Proceeding of the IEEE 28th
Canadian Conference on Electrical and Computer Engineering
Halifax, Canada, May 3-6, 2015

Robust Group Sparse Representation via Half-Quadratic Optimization for Face Recognition

Yong Peng and Bao-Liang Lu*, *Senior Member, IEEE*

***Abstract*— Sparse representation-based classifier (SRC), which represents a test sample with a linear combination of training samples, has shown promise in pattern classification. However, there are two shortcomings in SRC: (1) the ℓ_2-norm used to measure the reconstruction fidelity is noise-sensitive and (2) the ℓ_1-norm induced sparsity did not consider the correlation among the training samples. Furthermore, in real applications, face images with similar variations, such as illumination or expression, often have higher correlation than those from the same subject. Therefore, we propose to improve the performance of SRC from two aspects: (1) replace the noise-sensitive ℓ_2-norm with an M-estimator to enhance its robustness and (2) emphasize the sparsity of the number of classes instead of the number of training samples, which leads to the group sparsity. The proposed robust group sparse representation (RGSR) can be efficiently optimized via alternating minimization under the Half-Quadratic (HQ) framework. Extensive experiments on representative face data sets show that RGSR can achieve competitive performance in face recognition and outperforms several state-of-the-art methods in dealing with various types of noise such as corruption, occlusion and disguise.**

I. INTRODUCTION

Sparse representation (SR) (e.g., [1], [2]) is an efficient statistical signal modeling tool which has become a promising model in many machine learning and computer vision problems. When applied to image clustering or classification, SR represents an image using a small number of atoms parsimoniously chosen out of an over-complete dictionary. The ℓ_0-norm is the original definition of sparsity, which counts the number of non-zero elements in a vector. As the closest convex surrogate, the ℓ_1-norm is widely used as an alternate to measure the sparsity of representation coefficient. Many fast approaches have been proposed to optimize such ℓ_1-norm minimization SR models (e.g., [3]).

Recently, many studies [4], [5] have shown that the ℓ_1-norm induced sparse models perform well in low-correlation settings. However, if samples from the same class or manifold are highly correlated, the ℓ_1-norm minimization will encounter the stability problems. Generally, it tends to randomly select a single representative sample and ignore other correlated samples. This leads to a sparse solution but misses the correlated information in data, which often causes suboptimal performance. Specifically, for face recognition task in uncontrolled environment, the variation information (e.g., illumination and expression) may be more significant than the identity. In this case, it is possible that face images from different subjects with similar variations could have higher correlation than those from the same subject but with different variations. Therefore, we propose to consider the label information of training samples and emphasize the sparsity of the number of classes instead of the number of training samples, which leads the group sparsity.

Moreover, for most real-world applications, data are usually noisy or significantly corrupted. The original SR and its many variants usually use the sum of squared error or the ℓ_2-norm error function to measure the quality of signal reconstruction, which implicitly assumes that the noise follows the Gaussian distribution. However, it is not the case for real world problems which do not conform to the assumptions made by the model. The least-squares error is sensitive to outliers, which will greatly degrade the quality of approximation if there exists a single corrupted point. Therefore, it is necessary to replace the quadratic form of residuals by lowering down the weight of noisy or corrupted region of samples. Instead of minimizing the non-quadratic and possibly non-convex loss function, we propose to use the M-estimator [6] technique, which can be optimized by HQ minimization.

By conducting extensive experiments on representative face data sets, the results show that RGSR achieves competitive performance in classification. RGSR outperforms state-of-the-art methods in dealing with various types of noise such as corruption, occlusion and disguise.

The rest of this paper is organized as follows. In Section II, we give a brief overview of SRC and the HQ minimization. The proposed RGSR model will be presented in Section III. In Section IV, we conduct experiments to show the effectiveness of RGSR. Section V concludes the paper.

II. RELATED WORK

A. Sparse Representation-Based Classifier

The SRC was proposed in [7] for face recognition. Generally, the dictionary matrix $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_c]$ is formed by stacking the training samples together, where $\mathbf{A}_i$ is the subset of training samples from class i and c is the number of classes. For each test sample $\mathbf{y}$, the sparse representation coefficient can be computed via ℓ_1-norm regularized minimization problem

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\alpha} \|\mathbf{y} - \mathbf{A}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \qquad (1)$$

This work was partially supported by the National Basic Research Program of China (Grant No.2013CB329401), the National Natural Science Foundation of China (Grant No.61272248), and the Science and Technology Commission of Shanghai Municipality (Grant No.13511500200).

Yong Peng and Bao-Liang Lu are with the Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering and Department of Computer Science and Engineering, Shanghai Jiao Tong University. Corresponding author: Bao-Liang Lu (bllu@sjtu.edu.cn)

978-1-4799-5829-0/15/$31.00 ©2015 IEEE

where λ is the tradeoff parameter; then the classification is made by $\text{identity}(\mathbf{y}) = \arg\min_i\{error_i\}$, where $error_i = \|\mathbf{y}-\mathbf{A}_i\hat{\boldsymbol{\alpha}}_i\|_2$, $\hat{\boldsymbol{\alpha}} = [\hat{\boldsymbol{\alpha}}_1; \hat{\boldsymbol{\alpha}}_2; \cdots; \hat{\boldsymbol{\alpha}}_c]$ and $\hat{\boldsymbol{\alpha}}_i$ is the coefficient vector associated with the i-th class. It was claimed in [7] that the success of SRC is mainly caused by the ℓ_1-norm sparsity imposed on the coding efficient. However, this ℓ_1-norm induced sparsity treats each element in $\boldsymbol{\alpha}$ equally, which does not consider the correlation of samples in dictionary $\mathbf{A}$. Therefore, it performs well only when $\mathbf{A}$ is under low-correlation settings. In this paper, we use training data $\mathbf{X} \in \mathbb{R}^{d\times n}$ (d, n denote the dimensionality and number of training samples, respectively) as dictionary.

B. The Half-Quadratic Minimization

This section reviews the background of half-quadratic modeling based on conjugate function theory [8], [9] for convex or non-convex minimization.

Conjugate Function. Given a differentiable function $f(\mathbf{v})$: $\mathcal{S} \subseteq \mathbb{R}^n \to \mathbb{R}$, the conjugate $f^*(\mathbf{p})$: $\mathbb{R}^n \to \mathbb{R}$ of the function $f(\cdot)$ is defined as [10]

$$f^*(\mathbf{p}) = \inf_{\mathbf{v}\in S} \mathbf{p}^T\mathbf{v} - f(\mathbf{v}). \tag{2}$$

The domain of $f^*(\mathbf{p})$ is bounded above on $\mathcal{S}$ [10]. $f^*(\mathbf{p})$ is the pointwise supremum of a family of convex functions of $\mathbf{p}$, which is also a convex function. Based on conjugate function theory, a loss function in image restoration and signal recovery can be defined as [11], [12], [13]

$$f(\mathbf{v}) = \min_{\mathbf{p}}\{\psi(\mathbf{v},\mathbf{p}) + \varphi(\mathbf{p})\}, \tag{3}$$

where $f(\cdot)$ is a potential loss function such as a certain M-estimator, $\mathbf{v}$ is a set of adjustable parameters of a linear system, $\mathbf{p}$ is an auxiliary variable in HQ optimization, $\psi(\mathbf{v},\mathbf{p})$ is a quadratic function, and $\varphi(\cdot)$ is the dual potential function of $f(\cdot)$.

For face recognition application, we use the multiplicative form quadratic function of $\psi(\mathbf{v},\mathbf{p})$ as $\psi(\mathbf{v},\mathbf{p}) \doteq \sum_i p_i v_i^2$, where v_i is the coding residual for each pixel and p_i is the learned weight for such pixel. p_i will be a smaller value which can alleviate its influence if such pixel is corrupted. Thus, the learned $\mathbf{p}$ can adjust the influence of each pixel according to their corruption level. However, the widely used ℓ_2-norm loss function, actually employs the constant weight despite of whether the pixel is corrupted or not.

III. Robust Group Sparse Representation

Specifically, using group sparse representation w.r.t. a test sample $\mathbf{y}$, we have

$$\phi(\mathbf{e}) = \min_w\{\psi(\mathbf{e},\mathbf{w}) + \varphi(\mathbf{w})\}, \tag{4}$$

where $\mathbf{e} \triangleq \mathbf{X}\boldsymbol{\alpha} - \mathbf{y} \in \mathbb{R}^d$ and $\mathbf{w} \in \mathbb{R}^d$ are the coding residual and the pixel-level weight for face image, respectively. Here we consider ψ in multiplicative form $\psi(\mathbf{e},\mathbf{v}) = \sum_{i=1}^d w_i e_i^2$ which plays the role as *error detection* [14].

The first term in (1) which uses the ℓ_2-norm to measure coding residual can be easily dominated by a few outliers with large errors. This can be illustrated in Fig. 1, where the ℓ_2-norm induces more penalty for large fitting errors than the

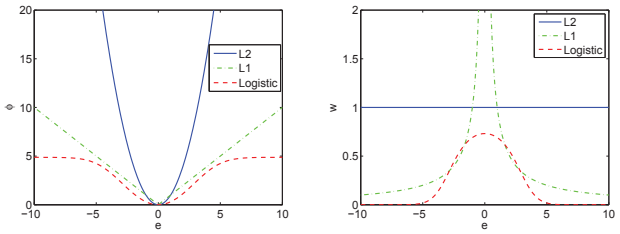

Fig. 1. Potential loss functions and their corresponding weight functions in half-quadratic minimization.

ℓ_1-norm and Logistic loss function (a type of M-estimator we use in this paper). Accordingly, the ℓ_2-norm uses a constant weight for both small and large errors. However, M-estimator can learn the weight $\mathbf{w}$ to adapt whether the pixel is corrupted or not, which can greatly alleviate the influence of outliers. In general, M-estimator uses small weight w_i for large e_i to make it robust to outliers as shown in Fig. 1.

Therefore, by replacing the ℓ_2-norm with M-estimator $\phi(\cdot)$, we can obtain the following objective

$$\min_\alpha \phi(\mathbf{X}\boldsymbol{\alpha} - \mathbf{y}) + \lambda\mathcal{R}(\boldsymbol{\alpha}), \tag{5}$$

where $\mathcal{R}(\boldsymbol{\alpha})$ is the group sparsity regularizer which will be explained later. Using the multiplicative form of ψ as $\psi(\mathbf{X}\boldsymbol{\alpha}-\mathbf{y},\mathbf{w}) = \sum_{i=1}^d w_i(y_i - \sum_{j=1}^n x_{ij}\alpha_j)^2$, we have the following minimization of the augmented objective

$$\min_{\alpha,w} \sum\nolimits_i (w_i(y_i - \sum\nolimits_j x_{ij}\alpha_j)^2 + \varphi(w_i)) + \lambda\mathcal{R}(\boldsymbol{\alpha}). \tag{6}$$

We will use $J(\boldsymbol{\alpha},\mathbf{w})$ to denote the above objective function. Following the HQ optimization framework [11], [14], a local minimizer $(\boldsymbol{\alpha},\mathbf{w})$ of $J(\boldsymbol{\alpha},\mathbf{w})$ can be alternately calculated by the following rules

$$w_i^{t+1} = \omega(y_i - \sum\nolimits_j x_{ij}\alpha_j^t), \tag{7}$$

$$\boldsymbol{\alpha}^{t+1} = \arg\min_\alpha \|\mathbf{W}^{1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}^t)\|_2^2 + \lambda\mathcal{R}(\boldsymbol{\alpha}^t), \tag{8}$$

where $\boldsymbol{\alpha}^t$ is an estimated coefficient vector for the t-th iteration, $\omega(\cdot)$ is the weight function derived from the conjugate of $\phi(\cdot)$. $\omega(\cdot)$ satisfies that

$$\psi(e_i,\omega(e_i)) + \varphi(\omega(e_i)) \le \psi(e_i,w_i) + \varphi(w_i). \tag{9}$$

Here, $\mathbf{W}$ is a diagonal matrix with each entry $W_{ii} = w_i^{t+1}$. The optimization of $\boldsymbol{\alpha}^{t+1}$ can be rewritten as the following regularized quadratic problem

$$\boldsymbol{\alpha}^{t+1} = \arg\min_\alpha \|\hat{\mathbf{X}}\boldsymbol{\alpha} - \hat{\mathbf{y}}\|_2^2 + \lambda\mathcal{R}(\boldsymbol{\alpha}), \tag{10}$$

where $\hat{\mathbf{X}} = \sqrt{\mathbf{W}}\mathbf{X}$ and $\hat{\mathbf{y}} = \sqrt{\mathbf{W}}\mathbf{y}$. The robust improvement of group sparse representation is given in Alg. 1.

Based on HQ framework [11], [14], we use the Logistic weight function to determine $\mathbf{w}$ for fair comparison with RSC in [15], whose loss function $\phi(\cdot)$ and weight function $\omega(\cdot)$ (as shown in Fig. 1) are respectively defined as

$$\phi(e_i) = \frac{-1}{2\mu}\ln\frac{1+\exp(\mu\delta - \mu e_i^2)}{1+\exp(\mu\delta)}, \tag{11}$$

$$\omega(e_i) = \frac{\exp(\mu\delta - \mu e_i^2)}{1+\exp(\mu\delta - \mu e_i^2)}, \tag{12}$$

where parameter μ controls the decreasing rate of weight and parameter δ is the demarcation point.

Algorithm 1 Robust Improvement Based on HQ

Input: Training data $\mathbf{X}$, test sample $\mathbf{y}$ and regularization parameter λ, initial guess $\boldsymbol{\alpha}^0$;

Output: Representation coefficient $\boldsymbol{\alpha}$, weight vector $\mathbf{w}$.

$t = 0$;
while not converged **do**
  $W_{ii}^{t+1} = \omega(y_i - \sum_{j=1}^{n} x_{ij}\alpha_j^t)$;
  $\hat{\mathbf{X}} = \sqrt{\mathbf{W}^{t+1}}\mathbf{X}$ and $\hat{\mathbf{y}} = \sqrt{\mathbf{W}^{t+1}}\mathbf{y}$;
  $\boldsymbol{\alpha}^{t+1} = \arg\min_\alpha \|\hat{\mathbf{X}}\boldsymbol{\alpha} - \hat{\mathbf{y}}\|_2^2 + \lambda\mathcal{R}(\boldsymbol{\alpha})$;
  $t = t + 1$;
end while

Consider $\boldsymbol{\alpha} = [\alpha_1^1, \cdots, \alpha_{|S_1|}^1, \cdots, \alpha_1^{|S_c|}, \cdots, \alpha_{|S_c|}^{|S_c|}]$, where $\{S_k\}, k = 1, 2, \cdots, c$ is the partition of training samples from different classes and $|S_k|$ is the number of samples in class k, and then the RGSR model can be reformulated as

$$\min_{\alpha,w} \sum\nolimits_i (w_i(y_i - \sum\nolimits_j x_{ij}\alpha_j)^2 + \varphi(w_i)) + \lambda \sum\nolimits_{S_k} \|\boldsymbol{\alpha}_{S_k}\|_2. \tag{13}$$

Obviously, (10) is equivalent to

$$\boldsymbol{\alpha}^{t+1} = \arg\min_\alpha \|\hat{\mathbf{X}}\boldsymbol{\alpha} - \hat{\mathbf{y}}\|_2^2 + \lambda \sum\nolimits_{S_k} \|\boldsymbol{\alpha}_{S_k}\|_2. \tag{14}$$

Set its derivative w.r.t. $\boldsymbol{\alpha}$ to zero and we can obtain a simple method for updating $\boldsymbol{\alpha}^{t+1}$ as

$$\hat{\boldsymbol{\alpha}} = (\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \lambda\mathbf{L})^{-1}\hat{\mathbf{X}}^T\hat{\mathbf{y}}, \tag{15}$$

$$\mathbf{L} = \begin{bmatrix} \frac{1}{2\|\boldsymbol{\alpha}_{S_1}\|_2}\mathbf{I}_{|S_1|} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \frac{1}{2\|\boldsymbol{\alpha}_{S_c}\|_2}\mathbf{I}_{|S_c|} \end{bmatrix}. \tag{16}$$

The whole procedure of optimizing the RGSR model is summarized in Alg. 2. The stop criteria for the outer loop and inner loop are respectively defined as

$$\|\mathbf{w}^{t+1} - \mathbf{w}^t\|_2 / \|\mathbf{w}^t\|_2 < \varepsilon_1$$
$$|\mathrm{obj}^{k+1} - \mathrm{obj}^k| / |\mathrm{obj}^k| < \varepsilon_2,$$

where obj is the objective value of (14), ε_1 and ε_2 are small positive values (0.05 and 0.001 in our experiments).

The diagram of RGSR is shown in Fig. 2. The convergence analysis of Alg. 2 will be given below.

Lemma 1: [16] For arbitrary two non-zero vectors $\mathbf{u}$ and $\mathbf{v}$, the following inequality holds

$$\|\mathbf{u}\|_2 - \frac{\|\mathbf{u}\|_2^2}{2\|\mathbf{v}\|_2} \le \|\mathbf{v}\|_2 - \frac{\|\mathbf{v}\|_2^2}{2\|\mathbf{v}\|_2}.$$

Theorem 2: The alternating optimization of objective $J(\boldsymbol{\alpha}, \mathbf{w})$ in (13) by Alg. 2 converges.

Proof: First we show that the inner loop monotonically decreases the objective of (14). It can be easily verified that (15) is the solution to the following problem

$$\hat{\boldsymbol{\alpha}} = \arg\min_\alpha \|\hat{\mathbf{X}}\boldsymbol{\alpha} - \hat{\mathbf{y}}\|_2^2 + \lambda\boldsymbol{\alpha}^T\mathbf{L}\boldsymbol{\alpha}$$

Algorithm 2 Robust Supervised Trace Lasso Model

Input: Training data $\mathbf{X}$, test sample $\mathbf{y}$, regularization parameter λ and initial guess $\mathbf{y}_{rec}^0$;

Output: Coefficient $\boldsymbol{\alpha}$ and feature weight vector $\mathbf{w}$.

$t = 0$;
// Outer loop for optimizing weight vector $\mathbf{w}$
while not converged **do**
  Compute residual $\mathbf{e}^{t+1} = \mathbf{y} - \mathbf{y}_{rec}^t$;
  Compute $\mathbf{w}^{t+1} = \omega(\mathbf{e}^t)$ based on the selected M-estimator and let $\mathbf{W}^{t+1} = \mathrm{Diag}(\mathbf{w}^{t+1})$;
  $\hat{\mathbf{X}} = \sqrt{\mathbf{W}^{t+1}}\mathbf{X}$ and $\hat{\mathbf{y}} = \sqrt{\mathbf{W}^{t+1}}\mathbf{y}$;
  // Inner loop for optimizing $\boldsymbol{\alpha}^{t+1}$ based on (14)
  $k = 0$;
  Initialize $\boldsymbol{\alpha}^k$;
  while not converged **do**
    Compute $\mathbf{L}^k$ based on (16);
    $\boldsymbol{\alpha}^{k+1} = (\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \lambda\mathbf{L}^k)^{-1}\hat{\mathbf{X}}^T\hat{\mathbf{y}}$;
    $k = k + 1$;
  end while
  Compute the reconstruction $\mathbf{y}_{rec}^{t+1} = \mathbf{X}\boldsymbol{\alpha}^{t+1}$;
  $t = t + 1$;
end while

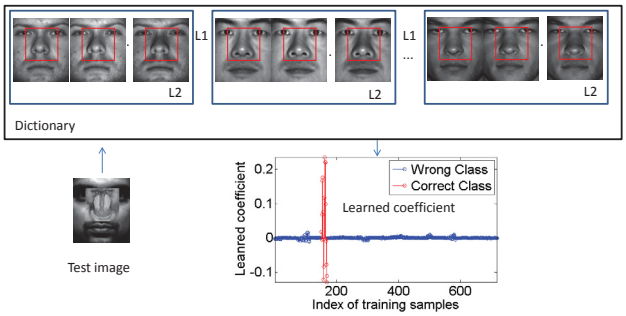

Fig. 2. The diagram of RGSR model. For an occluded test sample, RGSR learns the weight map which can mask the corresponding area of training samples shown in red rectangle. The learned coefficients for intra-class samples are measured by ℓ_2-norm, while that for inter-class samples are measured by ℓ_1-norm. This results in the structured sparsity.

and thus we have

$$\|\hat{\mathbf{X}}\hat{\boldsymbol{\alpha}} - \hat{\mathbf{y}}\|_2^2 + \lambda\hat{\boldsymbol{\alpha}}^T\mathbf{L}\hat{\boldsymbol{\alpha}} \le \|\hat{\mathbf{X}}\boldsymbol{\alpha} - \hat{\mathbf{y}}\|_2^2 + \lambda\boldsymbol{\alpha}^T\mathbf{L}\boldsymbol{\alpha}. \tag{17}$$

Based on Lemma 1, the following inequalities hold

$$\lambda\sum_{S_k}\|\hat{\boldsymbol{\alpha}}_{S_k}\|_2 - \lambda\sum_{S_k}\frac{\|\hat{\boldsymbol{\alpha}}_{S_k}\|_2^2}{2\|\boldsymbol{\alpha}_{S_k}\|_2} \le \lambda\sum_{S_k}\|\boldsymbol{\alpha}_{S_k}\|_2 - \lambda\sum_{S_k}\frac{\|\boldsymbol{\alpha}_{S_k}\|_2^2}{2\|\boldsymbol{\alpha}_{S_k}\|_2}$$

$$\lambda\sum_{S_k}\|\hat{\boldsymbol{\alpha}}_{S_k}\|_2 - \lambda\hat{\boldsymbol{\alpha}}^T\mathbf{L}\hat{\boldsymbol{\alpha}} \le \lambda\sum_{S_k}\|\boldsymbol{\alpha}_{S_k}\|_2 - \lambda\boldsymbol{\alpha}^T\mathbf{L}\boldsymbol{\alpha} \tag{18}$$

Add both sides of (17) and (18) together and we can obtain

$$\|\hat{\mathbf{X}}\hat{\boldsymbol{\alpha}} - \hat{\mathbf{y}}\|_2^2 + \lambda\sum_{S_k}\|\hat{\boldsymbol{\alpha}}_{S_k}\|_2 \le \|\hat{\mathbf{X}}\boldsymbol{\alpha} - \hat{\mathbf{y}}\|_2^2 + \lambda\sum_{S_k}\|\boldsymbol{\alpha}_{S_k}\|_2,$$

which means that the solution in optimizing $\boldsymbol{\alpha}$ satisfies $J(\boldsymbol{\alpha}^{t+1}, \mathbf{w}^{t+1}) \le J(\boldsymbol{\alpha}^t, \mathbf{w}^{t+1})$.

According to the property of weight function $\omega(\cdot)$ shown in inequality (9), we have that for a fixed $\boldsymbol{\alpha}^{t+1}$,

$J(\boldsymbol{\alpha}^t, \mathbf{w}^{t+1}) \leq J(\boldsymbol{\alpha}^t, \mathbf{w}^t)$. Combining with the recent conclusion $J(\boldsymbol{\alpha}^{t+1}, \mathbf{w}^{t+1}) \leq J(\boldsymbol{\alpha}^t, \mathbf{w}^{t+1})$, we get

$$J(\boldsymbol{\alpha}^{t+1}, \mathbf{w}^{t+1}) \leq J(\boldsymbol{\alpha}^t, \mathbf{w}^{t+1}) \leq J(\boldsymbol{\alpha}^t, \mathbf{w}^t).$$

Thus, $\{\cdots, J(\boldsymbol{\alpha}^t, \mathbf{w}^t), J(\boldsymbol{\alpha}^t, \mathbf{w}^{t+1}), J(\boldsymbol{\alpha}^{t+1}, \mathbf{w}^{t+1}), \cdots\}$ generated by Alg. 2 converges as $t \to \infty$. ■

IV. Experimental Studies

We evaluate the performance of RGSR on benchmark face data sets: AR [17] and Extended Yale B [18], [19]. We conduct experiments under two settings: (1) FR without occlusion but with variations such as illumination and expression changes and (2) FR with three types of occlusions: random pixel corruption/block occlusion and real disguise.

There are three parameters involved in RGSR model: the regularization parameter λ and the Logistic weight function related parameters (μ, δ). In this paper, λ is set as 0.001 by default. According to the properties of (μ, δ), smaller δ (larger μ) is encouraged if the image is grossly corrupted, which can group more pixels into outliers. For a corrupted image, the squared error vector is $\boldsymbol{\pi} = [e_1^2, e_2^2, \cdots, e_d^2]$ (e_i is the coding residual w.r.t. the i-th pixel) and its ascending sorted version is $\boldsymbol{\pi}_a$. We set δ as $\boldsymbol{\pi}_a(\lfloor \tau d \rfloor)$ and $\mu = c/\delta$. Thus, two new parameters (c, τ) are introduced to build the tight connection to the corruption level instead of using (μ, δ) directly [15]. In our experiments, the corruption level for the second setting is higher than the first one and smaller τ is preferred; thus we set (c, τ) respectively as (8,0.8) and (8,0.6) for both settings.

A. Face Recognition without Occlusion

In this experimental setting, we compare RGSR with nearest neighbor (NN), nearest subspace (NS), linear support vector machine, SRC [7], collaborative representation based classification (CRC) [20] and RSC [15].

Similar to general FR methods, we perform experiments in the PCA subspace, in which the Eigenface [21] features are used as input. By applying PCA to the training data, (14) will become $\min_\alpha \|\mathbf{P}(\hat{\boldsymbol{X}}\boldsymbol{\alpha} - \hat{\boldsymbol{y}})\|_2^2 + \lambda \sum_{S_k} \|\boldsymbol{\alpha}_{S_k}\|_2$, where $\mathbf{P}$ is the projection matrix.

1) AR: As in [7], a subset with only illumination and expression changes which contains 50 males and 50 females was chosen from the AR data set [17] in our experiments. For each subject, the seven images from Session 1 were used for training, the other seven images from Session 2 for testing. The image size is cropped to 60×43 pixels. The comparison of RGSR and its competing methods is given in Table I. RGSR achieves the best results among all methods in all dimensions. RGSR consistently performs better than RSC because the structured sparsity is encouraged than the ℓ_1-norm induced flat sparsity.

2) Extended Yale B: The Extended Yale B data set [18], [19] contains about 2414 frontal face images from 38 individuals. We used the cropped and normalized 54×48 images, which were taken under varying illuminations. We randomly split the database into two halves. One half (about 32 images per subject) was used as training samples, and the other half for testing. Table II shows the recognition rates versus feature dimension by the competing methods. RSGR has much performance improvement in higher dimensions. In this experiment, the training samples from each class are sufficient (about 32) and they are more uncorrelated in lower dimensional subspace when comparing with AR data set; thus the ℓ_1-norm is more appropriate to regularize the representation of samples with big variations. RGSR has limited improvement over RSC in higher dimensional subspace.

TABLE I
Face recognition rates on AR.

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%	76.0%
SRC[7]	73.5%	83.3%	90.1%	93.3%
CRC[20]	64.4%	80.5%	90.0%	93.4%
RSC[15]	71.4%	86.8%	94.0%	96.0%
RGSR	**73.7%**	**87.7%**	**94.4%**	**96.7%**

TABLE II
Face recognition rates on Extended Yale B.

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[7]	89.1%	95.1%	96.8%	97.9%
CRC[20]	74.0%	92.9%	96.5%	98.0%
RSC[15]	91.3%	**98.1%**	98.4%	99.4%
RGSR	88.2%	96.4%	**98.6%**	**99.6%**

B. Face Recognition with Occlusion

In this section, we test the robustness of RGSR to different types of occlusions including random pixel corruption, random block occlusion and real disguise.

1) Face Recognition with Random Pixel Corruption: To be identical to the experimental settings in [7], 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 face images are resized to 96×84 pixels. For each test image, we replaced a certain percentage of its pixels by uniformly distributed random values within $[0, 255]$. The corrupted pixels were randomly chosen from test image and the locations are unknown.

We compare RGSR with SRC, CRC, correntropy-based sparse representation (CESR) [22] and RSC. Fig. 3 shows the results of different models under the corruption level from 0% to 90%. All the models except CRC perform well when the corruption level is lower than 60%. However, when the percentage is more than 60%, the performance of SRC was greatly reduced. Even with 90% pixels corrupted, RGSR still obtains an acceptable accuracy (55.85%). A representative example of RSC and RGSR with 80% random pixel corruption is shown in Fig. 4. The corrupted face image is difficult to recognize even for human; however,

both RSC and RGSR can accurately estimate the weight map and recover the clean image. Both the corrupted pixels and shadow region are reflected in the learned weight maps. The reconstructed images are faithful to the original image but with better visual quality. From the learned coefficients, we find only one sample from the correct class plays a main role in reconstruction for RSC; while for RGSR, all the samples from the correct class have large coefficients. Therefore, the reconstructed face image by RGSR is cleaner than that by RSC especially for the right half face (lower illumination). The coefficients obtained by RGSR has obvious grouping effect and are smoother than those of RSC.

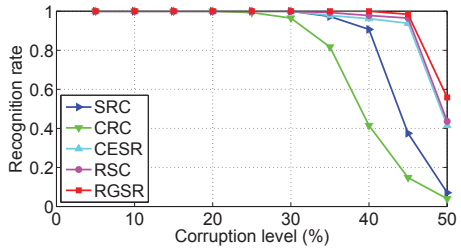

Fig. 3. Recognition rates versus different percentage of pixel corruption.

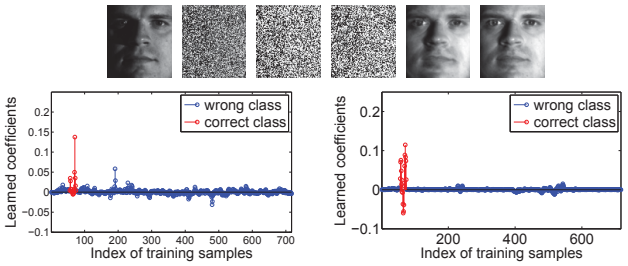

Fig. 4. Recognition under 80% random pixel corruption. **First-row** are respectively the original image, corrupted image, weight maps obtained via RSC and RGSR, reconstructed images via RSC and RGSR; **Second-row:** learned coefficients via RSC and RGSR (best viewed in color).

2) Face Recognition with Block Occlusion: In this part, we test the robustness of RGSR model to block occlusion. We also used the same experimental settings as in [7], i.e., Subsets 1 and 2 of Extended Yale B for training and Subset 3 for testing. The images were resized to 96×84 pixels. We compare RGSR with SRC, CRC, Gabor-SRC (use Gabor features to construct the occlusion dictionary) [23], CESR and RSC. Fig. 5 shows the change trend of different models under the level of the occluded area from 0% to 50%. Obviously, RGSR gets promising results even if the occlusion level is high. Fig. 6 gives a representative example under 40% random block occlusion. From the coefficients learned by RSC, we can find that many training samples from the wrong classes contribute to the reconstruction, which blurs the area around the lip in the reconstructed image. There are only three non-zero values w.r.t. the samples from correct class, which means that the ℓ_1-norm sparsity encourages to select representative samples when they are highly correlated. For RGSR, the reconstruction is mainly achieved by the training samples from the correct class because they have similar non-zero values and samples from wrong classes have near-zero values.

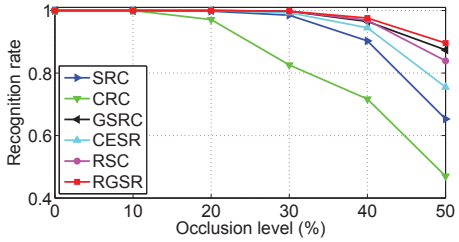

Fig. 5. Recognition rates versus different percentage of block occlusion.

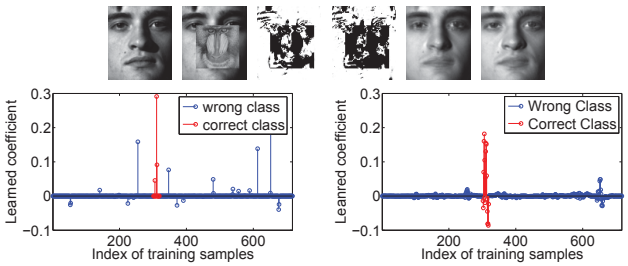

Fig. 6. Recognition under 40% block occlusion. **First-row** are respectively the original image, occluded image, weight maps obtained via RSC and RGSR, reconstructed images via RSC and RGSR; **Second-row:** learned coefficients via RSC and RGSR (best viewed in color).

3) Face Recognition with Real Disguise: A subset from AR data set is used in this experiment, which consists of 2,599 face images from 100 subjects (about 26 samples per subject), 50 males and 50 females. We conduct two tests: one follows the experimental setting in [7], while the other follows [15] and is more challenging. The images are resized to 42×30 pixels.

In the first test, 800 images (8 samples per subject) of non-occluded frontal views with various facial expressions in Session 1 and 2 were used for training, while two separate subsets (with sunglasses and scarves) of 200 images (1 sample per subject per Session, with neutral expression) for testing. The recognition rates of different models are listed in Table III. RGSR achieves 100% recognition rate under the sunglass disguise and 97.5% under the scarf disguise, which are respectively 13% and 38% improvements w.r.t. SRC. Though RSC performs well on both disguises, RGSR still has respectively 1.5% and 1% improvement over it.

TABLE III
RECOGNITION RATES ON AR WITH DISGUISE OCCLUSION.

Algorithms	Sunglasses	Scarves
SRC [7]	87.0%	59.5%
CRC [20]	68.5%	90.5%
GSRC [23]	93%	79%
CESR [22]	99%	42%
RSC [15]	98.5%	96.5%
RGSR	**100%**	**97.5%**

In the second test, we use more complex disguises (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 disguise images (3 images with various illuminations and sunglasses or scarves per subject per Session) in Session 1 and 2 for testing. Table IV shows the results of different competing models. RGSR obtains much improvement w.r.t. RSC, about 4.3% (Session 1) and 6.4% (Session 2) for the sunglass disguise; for the scarf disguise, the improvements are respectively 2% (Session 1) and 4% (Session 2). Fig. 7 illustrates the classification process of RGSR on a representative example. Compared to RSC, the reconstructed image by RGSR has better visual quality around the eye corner for the disguised test image, which can easily remove the sunglass disguise. The coefficients learned by RGSR have obvious grouping effect, which enforces training samples from the same class have similar coefficients. And there are samples from only a few wrong classes which have large values. But for RSC, the coefficients have large values across each class and correspondingly the coding residual for each class has similar variation tendency.

TABLE IV

RECOGNITION RATES ON AR WITH SUNGLASSES OR SCARF IN SESSION 1 AND SESSION 2.

Algorithms	Sg-s1	Sc-s1	Sg-s2	Sc-s2
SRC [7]	89.3%	32.3%	57.3%	12.7%
CRC [20]	43.7%	30.7%	17.7%	13.7%
GSRC [23]	87.3%	85.0%	45.0%	66.0%
CESR [22]	95.3%	38%	79%	20.7%
RSC [15]	94.7%	91.0%	80.3%	72.7%
RGSR	**99.0%**	**93.0%**	**86.7%**	**76.7%**

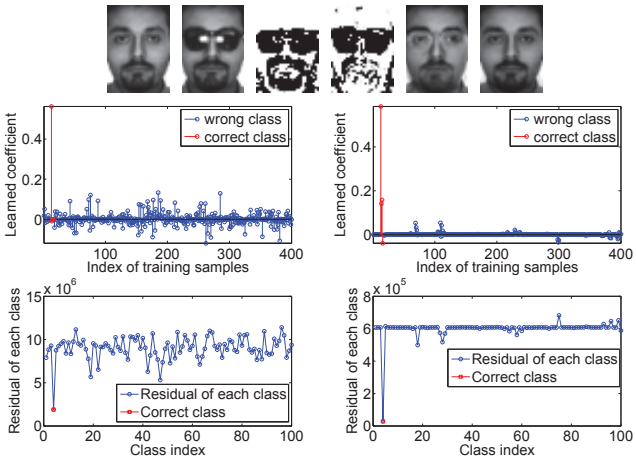

Fig. 7. An example of FR with disguise. **First-row** are respectively face without disguise, sunglass disguised test image, the weight maps obtained via RSC and RGSR, the reconstructed images via RSC and RGSR; **Mid-row**: learned coefficients associated with each training sample via RSC and RGSR; **Third-row:** residuals of each class via RSC and RGSR.

V. CONCLUSION

This paper proposed the robust group sparse representation-based classifier by improving SRC from two aspects: using robust M-estimator to measure the representation fidelity and the group sparsity constraint on the coefficients. The optimization method to proposed RGSR model is efficient and we provide its convergence analysis. The RGSR model was evaluated under different conditions, including variations of illuminations, expressions, occlusion and combined corruption. Our experimental results demonstrated that RGSR performs well especially under high-dimensional cases and outperforms many state-of-the-art methods including robust sparse coding.

REFERENCES

[1] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. R. Stat. Soc. B*, pp. 267–288, 1996.

[2] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *P. Natl. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197–2202, 2003.

[3] A. Yang, Z. Zhou, A. Balasubramanian, S. Sastry, and Y. Ma, "Fast ℓ1-minimization algorithms for robust face recognition," *IEEE TIP*, vol. 22, no. 8, pp. 3234–3246, 2013.

[4] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, 2006.

[5] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Stat.*, vol. 37, no. 4, pp. 1705–1732, 2009.

[6] P. J. Huber, *Robust statistics*. Springer, 2011.

[7] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE TPAMI*, vol. 31, no. 2, pp. 210–227, 2009.

[8] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE TPAMI*, vol. 14, no. 3, pp. 367–383, 1992.

[9] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE TIP*, vol. 4, no. 7, pp. 932–946, 1995.

[10] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[11] M. Nikolova and M. K. Ng, "Analysis of half-quadratic minimization methods for signal and image recovery," *SIAM J. Sci. Comput.*, vol. 27, no. 3, pp. 937–966, 2005.

[12] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. & Sim.*, vol. 4, no. 4, pp. 1168–1200, 2005.

[13] J. M. Bioucas-Dias and M. A. Figueiredo, "A new twist: two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE TIP*, vol. 16, no. 12, pp. 2992–3004, 2007.

[14] R. He, W.-S. Zheng, T. Tan, and Z. Sun, "Half-quadratic based iterative minimization for robust sparse representation," *IEEE TPAMI*, vol. 36, no. 2, pp. 261–275, 2013.

[15] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *CVPR*, 2011.

[16] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint ℓ2,1-norms minimization," in *NIPS*, 2010.

[17] A. M. Martinez, "The AR face database," *CVC Technical Report*, 1998.

[18] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE TPAMI*, vol. 23, no. 6, pp. 643–660, 2001.

[19] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE TPAMI*, vol. 27, no. 5, pp. 684–698, 2005.

[20] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *ICCV*, 2011.

[21] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[22] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE TPAMI*, vol. 33, no. 8, pp. 1561–1576, 2011.

[23] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with gabor occlusion dictionary," in *ECCV*, 2010.