Robust Sparse Coding for Face Recognition

Meng Yang Lei Zhang * Jian Yang David Zhang
Hong Kong Polytechnic Univ. Nanjing Univ. of Sci. & Tech. Hong Kong Polytechnic Univ.

Abstract

Recently the sparse representation (or coding) based classification (SRC) has been successfully used in face recognition. In SRC, the testing image is represented as a sparse linear combination of the training samples, and the representation fidelity is measured by the l_2-norm or l_1-norm of coding residual. Such a sparse coding model actually assumes that the coding residual follows Gaussian or Laplacian distribution, which may not be accurate enough to describe the coding errors in practice. In this paper, we propose a new scheme, namely the robust sparse coding (RSC), by modeling the sparse coding as a sparsity-constrained robust regression problem. The RSC seeks for the MLE (maximum likelihood estimation) solution of the sparse coding problem, and it is much more robust to outliers (e.g., occlusions, corruptions, etc.) than SRC. An efficient iteratively reweighted sparse coding algorithm is proposed to solve the RSC model. Extensive experiments on representative face databases demonstrate that the RSC scheme is much more effective than state-of-the-art methods in dealing with face occlusion, corruption, lighting and expression changes, etc.

1. Introduction

As a powerful tool for statistical signal modeling, sparse representation (or sparse coding) has been successfully used in image processing applications [16], and recently has led to promising results in face recognition [24, 25, 27] and texture classification [15]. Based on the findings that natural images can be generally coded by structural primitives (e.g., edges and line segments) that are qualitatively similar in form to simple cell receptive fields [18], sparse coding techniques represent a natural image using a small number of atoms parsimoniously chosen out of an over-complete dictionary. Intuitively, the sparsity of the coding coefficient vector can be measured by the l_0-norm of it (l_0-norm counts the number of nonzero entries in a vector). Since the combinatorial l_0-norm minimization is an NP-hard problem, the l_1-norm minimization, as the closest convex function to l_0-norm minimization, is widely employed in sparse coding, and it was shown that l_0-norm and l_1-norm minimizations are equivalent if the solution is sufficiently sparse [3]. In general, the sparse coding problem can be formulated as

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \quad \text{s.t.} \quad \|\boldsymbol{y} - D\boldsymbol{\alpha}\|_2^2 \leq \varepsilon, \tag{1}$$

where $\boldsymbol{y}$ is a given signal, D is the dictionary of coding atoms, $\boldsymbol{\alpha}$ is the coding vector of $\boldsymbol{y}$ over D, and $\varepsilon > 0$ is a constant.

Face recognition (FR) is among the most visible and challenging research topics in computer vision and pattern recognition [29], and many methods, such as Eigenfaces [21], Fisherfaces [2] and SVM [7], have been proposed in the past two decades. Recently, Wright *et al.* [25] applied sparse coding to FR and proposed the sparse representation based classification (SRC) scheme, which achieves impressive FR performance. By coding a query image $\boldsymbol{y}$ as a sparse linear combination of the training samples via the l_1-norm minimization in Eq. (1), SRC classifies the query image $\boldsymbol{y}$ by evaluating which class of training samples could result in the minimal reconstruction error of it with the associated coding coefficients. In addition, by introducing an identity matrix I as a dictionary to code the outlier pixels (e.g., corrupted or occluded pixels):

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \|[\boldsymbol{\alpha};\boldsymbol{\beta}]\|_1 \quad \text{s.t.} \quad \boldsymbol{y} = [D, I] \cdot [\boldsymbol{\alpha};\boldsymbol{\beta}], \tag{2}$$

the SRC method shows high robustness to face occlusion and corruption. In [9], Huang *et al.* proposed a sparse representation recovery method which is invariant to image-plane transformation to deal with the misalignment and pose variation in FR, while in [22] Wagner *et al.* proposed a sparse representation based method that could deal with face misalignment and illumination variation. Instead of directly using original facial features, Yang and Zhang [27] used Gabor features in SRC to reduce greatly the size of occlusion dictionary and improve a lot the FR accuracy.

The sparse coding model in Eq. (1) is widely used in literature. There are mainly two issues in this model. The first one is that whether the l_1-norm constraint $\|\boldsymbol{\alpha}\|_1$ is good enough to characterize the signal sparsity. The second one is

*Corresponding author. This research is supported by the Hong Kong General Research Fund (PolyU 5351/08E).

that whether the l_2-norm term $\|\boldsymbol{y} - D\boldsymbol{\alpha}\|_2^2 \leq \varepsilon$ is effective enough to characterize the signal fidelity, especially when the observation $\boldsymbol{y}$ is noisy or has many outliers. Many works have been done for the first issue by modifying the sparsity constraint. For example, Liu *et al.* [14] added a nonnegative constraint to the sparse coefficient $\boldsymbol{\alpha}$; Gao *et al.* [4] introduced a Laplacian term of coefficient in sparse coding; Wang *et al.* [23] used the weighted l_2-norm for the sparsity constraint. In addition, Ramirez *et al.* [19] proposed a framework of universal sparse modeling to design sparsity regularization terms. The Bayesian methods were also used for designing the sparsity regularization terms [11].

The above developments of sparsity regularization term in Eq. (1) improve the sparse representation in different aspects; however, to the best of our knowledge, little work has been done on improving the fidelity term $\|\boldsymbol{y} - D\boldsymbol{\alpha}\|_2^2$ except that in [24, 25] the l_1-norm was used to define the coding fidelity (i.e., $\|\boldsymbol{y} - D\boldsymbol{\alpha}\|_1$). In fact, the fidelity term has a high impact on the final coding results because it ensures that the given signal $\boldsymbol{y}$ can be faithfully represented by the dictionary D. From the viewpoint of maximum likelihood estimation (MLE), defining the fidelity term with l_2- or l_1-norm actually assumes that the coding residual $\boldsymbol{e} = \boldsymbol{y} - D\boldsymbol{\alpha}$ follows Gaussian or Laplacian distribution. But in practice this assumption may not hold well, especially when occlusions, corruptions and expression variations occur in the query face images. So the conventional l_2- or l_1-norm based fidelity term in sparse coding model Eq. (1) may not be robust enough in these cases. Meanwhile, these problems cannot be well solved by modifying the sparsity regularization term.

To improve the robustness and effectiveness of sparse representation, we propose a so-called robust sparse coding (RSC) model in this paper. Inspired by the robust regression theory [1, 10], we design the signal fidelity term as an MLE-like estimator, which minimizes some function (associated with the distribution of the coding residuals) of the coding residuals. The proposed RSC scheme utilizes the MLE principle to robustly regress the given signal with sparse regression coefficients, and we transform the minimization problem into an iteratively reweighted sparse coding problem. A reasonable weight function is designed for applying RSC to FR. Our extensive experiments in benchmark face databases show that RSC achieves much better performance than existing sparse coding based FR methods, especially when there are complicated variations of face images, such as occlusions, corruptions and expressions, etc.

The rest of this paper is organized as follows. Section 2 presents the proposed RSC model. Section 3 presents the algorithm of RSC and some analyses, such as convergence and complexity. Section 4 conducts the experiments, and Section 5 concludes the paper.

2. Robust Sparse Coding (RSC)

2.1. The RSC model

The traditional sparse coding model in Eq. (1) is equivalent to the so-called LASSO problem [20]:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - D\boldsymbol{\alpha}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_1 \leq \sigma, \tag{3}$$

where $\sigma > 0$ is a constant, $\boldsymbol{y} = [y_1; y_2; \cdots; y_n] \in \mathbb{R}^n$ is the signal to be coded, $D = [\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_m] \in \mathbb{R}^{n \times m}$ is the dictionary with column vector $\boldsymbol{d}_j$ being the j^{th} atom, and $\boldsymbol{\alpha}$ is the coding coefficient vector. In our problem of FR, the atom $\boldsymbol{d}_j$ is the training face sample (or its dimensionality reduced feature) and hence the dictionary D is the training dataset.

We can see that the sparse coding problem in Eq. (3) is essentially a sparsity-constrained least square estimation problem. It is known that only when the residual $\boldsymbol{e} = \boldsymbol{y} - D\boldsymbol{\alpha}$ follows the Gaussian distribution, the least square solution is the MLE solution. If $\boldsymbol{e}$ follows the Laplacian distribution, the MLE solution will be

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - D\boldsymbol{\alpha}\|_1 \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_1 \leq \sigma, \tag{4}$$

Actually Eq. (4) is essentially another expression of Eq. (2) because both of them can have the following Lagrangian formulation: $\min_{\boldsymbol{\alpha}} \{\|\boldsymbol{y} - D\boldsymbol{\alpha}\|_1 + \lambda \|\boldsymbol{\alpha}\|_1\}$[26].

In practice, however, the distribution of residual $\boldsymbol{e}$ may be far from Gaussian or Laplacian distribution, especially when there are occlusions, corruptions and/or other variations. Hence, the conventional sparse coding models in Eq. (3) (or Eq. (1)) and Eq. (4) (or Eq. (2)) may not be robust and effective enough for face image representation.

In order to construct a more robust model for sparse coding of face images, in this paper we propose to find an MLE solution of the coding coefficients. We rewrite the dictionary D as $D = [\boldsymbol{r}_1; \boldsymbol{r}_2; \cdots; \boldsymbol{r}_n]$, where row vector $\boldsymbol{r}_i$ is the i^{th} row of D. Denote by $\boldsymbol{e} = \boldsymbol{y} - D\boldsymbol{\alpha} = [e_1; e_2; \cdots; e_n]$ the coding residual. Then each element of $\boldsymbol{e}$ is $e_i = y_i - r_i\boldsymbol{\alpha}, i = 1, 2, \cdots, n$. Assume that $e_1, e_2, \cdots, e_n$ are independently and identically distributed according to some probability density function (PDF) $f_{\boldsymbol{\theta}}(e_i)$, where $\boldsymbol{\theta}$ denotes the parameter set that characterizes the distribution. Without considering the sparsity constraint of $\boldsymbol{\alpha}$, the likelihood of the estimator is $L_{\boldsymbol{\theta}}(e_1, e_2, \cdots, e_n) = \prod_{i=1}^n f_{\boldsymbol{\theta}}(e_i)$, and MLE aims to maximize this likelihood function or, equivalently, minimize the objective function: $-\ln L_{\boldsymbol{\theta}} = \sum_{i=1}^n \rho_{\boldsymbol{\theta}}(e_i)$, where $\rho_{\boldsymbol{\theta}}(e_i) = -\ln f_{\boldsymbol{\theta}}(e_i)$.

With consideration of the sparsity constraint of $\boldsymbol{\alpha}$, the MLE of $\boldsymbol{\alpha}$, namely the robust sparse coding (RSC), can be formulated as the following minimization

$$\min_{\boldsymbol{\alpha}} \sum\nolimits_{i=1}^n \rho_{\boldsymbol{\theta}}(y_i - \boldsymbol{r}_i\boldsymbol{\alpha}) \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_1 \leq \sigma, \tag{5}$$

In general, we assume that the unknown PDF $f_{\boldsymbol{\theta}}(e_i)$ is symmetric, and $f_{\boldsymbol{\theta}}(e_i) < f_{\boldsymbol{\theta}}(e_j)$ if $|e_i| > |e_j|$. So $\rho_{\boldsymbol{\theta}}(e_i)$ has the following properties: $\rho_{\boldsymbol{\theta}}(0)$ is the global minimal of $\rho_{\boldsymbol{\theta}}(e_i)$; $\rho_{\boldsymbol{\theta}}(e_i) = \rho_{\boldsymbol{\theta}}(-e_i)$; $\rho_{\boldsymbol{\theta}}(e_i) < \rho_{\boldsymbol{\theta}}(e_j)$ if $|e_i| > |e_j|$. Without loss of generality, we let $\rho_{\boldsymbol{\theta}}(0) = 0$.

Form Eq. (5), we can see that the proposed RSC model is essentially a sparsity-constrained MLE problem. In other words, it is a more general sparse coding model, while the conventional sparse coding models in Eq. (3) and Eq. (4) are special cases of it when the coding residual follows Gaussian and Laplacian distributions, respectively.

By solving Eq. (5), we can get the MLE solution to $\boldsymbol{\alpha}$ with sparsity constraint. Clearly, one key problem is how to determine the distribution $\rho_{\boldsymbol{\theta}}$ (or $f_{\boldsymbol{\theta}}$). Explicitly taking $f_{\boldsymbol{\theta}}$ as Gaussian or Laplacian distribution is simple but not effective enough. In this paper, we do not determine $\rho_{\boldsymbol{\theta}}$ directly to solve Eq. (5). Instead, with the above mentioned general assumptions of $\rho_{\boldsymbol{\theta}}$, we transform the minimization problem in Eq. (5) into an iteratively reweighted sparse coding problem, and the resulted weights have clear physical meaning, i.e., outliers will have low weight values. By iteratively computing the weights, the MLE solution of RSC could be solved efficiently.

2.2. The distribution induced weights

Let $F_{\boldsymbol{\theta}}(\boldsymbol{e}) = \sum_{i=1}^{n} \rho_{\boldsymbol{\theta}}(e_i)$. We can approximate $F_{\boldsymbol{\theta}}(\boldsymbol{e})$ by its first order Taylor expansion in the neighborhood of $\boldsymbol{e}_0$: $\tilde{F}_{\boldsymbol{\theta}}(\boldsymbol{e}) = F_{\boldsymbol{\theta}}(\boldsymbol{e}_0) + (\boldsymbol{e} - \boldsymbol{e}_0)^T F'_{\boldsymbol{\theta}}(\boldsymbol{e}_0) + \boldsymbol{R}_1(\boldsymbol{e})$, where $\boldsymbol{R}_1(\boldsymbol{e})$ is the high order residual term, and $F'_{\boldsymbol{\theta}}(\boldsymbol{e})$ is the derivative of $F_{\boldsymbol{\theta}}(\boldsymbol{e})$. Denote by $\rho'_{\boldsymbol{\theta}}$ the derivative of $\rho_{\boldsymbol{\theta}}$, and then $F'_{\boldsymbol{\theta}}(\boldsymbol{e}_0) = [\rho'_{\boldsymbol{\theta}}(e_{0,1}); \rho'_{\boldsymbol{\theta}}(e_{0,2}); \cdots; \rho'_{\boldsymbol{\theta}}(e_{0,n})]$, where $e_{0,i}$ is the i^{th} element of $\boldsymbol{e}_0$.

In sparse coding, it is usually expected that the fidelity term is strictly convex. So we approximate the residual term as $R_1(\boldsymbol{e}) = 0.5(\boldsymbol{e}-\boldsymbol{e}_0)^T W(\boldsymbol{e}-\boldsymbol{e}_0)$, where W is a diagonal matrix for that the elements in $\boldsymbol{e}$ are independent and there is no cross term between e_i and e_j, $i \neq j$, in $F_{\boldsymbol{\theta}}(\boldsymbol{e})$. Since $F_{\boldsymbol{\theta}}(\boldsymbol{e})$ reaches its minimal value (i.e., 0) at $\boldsymbol{e} = \boldsymbol{0}$, we also require that $\tilde{F}_{\boldsymbol{\theta}}(\boldsymbol{e})$ has its minimal value at $\boldsymbol{e} = \boldsymbol{0}$. Letting $\tilde{F}'_{\boldsymbol{\theta}}(\boldsymbol{0}) = \boldsymbol{0}$, we have the diagonal element of W as

$$W_{i,i} = \omega_{\boldsymbol{\theta}}(e_{0,i}) = \rho'_{\boldsymbol{\theta}}(e_{0,i})/e_{0,i}. \tag{6}$$

According to the properties of $\rho_{\boldsymbol{\theta}}(e_i)$, $\rho'_{\boldsymbol{\theta}}(e_i)$ will have the same sign as e_i. So each $W_{i,i}$ is a non-negative scalar. Then $\tilde{F}_{\boldsymbol{\theta}}(\boldsymbol{e})$ can be written as $\tilde{F}_{\boldsymbol{\theta}}(\boldsymbol{e}) = \frac{1}{2}\left\|W^{1/2}\boldsymbol{e}\right\|_2^2 + b$, where b is a scalar value determined by $\boldsymbol{e}_0$. Since $\boldsymbol{e} = \boldsymbol{y} - D\boldsymbol{\alpha}$, the RSC model in Eq. (5) can be approximated by

$$\min_{\boldsymbol{\alpha}} \left\|W^{1/2}\left(\boldsymbol{y} - D\boldsymbol{\alpha}\right)\right\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_1 \leq \sigma, \tag{7}$$

which is clearly a weighted LASSO problem. Because the weight matrix W needs to be estimated using Eq. (6), Eq. (7) is a local approximation of the RSC in Eq. (5) at $\boldsymbol{e}_0$, and the minimization procedure of RSC can be transformed into an iteratively reweighted sparse coding problem with W being updated using the residuals in previous iteration via Eq. (6). Each $W_{i,i}$ is a non-negative scalar, so the weighted LASSO in each iteration is a convex problem, which could be solved easily by methods such as l_1-ls [12].

Since W is a diagonal matrix, its element $W_{i,i}$ (i.e., $\omega_{\boldsymbol{\theta}}(e_i)$) is the weight assigned to each pixel of the query image $\boldsymbol{y}$. Intuitively, in FR the outlier pixels (e.g. occluded or corrupted pixels) should have low weight values. Thus, with Eq. (7) the determination of distribution $\rho_{\boldsymbol{\theta}}$ is transformed into the determination of weight W. Considering the logistic function has properties similar to the hinge loss function in SVM [28], we choose it as the weight function

$$\omega_{\boldsymbol{\theta}}(e_i) = \exp\left(\mu\delta - \mu e_i^2\right) / \left(1 + \exp\left(\mu\delta - \mu e_i^2\right)\right) \tag{8}$$

where μ and δ are positive scalars. Parameter μ controls the decreasing rate from 1 to 0, and δ controls the location of demarcation point. With Eq. (8), Eq. (6) and $\rho_{\boldsymbol{\theta}}(0) = 0$, we could get

$$\rho_{\boldsymbol{\theta}}(e_i) = \frac{-1}{2\mu}\left(\ln\left(1 + \exp\left(\mu\delta - \mu e_i^2\right)\right) - \ln\left(1 + \exp\mu\delta\right)\right) \tag{9}$$

The original sparse coding models in Eqs. (3) and (4) can be interpreted by Eq. (7). The model in Eq. (3) is the case by letting $\omega_{\boldsymbol{\theta}}(e_i) = 2$. The model in Eq. (4) is the case by letting $\omega_{\boldsymbol{\theta}}(e_i) = 1/|e_i|$. Compared with the models in Eqs. (3) and (4), the proposed weighted LASSO in Eq. (7) has the following advantage: outliers (usually the pixels with big residuals) will be adaptively assigned with low weights to reduce their affects on the regression estimation so that the sensitiveness to outliers can be greatly reduced. The weight function of Eq. (8) is bounded in $[0, 1]$. Although the model in Eq. (4) also assigns low weight to outliers, its weight function is not bounded. The weights of pixels with very small residuals will have nearly infinite values. This reduces the stability of the coding process.

The convexity of the RSC model (Eq. (5)) depends on the form of $\rho_{\boldsymbol{\theta}}(e_i)$ or the weight function $\omega_{\boldsymbol{\theta}}(e_i)$. If we simply let $\omega_{\boldsymbol{\theta}}(e_i) = 2$, the RSC degenerates to the original sparse coding problem (Eq. (3)), which is convex but not effective. The RSC model is not convex with the weight function defined in Eq. (8). However, for FR, a good initialization can always be got, and our RSC algorithm described in next section could always find a local optimal solution, which has very good FR performance as validated in the experiments in Section 4.

3. Algorithm of RSC

As discussed in Section 2.2, the implementation of RSC can be an iterative process, and in each iteration it is a convex l_1-minimization problem. In this section we propose

such an iteratively reweighted sparse coding (IRSC) algorithm to solve the RSC minimization.

3.1. Iteratively reweighted sparse coding (IRSC)

Although in general the RSC model can only have a locally optimal solution, fortunately in FR we are able to have a very reasonable initialization to achieve good performance. When a testing face image $\boldsymbol{y}$ comes, in order to initialize the weight, we should firstly estimate the coding residual $\boldsymbol{e}$ of $\boldsymbol{y}$. We can initialize $\boldsymbol{e}$ as $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{y}_{ini}$, where $\boldsymbol{y}_{ini}$ is some initial estimation of the true face from observation $\boldsymbol{y}$. Because we do not know which class the testing face image $\boldsymbol{y}$ belongs to, a reasonable $\boldsymbol{y}_{ini}$ can be set as the mean image of all training images. In the paper, we simply compute $\boldsymbol{y}_{ini}$ as

$$\boldsymbol{y}_{ini} = \boldsymbol{m}_D, \tag{10}$$

where $\boldsymbol{m}_D$ is the mean image of all training samples.

With the initialized $\boldsymbol{y}_{ini}$, our algorithm to solve the RSC model, namely Iteratively Reweighted Sparse Coding (IRSC), is summarized in Algorithm 1.

When RSC converges, we use the same classification strategy as in SRC [25] to classify the face image $\boldsymbol{y}$.

3.2. The convergence of IRSC

The weighted sparse coding in Eq. (7) is a local approximation of RSC in Eq. (5), and in each iteration the objective function value of Eq. (5) decreases by the IRSC algorithm. Since the original cost function of Eq. (5) is lower bounded ($\geq$0), the iterative minimization procedure in IRSC will converge.

The convergence is achieved when the difference of the weight between adjacent iterations is small enough. Specifically, we stop the iteration if the following holds:

$$\left\|W^{(t)} - W^{(t-1)}\right\|_2 \Big/ \left\|W^{(t-1)}\right\|_2 < \gamma, \tag{12}$$

where γ is a small positive scalar.

3.3. Complexity analysis

The complexity of both SRC and the proposed IRSC mainly lies in the sparse coding process, i.e., Eq. (3) and Eq. (7). Suppose that the dimensionality n of face feature is fixed, the complexity of sparse coding model Eq. (3) basically depends on the number of dictionary atoms, i.e. m. The empirical complexity of commonly used l_1-regularized sparse coding methods (such as l_1-ls [12]) to solve Eq. (3) or Eq. (7) is $O(m^\varepsilon)$ with $\varepsilon \approx 1.5$ [12]. For FR without occlusion, SRC [25] performs sparse coding once and then uses the residuals associated with each class to classify the face image, while RSC needs several iterations (usually 2 iterations) to finish the coding. Thus in this case, RSC's complexity is higher than SRC.

Algorithm 1 Iteratively Reweighted Sparse Coding

Input: Normalized test sample $\boldsymbol{y}$ with unit l_2-norm, dictionary D (each column of D has unit l_2-norm) and $\boldsymbol{y}_{rec}^{(1)}$ initialized as $\boldsymbol{y}_{ini}$.

Output: $\boldsymbol{\alpha}$

Start from $t = 1$:

1: Compute residual $\boldsymbol{e}^{(t)} = \boldsymbol{y} - \boldsymbol{y}_{rec}^{(t)}$.

2: Estimate weights as

$$\omega_{\boldsymbol{\theta}}\left(e_i^{(t)}\right) = \frac{\exp\left(\mu^{(t)}\delta^{(t)} - \mu^{(t)}(e_i^{(t)})^2\right)}{1 + \exp\left(\mu^{(t)}\delta^{(t)} - \mu^{(t)}(e_i^{(t)})^2\right)}, \tag{11}$$

where $\mu^{(t)}$ and $\delta^{(t)}$ are parameters estimated in the t^{th} iteration (please refer to Section 4.1 for the setting of them).

3: Sparse coding:

$$\boldsymbol{\alpha}^* = \min_{\boldsymbol{\alpha}} \left\|(W^{(t)})^{1/2}(\boldsymbol{y} - D\boldsymbol{\alpha})\right\|_2^2 \quad \text{s.t.}\ \|\boldsymbol{\alpha}\|_1 \leq \sigma,$$

where $W^{(t)}$ is the estimated diagonal weight matrix with $W_{i,i}^{(t)} = \omega_{\boldsymbol{\theta}}(e_i^{(t)})$.

4: Update the sparse coding coefficients:

If $t = 1$, $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^*$;

If $t > 1$, $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^{(t-1)} + \eta^{(t)}\left(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^{(t-1)}\right)$;

where $0 < \eta^{(t)} < 1$ is the step size, and a suitable $\eta^{(t)}$ should make $\sum_{i=1}^n \rho_{\boldsymbol{\theta}}(e^{(t)}) < \sum_{i=1}^n \rho_{\boldsymbol{\theta}}(e^{(t-1)})$. $\eta^{(t)}$ can be searched from 1 to 0 by the standard line-search process [8]. (Since both $\boldsymbol{\alpha}^{(t-1)}$ and $\boldsymbol{\alpha}^*$ belong to the convex set $Q = \{\|\boldsymbol{\alpha}\|_1 \leq \sigma\}$, $\boldsymbol{\alpha}^{(t)}$ will also belong to Q.

5: Compute the reconstructed test sample:

$\boldsymbol{y}_{rec}^{(t)} = D\boldsymbol{\alpha}^{(t)}$,

and let $t = t + 1$.

6: Go back to step 1 until the condition of convergence (described in Section 3.2) is met, or the maximal number of iterations is reached.

For FR with occlusion or corruption, SRC needs to use an identity matrix to code the occluded or corrupted pixels, as shown in Eq. (2). In this case SRC's complexity is $O((m+n)^\varepsilon)$. Considering the fact that n is often much greater than m in sparse coding based FR (e.g. $n = 8086$, $m = 717$ in the experiments with pixel corruption and block occlusion in [25]), the complexity of SRC becomes very high when dealing with occlusion and corruption.

The computational complexity of our proposed RSC is $O(k(m)^\varepsilon)$, where k is the number of iteration. Note that k depends on the percentage of outliers in the face image. By our experience, when there is a small percentage of outliers, RSC will converge in only two iterations. If there is a big percentage of outliers (e.g. occlusion, corruption, etc.), RSC could converge in 10 iterations. So for FR with occlu-

sion, the complexity of RSC is generally much lower than SRC. In addition, in the iteration of IRSC we can delete the element y_i that has very small weight because this implies that y_i is an outlier. Thus the complexity of RSC can be further reduced (i.e., in FR with real disguise on the AR database, about 30% pixels could be deleted in each iteration in average).

4. Experimental Results

In this section, we perform experiments on benchmark face databases to demonstrate the performance of RSC (source codes accompanying this work are available at http://www.comp.polyu.edu.hk/~cslzhang/code.htm). We first discuss the parameter selection of RSC in Section 4.1; in Section 4.2, we test RSC for FR without occlusion on three face databases (Extended Yale B [5, 13], AR [17], and Multi-PIE [6]). In Section 4.3, we demonstrate the robustness of RSC to random pixel corruption, random block occlusion and real disguise. All the face images are cropped and aligned by using the locations of eyes, which are provided by the face databases (except for Multi-PIE, for which we manually locate the positions of eyes). For all methods, the training samples are used as the dictionary D in sparse coding.

4.1. Parameter selection

In the weight function Eq. (8), there are two parameters, δ and μ, which need to be calculated in Step 2 of IRSC. δ is the parameter of demarcation point. When the square of residual is larger than δ, the weight value is less than 0.5. In order to make the model robust to outliers, we compute the value of δ as follows.

Denote by $\boldsymbol{\psi} = \left[(e_1)^2, (e_2)^2, \cdots, (e_n)^2\right]$. By sorting $\boldsymbol{\psi}$ in an ascending order, we get the re-ordered array $\boldsymbol{\psi}_a$. Let $k = \lfloor \tau n \rfloor$, where scalar $\tau \in (0, 1]$, and $\lfloor \tau n \rfloor$ outputs the largest integer smaller than τn. We set δ as

$$\delta = \boldsymbol{\psi}_a(k) \tag{13}$$

Parameter μ controls the decreasing rate of weight value from 1 to 0. Here we simply let $\mu = c/\delta$, where c is a constant. In the experiments, if no specific instructions, c is set as 8; τ is set as 0.8 for FR without occlusion, and 0.5 for FR with occlusion. In addition, in our experiments, we solve the (weighted) sparse coding (in Eq. (2), Eq. (3) or Eq.(7)) by its unconstrained Lagrangian formulation. Take Eq. (3) as an example, its Lagrangian form is $\min_{\boldsymbol{\alpha}} \left\{ \|\boldsymbol{y} - D\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 \right\}$), and the default value for the multiplier, λ, is 0.001.

4.2. Face recognition without occlusion

We first validate the performance of RSC in FR with variations such as illumination and expression changes but without occlusion. We compare RSC with the popular methods such as nearest neighbor (NN), nearest subspace (NS), linear support vector machine (SVM), and the recently developed SRC [25].

Dim	30	84	150	300
NN	66.3%	85.8%	90.0%	91.6%
NS	63.6%	94.5%	95.1%	96.0%
SVM	**92.4%**	94.9%	96.4%	97.0%
SRC [25]	90.9%	95.5%	96.8%	98.3%
RSC	91.3%	**98.1%**	**98.4%**	**99.4%**

Table 1. Face recognition rates on the Extended Yale B database

In the experiments, PCA (i.e., Eigenfaces [21]) is used to reduce the dimensionality of original face features, and the Eigenface features are used for all the competing methods. Denote by P the subspace projection matrix computed by applying PCA to the training data. Then in RSC, the sparse coding in step 3 of IRSC becomes: $\boldsymbol{\alpha}^* = \min_{\boldsymbol{\alpha}} \left\| P(W^{(t)})^{1/2} (\boldsymbol{y} - D\boldsymbol{\alpha}) \right\|_2^2 \quad \text{s.t.} \ \|\boldsymbol{\alpha}\|_1 \leq \sigma.$

1) Extended Yale B Database: The Extended Yale B [5, 13] database contains about 2,414 frontal face images of 38 individuals. We used the cropped and normalized 54×48 face images, which were taken under varying illumination conditions. We randomly split the database into two halves. One half (about 32 images per person) was used as the dictionary, and the other half for testing. Table 1 shows the recognition rates versus feature dimension by NN, NS, SVM, SRC and RSC. It can be seen that RSC achieves better results than the other methods in all dimensions except that RSC is slightly worse than SVM when the dimension is 30. When the dimension is 84, RSC achieves about 3% improvement of recognition rate over SRC. The best recognition rate of RSC is 99.4%, compared to 91.6% for NN, 96.0% for NS, 97.0% for SVM, and 98.3% for SRC.

2) AR database: As in [25], a subset (with only illumination and expression changes) that contains 50 males and 50 females was chosen from the AR dataset [17]. For each subject, the seven images from Session 1 were used for training, with other seven images from Session 2 for testing. The size of image is cropped to 60×43. The comparison of RSC and its competing methods is given in Table 2. Again, we can see that RSC performs much better than all the other four methods in all dimensions except that RSC is slightly worse than SRC when the dimension is 30. Nevertheless, when the dimension is too low, all the methods cannot achieve very high recognition rate. On other dimensions, RSC outperforms SRC by about 3%. SVM does not give good results in this experiment because there are not enough training samples (7 samples per class here) and there are high variations between training set and testing set. The maximal recognition rates of RSC, SRC, SVM, NS and NN are 96.0%,

Dim	30	54	120	300
NN	62.5%	68.0%	70.1%	71.3%
NS	66.1%	70.1%	75.4%	76.0%
SVM	66.1%	69.4%	74.5%	75.4%
SRC [25]	**73.5%**	83.3%	90.1%	93.3%
RSC	71.4%	**86.8%**	**94.0%**	**96.0%**

Table 2. Face recognition rates on the AR database

Dim	Sim-S1	Sim-S3	Sur-S2	Sqi-S2
NN	88.7%	47.3%	40.1%	49.6%
NS	89.6%	48.8%	39.6%	51.2%
SVM	88.9%	46.3%	25.6%	47.7%
SRC [25]	93.7%	60.3%	51.4%	58.1%
RSC	**97.8%**	**75.0%**	**68.8%**	**64.6%**

Table 3. Face recognition rates on Multi-PIE database. ('Sim-S1'('Sim-S3'): set with smile in Session 1 (3);'Sur-S2'('Sqi-S2'): set with surprise (squint) in Session 2).

93.3%, 75.4%, 76.0% and 71.3%, respectively.

3) Multi PIE database: The CMU Multi-PIE database [6] contains images of 337 subjects captured in four sessions with simultaneous variations in pose, expression, and illumination. Among these 337 subjects, all the 249 subjects in Session 1 were used as training set. To make the FR more challenging, four subsets with both illumination and expression variations in Sessions 1, 2 and 3, were used for testing. For the training set, as in [22] we used the 7 frontal images with extreme illuminations $\{0, 1, 7, 13, 14, 16, 18\}$ and neutral expression (refer to Fig. 1(a) for examples). For the testing set, 4 typical frontal images with illuminations $\{0, 2, 7, 13\}$ and different expressions (smile in Sessions 1 and 3, squint and surprise in Session 2) are used (refer to Fig. 1(b) for examples with surprise in Session 2, Fig. 1(c) for examples with smile in Session 1, and Fig. 1(d) for examples with smile in Session 3). Here we used the Eigenface with dimensionality 300 as the face feature for sparse coding. Table 3 lists the recognition rates in four testing sets by the competing methods.

From Table 3, we can see that RSC achieves the best performance in all tests, and SRC performs the second best. In addition, all the methods achieve their best results when Smi-S1 is used for testing because the training set is also from Session 1. The highest recognition rate of RSC on Smi-S1 is 97.8%, more than 4% improvement over SRC. From testing set Smi-S1 to set Smi-S3, the variations increase because of the longer data acquisition time interval (refer to Fig. 1(c) and Fig. 1(d)). The recognition rate of RSC drops by 22.8%, while those of NN, NS, SVM and SRC drop by 41.4%, 40.8%, 42.6% and 33.4%. This validates that RSC is much more robust to face variations than the other methods. For the testing sets Sur-S2 and Sqi-S2,

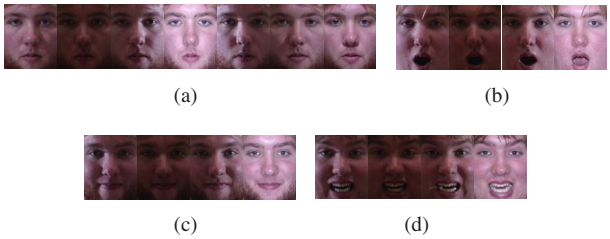

Figure 1. A subject in Multi-PIE database. (a) Training samples with only illumination variations. (b) Testing samples with surprise expression and illumination variations. (c) and (d) show the testing samples with smile expression and illumination variations in Session 1 and Session 3, respectively.

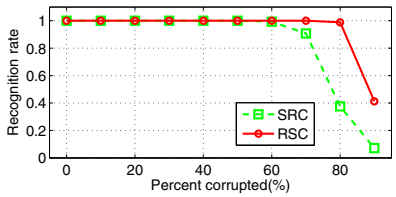

Figure 2. The recognition rate curves of RSC and SRC versus different percentage of corruption.

RSC's recognition rates are 17.4% and 6.5% higher than those of SRC, respectively. Meanwhile, we could also see that FR with surprise expression change is much more difficult than FR with the other two expression changes.

4.3. Face recognition with occlusion

One of the most interesting features of sparse coding based FR in [25] is its robustness to face occlusion by adding an occlusion dictionary (an identity matrix). Thus, in this subsection we test the robustness of RSC to different kinds of occlusions, such as random pixel corruption, random block occlusion and real disguise. In the experiments of random corruption, we compare our proposed RSC with SRC [25]. In the experiments of block occlusion and real disguise, we compare RSC with SRC and the recently developed Gabor-SRC (GSRC) [27].

1) FR with pixel corruption: To be identical to the experimental settings in [25], we used Subsets 1 and 2 (717 images, normal-to-moderate lighting conditions) of the Extended Yale B database for training, and used Subset 3 (453 images, more extreme lighting conditions) for testing. The images were resized to 96×84 pixels. For each testing image, we replaced a certain percentage of its pixels by uniformly distributed random values within $[0, 255]$. The corrupted pixels were randomly chosen for each test image and the locations are unknown to the algorithm.

Fig. 2 shows the results of RSC and SRC under the percentage of corrupted pixels from 0% to 90%. It can be seen

Occlusion	0%	10%	20%	30%	40%	50%
SRC [25]	1	1	0.998	0.985	0.903	0.653
GSRC [27]	1	1	1	0.998	0.965	**0.874**
RSC	**1**	**1**	**1**	**0.998**	**0.969**	0.839

Table 4. The recognition rates of RSC, SRC and GSRC under different levels of block occlusion.

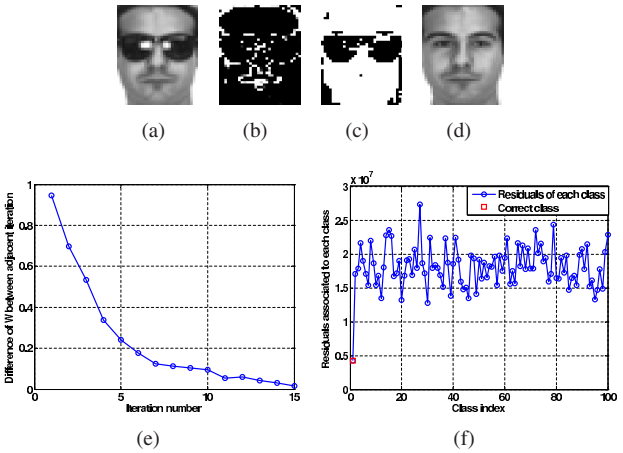

Figure 3. An example of face recognition with disguise using RSC. (a) A test image with sunglasses. (b) The initialized weight map (binarized). (c) The weight map (binarized) when IRSC converges. (d) A template image of the identified subject. (e) The convergence curve of IRSC. (f) The residuals of each class by RSC.

that when the percentage of corrupted pixels is between 10% and 60%, both RSC and SRC correctly classify all the testing images. However, when the percentage of corrupted pixels is more than 60%, the advantage of RSC over SRC is clear. Especially, RSC can still have a recognition rate of 98.9% when 80% pixels are corrupted, while SRC only has a recognition rate of 37.5%.

2) FR with block occlusion: In this part, we test the robustness of RSC to block occlusion. We also used the same experimental settings as in [25], i.e. Subsets 1 and 2 of Extended Yale B for training and Subset 3 for testing. The images were resized to 96×84. Here we set $\tau = 0.7$. Table 4 lists the results of RSC, SRC and GSRC. We see that RSC achieves much higher recognition rates than SRC when the occlusion percentage is larger than 30% (more than 18 % (6%) improvement at 50% (40%) occlusion). Compared to GSRC, RSC still gets competing results without using Gabor features.

3) FR with real face disguise: A subset from the AR database is used in this experiment. This subset consists of 2,599 images from 100 subjects (about 26 samples per class), 50 males and 50 females. We do two tests: one follows the experimental setting in [25], while the other one is more challenging. The images were resized to 42×30. (For simplicity, we let $\delta = 120$, $\mu = 0.1$, $\lambda = 100$, and did not normalize the face images to have unit l_2-norm).

In the first test, 799 images (about 8 samples per subject) of non-occluded frontal views with various facial expressions in Sessions 1 and 2 were used for training, while two separate subsets (with sunglasses and scarf) of 200 images (1 sample per subject per Session, with neutral expression) for testing. Fig. 3 illustrates the classification process of RSC by using an example. Fig. 3(a) shows a test image with sunglasses; Figs. 3(b) and 3(c) show the initialized and converged weight maps (which are binarized for better illustration), respectively; Fig. 3(d) shows a template image of the identified subject. The convergence of the IRSC process is shown in Fig. 3(e) and Fig. 3(f) plots the residuals of each class. The detailed FR results of RSC, SRC and GSRC are listed in Table 5. We see that RSC achieves a recognition rate of 99% in FR with sunglasses, 6% and 12% higher than that of GSRC and SRC, while in FR with scarf, much more improvement is obtained (18% and 37.5% higher than GSRC and SRC).

In the second test, we conduct FR with more complex disguise (disguise with variations of illumination and longer data acquisition interval). 400 images (4 neutral images with different illuminations per subject) of non-occluded frontal views in Session 1 were used for training, while the disguised images (3 images with various illuminations and sunglasses or scarf per subject per Session) in Sessions 1 and 2 for testing. Table 6 shows the results of RSC, GSRC and SRC. Clearly, RSC achieves the best results in all the cases. Compared to SRC, RSC advances much on the testing set with scarf, about 60% improvement in each session. Compared to GSRC, over 6% improvement is achieved by RSC for scarf disguise. For the testing set with sunglasses in Session 2, the recognition rate of RSC is about 35% higher than that of GSRC. Surprisingly, GSRC has lower recognition rates than SRC in the testing sets with sunglasses. This is possibly because that Gabor analysis needs relatively high resolution images and favors regions that are rich in local features, i.e. the eyes. In addition, the average drop of RSC's recognition rate from Session 1 to Session 2 is about 16%, compared to 25% for SRC and 30% for GSRC. We also compute the running times of SRC and RSC (both sparse coding by l_1-ls [12] in Matlab with machine of 3.16 GHz and 3.25G RAM), which are 60.08 s (SRC) and 21.30 s (RSC) in average, validating RSC has lower computational cost than SRC in that case.

5. Conclusion

This paper presented a novel robust sparse coding (RSC) model and an effective iteratively reweighted sparse coding (IRSC) algorithm for RSC. One important advantage of RSC is its robustness to various types of outliers (i.e., occlusion, corruption, expression, etc.) because RSC seeks

Algorithms	SRC [25]	GSRC [27]	**RSC**
Sunglasses	87.0%	93%	**99%**
Scarf	59.5%	79%	**97%**

Table 5. Recognition rates of RSC, GSRC and SRC on the AR database with disguise occlusion.

Algorithms	sg-1	sc-1	sg-2	sc-2
SRC [25]	89.3%	32.3%	57.3%	12.7%
GSRC [27]	87.3%	85%	45%	66%
RSC	**94.7%**	**91.0%**	**80.3%**	**72.7%**

Table 6. Recognition rates of RSC, GSRC and SRC on the AR database with sunglasses (sg-X) or scarf (sc-X) in Session X.

for an MLE (maximum likelihood estimation) solution of the sparse coding problem. Its associated IRSC algorithm is essentially a sparsity-constrained robust regression process. We evaluated the proposed method on different conditions, including variations of illumination, expression, occlusion and corruption as the combination of them. The extensive experimental results clearly demonstrated that RSC outperforms significantly previous methods, such as SRC and GSRC, while its computational complexity is comparable or less than SRC.

References

[1] R. Andersen. *Modern methods for robust regression, series: Quantitative applications in the social sciences*. SAGE Publications, 2008. 626

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriengman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE PAMI*, 19(7):711–720, 1997. 625

[3] D. Donoho. For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. Pure and Applied Math.*, 59(6):797–829, 2006. 625

[4] S. H. Gao, I. W. H. Tsang, L. T. Chia, and P. L. Zhao. Local features are not lonely-laplacian sparse coding for image classification. In *CVPR*, 2010. 626

[5] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE PAMI*, 23(6):643–660, 2001. 629

[6] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28:807–813, 2010. 629, 630

[7] B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machine: Global versus component-based approach. In *ICCV*, 2001. 625

[8] J. Hiriart-Urruty and C. Lemarechal. *Convex analysis and minimization algorithms*. Springer-Verlag, 1996. 628

[9] J. Z. Huang, X. L. Huang, and D. Metaxas. Simultaneous image transformation and sparse representation recovery. In *CVPR*, 2008. 625

[10] P. J. Huber. Robust regression: Asymptotics, conjectures and monte carlo. *Ann. Stat.*, 1(5):799–821, 1973. 626

[11] S. H. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE SP*, 56(6):2346–2356, 2008. 626

[12] S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. A interior-point method for large-scale l_1-regularized least squares. *IEEE Journal on Selected Topics in Signal Processing*, 1(4):606–617, 2007. 627, 628, 631

[13] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE PAMI*, 27(5):684–698, 2005. 629

[14] Y. N. Liu, F. Wu, Z. H. Zhang, Y. T. Zhuang, and S. C. Yan. Sparse representation using nonnegative curds and whey. In *CVPR*, 2010. 626

[15] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *NIPS*, 2009. 625

[16] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE IP*, 17(1):53–69, 2008. 625

[17] A. Martinez and R. benavente. The AR face database. Technical Report 24, CVC, 1998. 629

[18] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997. 625

[19] I. Ramirez and G. Sapiro. Universal sparse modeling. Technical report, arXiv:1003.2941v1[cs.IT], University of Minnesota, 2010. 626

[20] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996. 626

[21] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991. 625, 629

[22] A. Wagner, J. Wright, A. Ganesh, Z. H. Zhou, and Y. Ma. Towards a practical face recognition system: Robust registration and illumination by sparse representation. In *CVPR*, 2009. 625, 630

[23] J. J. Wang, J. C. Yang, K. Yu, F. J. Lv, T. Huang, and Y. H. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 626

[24] J. Wright and Y. Ma. Dense error correction via l_1 minimization. *IEEE Transactions on Information Theory*, 56(7):3540–3560, 2010. 625, 626

[25] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE PAMI*, 31(2):210–227, 2009. 625, 626, 628, 629, 630, 631, 632

[26] J. Yang and J. Zhang. Alternating direction algorithms for l_1-problems in compressive sensing. Technical report, Rice University, 2009. 626

[27] M. Yang and L. Zhang. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In *ECCV*, 2010. 625, 630, 631, 632

[28] J. Zhang, R. Jin, Y. M. Yang, and A. G. Hauptmann. Modified logistic regression: An approximation to SVM and its applications in large-scale text categorization. In *ICML*, 2003. 627

[29] W. Zhao, R. Chellppa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Survey*, 35(4):399–458, 2003. 625