par for most trials, but the averages still indicate that our method is better. For the CK database, we considered a critically low-dimensional image size of 7x7 (49 features). Invariably all times, our method outperforms the SRC algorithm in mean, low and high accuracies. Further, unlike the SRC algorithm, our method exhibits a clear trend of increase in accuracy with increased *J*.

(ii) ***Recognition in Low-Dimensional Feature Space:*** To demonstrate the performance of our algorithm in critically low-dimensional feature space, we apply linear random measurement on 32x32 database images (1024 features), retaining only 40% to 10% values (feature space of 409 to102 points) and evaluate the recognition results. We then downsample the original 32x32 images to 16x16 (256 features) and repeat the process for measurements from 60% to 10%. The effective feature dimensions vary from 153 to as low as just 25 points. Operating in such a low dimensional space is certainly challenging for any database, especially for a large database like CK. Table 4 tabulates the results; where the recognition rate is the average for 3 trials, with *J*=5, 4, and 11 for CMU, JAFFE and CK databases respectively. For this simulation, we used the TV minimization [12]. Clearly, even with 25 feature points, the recognition rate is as high as 94.35%, 97.69% and 97.818% for the three databases respectively.

(iii) ***Robustness of recognition w.r.t. expressions:*** We further designed two types of tests, one where similar expressions are present in both the training and the test sets, and the other where there is no common expression for the training and the test images. We experimented with three expressions (surprise, happiness and neutral) for each database and the results (averaging over 3 trials) are shown in Fig. 7. In all the cases, the performance is still very good: the worst case is only a loss of around 0.23%, 0.4% and 0.79% for CMU, JAFFE and CK databases respectively for the "surprise" expression. For the neutral expression, there is virtually no loss in accuracy (except for JAFFE where the loss is merely 0.05%).

Table 1: Recognition rate (%) for 10 trials on the CMU database with 32x32 image size.

J_k	**Proposed algorithm**			**SRC**		
	High	**Low**	**Avg**	**High**	**Low**	**Avg**
4	100	97.48	98.95	100	97.68	98.9
5	100	99.67	99.91	100	99.12	99.8
6	100	99.69	99.97	100	98.76	99.75
7	100	100	100	100	98.30	99.74
8	100	100	100	100	99.31	99.87
9	100	100	100	100	100	100
10	100	100	100	100	98.73	99.49

Table 2: Recognition rate (%) for 40 trials on the JAFFE database with 32x32 image size.

J_k	**Proposed algorithm**			**SRC**		
	High	**Low**	**Avg**	**High**	**Low**	**Avg**
2	95.89	81.18	89.94	95.11	82.1	90.1
3	98.13	88.13	93.22	98.13	87.0	92.1
4	98.67	90.67	95.12	98.24	90.2	95.13
5	100	93.57	96.12	100	89	96.01

Table 3: Recognition rate (%) for 5 trials on the CK database with mere 7x7 image size.

J_k	**Proposed algorithm**			**SRC**		
	High	**Low**	**Avg**	**High**	**Low**	**Avg**
5	96.2	94.01	95.47	89.3	93.4	91.41
6	97.43	94.63	95.93	94.04	91.3	93.77
7	97.35	95.21	96.15	91.89	94.9	93.29
8	97.9	95.23	96.49	94.43	81.0	89.78
9	98.01	95.28	96.90	97.73	95.4	96.29
10	98.63	95.69	97.14	98.1	94.1	95.64

Table 4: Recognition rate (%) for databases with low-dimensional features. "%M" gives the percentage of measurements taken and "ED" refers to "effective dimension".

Image Size	**%M**	**ED**	**CMU**	**JAFFE**	**CK**
32x32 =1024 pixels	10	102	99.23	97.69	98.425
	20	204	99.45	98.46	98.69
	30	307	99.67	98.69	98.91
	40	409	99.78	98.69	99.01
16x16 = 256 pixels	10	25	94.35	97.69	97.818
	20	51	99.45	98.22	98.303
	30	76	99.67	98.46	98.546
	40	102	99.67	98.46	98.546
	50	128	99.78	98.69	98.939
	60	153	99.78	99.69	98.939

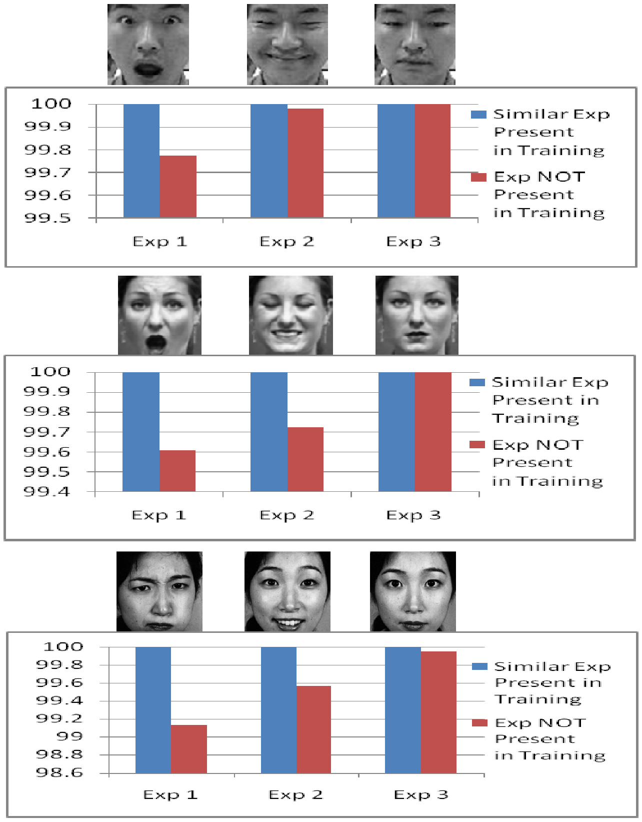


Figure 7: The recognition rate with and without the presence of similar expressions in the training set - (Surprise (Exp1), Happiness (Exp2) and Neutral (Exp3)). For the CMU (top), CK (middle), and JAFEE (bottom) databases.

6. Conclusion and Future Work

We proposed a novel technique based on compressive sensing for expression-invariant face recognition. The approach exploits the correlation of images from the same subject through joint sparsity models in designing novel algorithms for feature extraction and face recognition. Thorough analysis of the proposed algorithms and their performance evaluation, with comparison to the state-of-the-art, were performed to demonstrate the claimed advantages. We are currently working towards the following extension of the proposed method: handling illumination changes and pose variations. In addition, the approach is general in nature and thus can be applied to other problems involving multiple views of the scene, which is another direction we are pursuing.

References

[1] P. Belhumeur, J. Hespanha, D. Kriegman, “Eigenfaces vs. Fisherfaces: recognition using class specific linear projection”, in: European Conference on Computer Vision, 1996, pp. 45--58.

[2] W. Zhao , R. Chellppa , P. J. Phillips , A. Rosenfeld “Face Recognition: A Literature Survey” ACM Computing Surveys, Vol. 35, No. 4, December 2003, pp. 399–458.

[3] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. “Robust face recognition via sparse representation”. *IEEE Trans. PAMI* [DOI 10.1109/TPAMI.2008.79].

[4] J. Huang, X. Huang, D. Metaxas, “Simultaneous Image Transformation and Sparse Representation Recovery” IEEE Conf. on CVPR, Anchorage,AK, June 2008.

[5] D. Baron, M. Duarte, S. Sarvotham, M. B. Wakin, and R. G. Baraniuk, “Distributed compressed sensing,” Tech. Rep. TREE0612, Rice University, Online at: http://dsp.rice.edu/cs/.

[6] E J. Candès and M B. Wakin “An Introduction to Compressive Sampling”, *IEEE Signal Proc. Magazine*, Vol. 25, Issue 2, March 2008, pp. 21-31.

[7] X. Liu, T.Chen and B.V.K. Vijaya Kumar, “Face Authentication for Multiple Subjects Using Eigenflow” Pattern Recognition, Volume 36, Issue 2, February 2003, pp. 313-328.

[8] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba "Coding Facial Expressions with Gabor Wavelets" *IEEE* Int Conf. on Auto. Face and Gesture Recognition, Nara, Japan, April,1998.

[9] Kanade, T., Cohn, J.F., & Tian, Y. “Comprehensive database for facial expression analysis” *IEEE Int. Conf. on Automatic Face and Gesture Recognition,*Grenoble, France, 2000.

[10] M. F. Duarte, S. Sarvotham, D. Baron, M. B. Wakin and R. G. Baraniuk, “Distributed Compressed Sensing of Jointly Sparse Signals”, *39th Asilomar Conference on Signals, Systems and Computer (IEEE Cat. No.05CH37761)*, 2005, pp. 1537-41.

[11] M. Figueiredo, R. Nowak, and S. Wright. “Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems”. *IEEE Journal on Selected Topics in Signal Processing*, 2007, Vol 1, Issue 4, pp. 586-597.

[12] E. Candès and J. Romberg, “Practical signal recovery from random projections”. Wavelet Applications in Signal and Image Processing XI, Proc. SPIE Conf. 5914.

[13] E. Candès, J. Romberg, and T. Tao. “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information”. *IEEE Trans. Inf. Theory*, 52:489–509, 2006.

[14] E. Candès and T. Tao. “Near-optimal signal recovery from randomprojections: Universal encoding strategies?” *IEEE Trans. on Information Theory*, 52(12):5406–5425, 2006.

[15] D. L. Donoho, “Compressed sensing,” *IEEE Trans. Inform. Theory*, vol. 52,July 2006, pp. 1289–1306.

[16] E. Cand`es, J. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” Comm. on Pure and Applied Math, vol. 59, no. 8, 2006, pp. 1207–1223.

[17] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinvesky. “A method for large-scale ℓ1-regularized least squares problems with applications in signal processing and statistics,” *IEEE* J. Selected Topics in Signal Processing, 1(4):606-617, Dec 2007.

[18] E. Cand`es and J. Romberg, “l^1-magic: Recovery of Sparse Signals via Convex Programming” User Guide, 11-magic software, Available at http://www.acm.caltech.edu/l1magic/.