Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

A Robust Group-Sparse Representation Variational Method with applications to Face Recognition

(Article begins on next page)

26 June 2024

A Robust Group-Sparse Representation Variational Method with applications to Face Recognition

Fritz Keinert, Damiana Lazzaro, Serena Morigi

September 6, 2021

Abstract

In this paper we propose a Group-Sparse Representation based method with applications to Face Recognition (GSR-FR). The novel sparse representation variational model includes a non-convex sparsity-inducing penalty and a robust non-convex loss function. The penalty encourages group sparsity by using approximation of the ℓ_0-quasinorm, and the loss function is chosen to make the algorithm robust to noise, occlusions and disguises. The solution of the non-trivial non-convex optimization problem is efficiently obtained by a majorization-minimization strategy combined with forward-backward splitting, which in particular reduces the solution to a sequence of easier convex optimization sub-problems. Extensive experiments on widely used face databases show the potentiality of the proposed model and demonstrate that the GSR-FR algorithm is competitive with state-of-the-art methods based on sparse representation, especially for very low dimensional feature spaces.

1 Introduction

Given a sample vector $b \in \mathbb{R}^M$, the goal of the sparse representation problem is to provide a sparse approximation of b over a known dictionary $A \in \mathbb{R}^{M \times N}$ ($M < N$), by a linear combination of a few columns of A, referred to as atoms. Since real samples contain in general noise or outliers, the sparse representation model includes a small possible corruption modeled by a vector $e \in \mathbb{R}^M$ and reads as

$$b = A\alpha + e. \tag{1}$$

A sparse representation method demands that the solution $\alpha \in \mathbb{R}^N$ of the underdetermined linear system (1) is a sparse vector by imposing on α a suitable sparsity-inducing penalty term. Therefore the coefficients α can be recovered by solving a minimization problem of the form

$$\alpha^* = \arg \min_{\alpha \in \mathbb{R}^N} F(\alpha) \quad \text{subject to} \quad H(A\alpha - b) \leq \epsilon, \tag{2}$$

where $F(\cdot)$ represents a sparsity-inducing *penalty* function, and $H(\cdot)$ is a suitable loss function, named *fidelity* term, which ensures that the given observation b can be faithfully represented by the dictionary A.

Common choices for imposing the sparsity in the solution are the ℓ_0-quasinorm (number of nonzero elements in the vector), which leads to an intractable numerical problem, or the ℓ_1-norm ($F(\cdot) = \|\cdot\|_1$). It's well known that using a regularization term which closely approximates the ℓ_0-quasinorm can better recover vectors with more nonzero coefficients than using the ℓ_1-norm, see [13] and references therein.

More recently, to improve the robustness, especially in case of noise corruptions (e.g. outliers), the popular ℓ_2-norm fidelity term ($H(\cdot) = \|\cdot\|_2^2$) has been replaced by the ℓ_1-norm fidelity term ($H(\cdot) = \|\cdot\|_1$), which assumes that the residual ($b - A\alpha$) follows a Laplacian distribution instead of a Gaussian distribution, [36].

In addition to the sparsity requirement, in our proposal we farther assume that α is a group-sparse vector, i.e. that the few large magnitude values tend to form clusters (groups), that the group sizes are known, and that the groups have well-defined boundaries. We propose a variational sparse representation model in the form (2) where the penalty is a superposition of penalties for groups of contiguous coefficients. The individual terms in the superposition are concave functions of the group magnitudes, which can be viewed as encouraging group-sparsity. More specifically, we solve an equivalent unconstrained form of the sparse representation problem (2) which reads as

$$\alpha^* = \arg \min_{\alpha \in \mathbb{R}^N} \left\{ J(\alpha; \mu, \lambda) = F_\mu(\alpha) + \frac{1}{\lambda} H(A\alpha - b) \right\}, \tag{3}$$

where $\lambda > 0$ is a parameter that controls the trade-off between the two terms, F_μ is a sparsity-inducing penalty which is a non-separable, non-convex parametric function depending on a parameter μ to control the degree of concavity; as $\mu \to 0$, F_μ approaches the ℓ_0-quasinorm for groups. That is, $F_\mu(\alpha)$ approximately counts the number of groups in which α has non-negligible coefficients. The data fidelity H is a non-convex function which improves the robustness in case of noisy data.

The objective function $J(\alpha; \mu, \lambda)$ in (3) is thus, in general, non-convex. To avoid the numerical intricacies arising from non-convex optimization, we approximate α^*, defined as a minimizer of the cost function $J(\alpha; \mu, \lambda)$, by exploiting a graduated optimization strategy which enables us to avoid settling into local minima and reduces the non-trivial optimization problem to a sequence of successively more "non-convex" optimization problems. The solution is efficiently obtained by an iterative majorization-minimization method combined with a forward-backward splitting strategy.

Sparse representation theory has attracted much attention in recent years, and techniques for finding sparse solutions have found wide use in pattern recognition, computer vision, and image analysis. In particular, in the Face Recognition (FR) context, where the basic goal is to identify a person from his/her face image, given a collection of sample faces.

Challenging FR contexts present test images containing variations that are not present in the training images. These variations, including illumination, pose, facial expressions, occlusions and disguises, are mathematically considered as outliers, and the development of methods which are robust to outliers is still a challenging problem.

The theory of sparse representation and compressed sensing has shown promise in handling this variability in the FR context. In [36], Wright proposed a Sparse-Representation-based Classification (SRC) method which works equally well for any reasonable choice of facial features and it is more robust to occlusions. The basic idea in SRC is that each atom of the dictionary A in (1) represents a training face, and the test face b is approximately represented as a sparse linear combination of all the given sample faces. In [36], an approximate solution is obtained by solving (2), where F is the ℓ_1-norm and H the ℓ_2-norm.

The pioneering SRC idea has been modified by a number of authors (see Section 2.2), mainly by proposing different choices for the penalty and the fidelity terms in (2) as discussed in Section 2.

In this paper we propose to solve the FR problem by the variational model (3) which perfectly fits the usual group-sparsity a priori knowledge that the training samples in the dictionary A are sorted such that samples from the same class are contiguous. The robustness to facial occlusions is enforced by the data fidelity term which allows us to differently weight the outlier/non-outlier pixels to suppress/enhance their influence on the recognition process, according to the Robust Sparse Coding (RSC) approach [38].

In summary, the two significant contributions of our work are the following

- We propose a new general group-sparse representation model which integrates a penalty term F_μ designed to encourage group sparsity, matching the *a priori* knowledge that pre-specified disjoint blocks of variables (training faces) should be selected or ignored simultaneously.

- We devise a suitable efficient optimization method based on a robust graduated optimization strategy for solving the non-convex nonseparable minimization problem (3). The proposed algorithm, called *Group-Sparse Robust Face Recognition* (GSR-FR), is then successfully applied to solve the FR problem.

Extensive experiments on widely used face databases demonstrate that the proposed algorithm GSR-FR is competitive with the state-of-the-art sparse representation based methods. The use of a non-convex nonseparable sparsity-inducing function leads to better recognition rates than comparable methods, especially in the case of low dimensional feature spaces, and in the presence of occlusions.

The paper is organized as follows. Section 2 briefly reviews sparse representation for FR and some related works. Section 3 introduces the novel sparse-based representation model, and in Section 4 we devise a majorization-minimization based numerical method for its optimization. Implementation details on the GSR-FR algorithm are described in Section 5 and a discussion on its convergence is presented in Section 6. The proposed approach is validated in Section 7 and conclusions are drawn in Section 8.

2 Face recognition based on sparse representation

The face recognition process is usually split into two steps. The first is *classification*, that is, to identify the most likely candidate among the samples in the database. The second is *validation*, which verifies whether the test face actually represents the same person.

In a closed-set recognition protocol the system attempts to determine the identity of an unidentified individual, known to be in the database. In a more challenging open-set protocol, the system correctly identify identities that are present in the database while detecting and rejecting as unknown all other probe samples with identities that are not present in the database.

Let us consider a set of sample facial images of K individuals, called *training faces*, which consists of n_κ images of the κth individual, for $1 \leq \kappa \leq K$. Each group of images of the same person is called a *class*. The total number of training faces is $N = \sum_\kappa n_\kappa$. The training images represent well-aligned slightly different poses taken under varying illumination of each subject. Each image is column-major vectorized into an atom of length M.

The given n_κ atoms from the κth class are stacked as columns of a submatrix A_κ of size $M \times n_\kappa$. The *training dictionary matrix* $A \in \mathbb{R}^{M \times N}$ is defined as the concatenation of N training samples of all K classes $A = [A_1, \cdots, A_K]$.

In the sparse representation model (1), each column of the training matrix A is considered as a basis vector, and the test face b is approximately represented as a sparse linear combination of all the given sample faces. If the test face belongs to class κ, we expect that α^* is a sparse but structured vector whose nonzero entries are mostly associated with the κth class.

Given a *test face b*, the task in FR is to classify and validate this test face. Once the sparse coefficient vector α^* is recovered by solving the minimization problem (2), the classification of the test face b is done by determining the class κ for which the ℓ_2-norm of the residual is minimal, i.e., by solving the minimization problem

$$\min_\kappa \|A_\kappa \delta_\kappa(\alpha^*) - b\|_2, \tag{4}$$

where $\delta_\kappa(\alpha^*)$ denotes the coefficients of α^* which correspond to class κ.

The validation is carried out, following [36], by computing the *Sparsity Concentration Index* (SCI) defined as

$$\mathrm{SCI}(\alpha^*) = \frac{K \cdot \max_\kappa \|\delta_\kappa(\alpha^*)\|_1 / \|\alpha^*\|_1 - 1}{K - 1}. \tag{5}$$

If all coefficients in α^* are concentrated in a single class κ, then $SCI(\alpha^*) = 1$. If the coefficients are spread evenly among all the classes, then $SCI(\alpha^*) = 0$. The result is considered to be validated if SCI exceeds a given threshold. In an open-set scenario we can introduce a suitable threshold on the similarity score SCI to reject unknowns.

2.1 Feature extraction

A generally used preliminary procedure to reduce the data dimension and to cut down on the computational effort, is the so-called *feature extraction*. Typically, the original images contain tens of thousands of pixels each, which are reduced to hundreds or even tens of features. Commonly used features are obtained by Eigenfaces, Fisherfaces, Laplacianfaces, downsampling (randomly or by local regions), and other techniques [16]. These traditional features are trivially obtained but demand for robust classifiers. In contrast, the recently introduced Deep Neural Network (DNN) features [32], naturally integrate low/mid/high level features, and rely on sophisticated neural network frameworks, but provide good performance also

for simple classifiers, like Nearest Neighbors Search (NNS). To assess the performance of our proposed classifier, in the experimental session we used Principal Component Analysis (PCA) and downsampling for feature extraction in the closed-set recognition protocol, and DNN feature extraction in the more challenging open-set recognition protocol.

Therefore, from now on, we will consider the dictionary represented by the matrix $A \in \mathbb{R}^{M \times N}$ in the feature space, as well as the test image $b \in \mathbb{R}^M$, so that $M \ll N$ is the dimension of the feature space, which is significantly smaller than the dimension of the vectorized images.

2.2 Related Works

Face recognition is a very active field of research due to its practical applications in many areas such as biometric identification, information security, video surveillance, multimedia retrieval, etc. A large variety of methods have been proposed to solve this problem, such as Principal Component Analysis (PCA) [21], Linear Discriminant Analysis (LDA) [22], Support Vector Machine (SVM) [18], Local Binary Patterns (LBP) [1], Local Derivative Pattern (LDP) [42], and Learning-Based Descriptors [5]. Although these methods perform well when the test image is captured under controlled situations, their performance degrades significantly when the test image contains variations that are not present in the training images. The theory of sparse representation and compressed sensing has shown promise in handling this variability in the FR context.

In the following we briefly review only the sparse representation methods in the form of (2) which have been applied to FR problem, with particular care to those approaches exploiting sparsity or group-sparsity and local consistency.

The pioneering work on *Sparse-Representation-based Classification* (SRC) [36] proposed the minimization of the convex functional (2) with $\ell_1 - \ell_2$ norms for $F(\cdot)$ and $H(\cdot)$, respectively. SRC has achieved attractive performance in robust FR and has motivated a large amount of works that modified one or both of the two terms in the functional (2). Variants to the ℓ_2-norm H function have been proposed to improve robustness of the solution to data noise by giving less weight to outliers. In [8] the authors used the ℓ_1-norm as for the fidelity H which assumes a Laplace distribution for the residual, thus lowering the influence of large errors. Yang *et al.* in [38] introduced the *Robust Sparse Coding (RSC)* where they proposed a different more effective non-convex fidelity term H which has been generalized into the *Iteratively Reweighted Regularized Robust Coding (IR³C)* [41], called RRC_L₁. In [30] a non-convex M-estimator is proposed as fidelity term to enforce the robustness of the sparse representation model which is minimized in the Half Quadratic framework. A further improvement to the robustness to large outliers and non-Gaussian noise has been proposed in [17], by incorporating in the fidelity term a maximum correntropy criterion.

Unlike these sparse representation methods which use the 1-D pixel-based error model to address the face classification problem, the works [29, 37, 40, 46, 47, 19] introduced the local consistency concept (nearby data points share the same properties) by imposing that the error images have low-rank or approximately low-rank structure. As a convex relaxation of the rank-function, in [40] the authors proposed the minimization of the nuclear norm of the error matrix. In [37] a weighted nuclear norm is introduced to make the model more robust to noise.

All the cited works so far addressed the sparsity, pixel-wise or structured-wise, of the error image in the fidelity term H to enforce robustness of the recognition model, considering ℓ_1-norm as the penalty function F.

Many research efforts have been instead devoted to devising variants of the penalty function F to enforce sparsity or group-sparsity more severely on the coefficients α, approaching to the maximally sparsifying ℓ_0-quasinorm.

The *Block Sparse Bayesian Learning method (BSBL)* in [45] is based on this block sparsity notion, and it evaluates the representation coefficients by exploiting the correlation within blocks.

Various authors [10, 33] proposed methods based on group-sparse representation that impose the $\ell_{2,1}$-mixed-norm penalty on the reconstruction coefficients, when the training samples and the test sample are organized into classes.

In [35, 11] and [34] the similarity information between the query sample and distinct classes has been taken into account by weighting in the penalty term the ℓ_1-norm and the $\ell_{2,1}$-norm, respectively.

In [39], a Joint Representation and Pattern Learning (JRPL) model is proposed, which captures structured information and prior knowledge of image features.

Kernel Sparse Representation for Classification (KSRC), proposed in [14], represents a different class of variational sparse representation methods which aim to capture the non-linear distribution of the face images within the data. KSRC and its variants have been proposed for the FR problem in [43, 12].

3 The robust Group-Sparse Representation Model

In this section, we introduce and motivate our choices for the penalty function F_μ and the loss function H in (3).

3.1 Group sparsity-inducing penalty $F_\mu(\alpha)$

Group or block sparsity relies on the *a priori* knowledge that when the sample belongs to a class κ, the vector α^* is not only sparse, but the nonzero elements are mostly associated with the κth class and correlated in amplitude. In this section we devise the following penalty function which encourages group sparsity

$$F_\mu(\alpha) = \sum_{\kappa=1}^{K} \psi_\mu(\|\delta_\kappa(\alpha)\|_1), \qquad (6)$$

where $\delta_\kappa(\alpha)$ refers to the vector of coefficients of α belonging to class κ and ψ_μ is a parameterized non-convex function with parameter $\mu \in \mathbb{R}_+ \setminus \{0\}$ which, as $\mu \to 0$, tends to the ℓ_0-quasinorm. The use of the non-convex ψ_μ encourages the sparsity of the solution, while the argument of ψ_μ in (6) enforces solutions whose nonzero coefficients are located mostly in one of the classes, making the function $F_\mu(\alpha)$ non-separable and non-convex.

Several families of non-convex ψ functions have been proposed, including the ℓ_q-norms for $q < 1$, log-sum, exp, and atan, see [24],[4],[25]. In our model we consider the following parameterized log-exp family of functions $\psi_\mu(t) : \mathbb{R} \to \mathbb{R}_+$, proposed in [28],

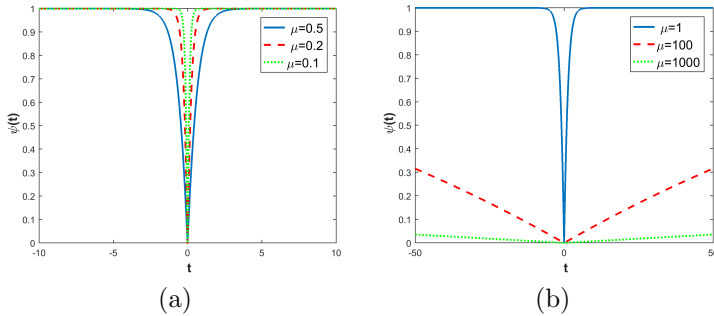$$\psi_\mu(t) = \frac{1}{\log(2)} \log\left(\frac{2}{1 + e^{-|t|/\mu}}\right), \qquad (7)$$

Figure 1: Log-exp sparsity-inducing function $\psi_\mu(t)$ for several values of μ: (a) for small μ values $\psi_\mu(t)$ approximates the ℓ_0-quasinorm; (b) for large μ values $\psi_\mu(t)$ tends to a line.
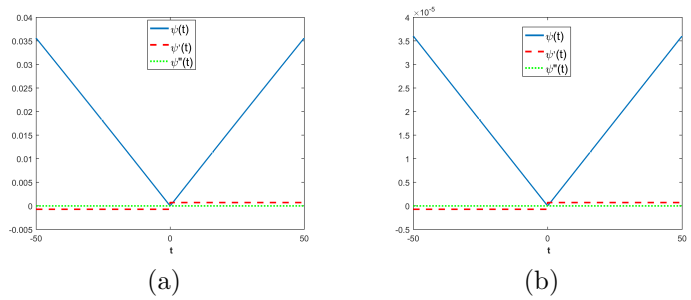

Figure 2: First (dashed line) and second (dotted line) derivatives of $\psi_\mu(t)$ (solid line) for large μ values: (a) $\mu = 1000$; (b) $\mu = 10^6$.

characterized by the properties in the Prop.1.

Proposition 1. *Let $\psi_\mu(t)$ be defined as in* (7). *Then the following properties are satisfied:*

P1)
$$
\begin{cases}
\psi_\mu(t) \in \mathcal{C}^2(\mathbb{R}_+ \setminus \{0\}) \\
(\psi_\mu \text{ twice continuously differentiable in } t \text{ on } \mathbb{R}_+ \setminus \{0\}) \\
\psi_\mu(t) \in \mathcal{C}^0(\mathbb{R}_+) \\
(\psi_\mu \text{ continuous in } t \text{ on } \mathbb{R}_+)
\end{cases}
$$

P2) $\psi'_\mu(t) > 0 \quad \forall t \in \mathbb{R}_+ \setminus \{0\}$
(ψ_μ strictly increasing in t on $\mathbb{R}_+ \setminus \{0\}$)

P3) $\psi''_\mu(t) \leq 0 \quad \forall t \in \mathbb{R} \setminus \{0\}$
(ψ_μ concave in t on $\mathbb{R}_+ \setminus \{0\}$)

P4) $\lim_{\mu \to 0} \psi_\mu(t) = \|t\|_0, \quad \lim_{\mu \to +\infty} \psi_\mu(t) = 0,$

P5) $\lim_{t \to \infty} \psi_\mu(t) = 1$

The properties in Prop. 1 are derived from simple investigations of the first and second derivatives of the penalty function ψ_μ, illustrated in Fig.2 and calculated respectively as

$$
\psi'_\mu(t) = \frac{1}{\mu \log(2)} \frac{1}{1 + e^{|t|/\mu}}, \tag{8}
$$

$$
\psi''_\mu(t) = -\frac{1}{\mu^2 \log(2)} \frac{e^{|t|/\mu}}{\left(1 + e^{|t|/\mu}\right)^2}. \tag{9}
$$

Property P4) (left) states that as $\mu \to 0$ ψ_μ approximates the ℓ_0-quasinorm more closely than the other families mentioned. As $\mu \to \infty$, $\psi''_\mu(t) \to 0$ and thus $\psi_\mu(t)$ tends to a line, which is a convex function.

Plots of the $\psi_\mu(t)$ function for different values of the μ parameter are illustrated in Fig.1. In particular for small values of μ in Fig.1(a) and for large values of μ in Fig.1(b).

The following results motivate the clustering property of the proposed penalty in (6), and the fact that it is a good approximant of the ℓ_0-quasinorm penalty.

Lemma 2. *Let ψ_μ in* (7) *defined for $t \geq 0$, be strictly concave, with $\psi_\mu(0) = 0$. Then for any arbitrary $\mu > 0$ and $\alpha_j > 0$,*

$$
\psi\left(\sum_{j=1}^{s} \alpha_j\right) < \sum_{j=1}^{s} \psi(\alpha_j), \quad j = 1, \dots, s. \tag{10}
$$

Proof. A strictly concave function satisfies $\psi_\mu((1-\lambda)a + \lambda b) > (1 - \lambda)\psi_\mu(a) + \lambda\psi_\mu(b)$, for any $0 < \lambda < 1$, $a < b$. Set $a = 0$, $b = \sum \alpha_j$, $\lambda = \alpha_1 / \sum \alpha_j$, we get

$$
\psi_\mu(\alpha_1) > \frac{\alpha_1}{\sum \alpha_j} \psi_\mu(\sum \alpha_j). \tag{11}
$$

Do the same for all the other α_j, and add. $\square$

Proposition 3. *Let ψ_μ satisfy the assumptions in Lemma 2. Then for any vector $\alpha \in \mathbb{R}^N$ we have*

$$
F_\mu(\alpha) \leq \|\alpha\|_0 \tag{12}
$$

and for a vector $t \in \mathbb{R}^K$, adding up the values of ψ_μ for each component $t_\kappa = \|\delta_\kappa(\alpha)\|_1$ we approximate the ℓ_0-quasinorm for vectors:

$$
\lim_{\mu \to 0} \sum_{\kappa=1}^{K} \psi_\mu(t_\kappa) = \|t\|_0. \tag{13}
$$

Proof.

$$
F_\mu(\alpha) = \sum_\kappa \psi_\mu(\|\delta_\kappa(\alpha)\|_1) = \sum_\kappa \psi_\mu \left(\sum_{j=\kappa n_\kappa + 1}^{(\kappa+1)n_\kappa} \alpha_j \right) =
$$
$$
\leq \sum_\kappa \sum_{j=\kappa n_\kappa + 1}^{(\kappa+1)n_\kappa} \psi_\mu(\alpha_j) \tag{14}
$$
$$
= \sum_i^{N} \psi_\mu(\alpha_i).
$$

In (14) we applied the result in lemma 2. The result (12) follows from the fact that $0 \leq \psi_\mu(t) \leq 1$, while (13) follows from Prop. 1 - P4) left. $\square$

The function F_μ defined in (6) likewise approximately counts the number of groups in which a vector α has non-negligible coefficients. Minimizing F_μ encourages the coefficients to concentrate inside a small number of groups.

3.2 Loss function $H(r)$

Let us assume that the representation residual r_i, defined as $r_i = (A)_i \alpha - b_i$ with $(A)_i$ the ith row vector of A, is independently and identically distributed. Then we define the function $\rho : \mathbb{R} \to \mathbb{R}_+$, introduced in [38], as follows

$$
\rho(r_i; s, d) = -\frac{1}{2s} \left[\log(1 + e^{s(d-r_i^2)} - \log(1 + e^{sd}) \right], \tag{15}
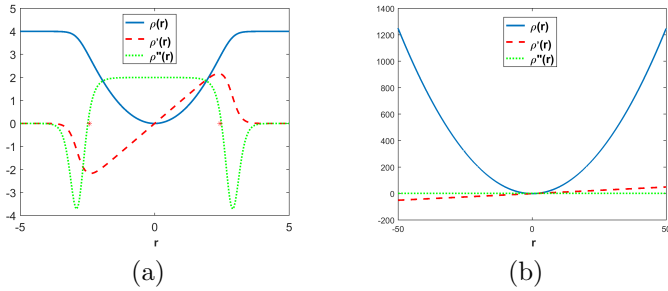$$

4

Figure 3: The loss function $\rho(r)$ (solid line) for $sd = 8$, the first and second derivatives (dashed and dotted lines respectively), the inflection points r^* (stars): (a) for $s = 1$; (b) for $s = 10^{-7}$.

where s, d are positive scalars. In particular, the parameter s controls the sharpness of the transition between convex and non-convex regimes, and the d parameter controls the point of transition. In practice, the loss function ρ assigns a smaller weight to the residual in the convex region, and a larger weight to larger residuals in concave region. The loss function ρ together with its first and second derivatives is illustrated in Fig. 3 for a moderate value s on the left, and for s near 0 on the right.

Proposition 4. *Let $\rho(r; s, \delta)$ be defined as in (15). Then the following properties are satisfied:*

P1) $\begin{array}{l} \rho(r; s, \delta) \in \mathcal{C}^2(\mathbb{R}_+ \setminus \{0\}) \\ (\rho \text{ twice continuously differentiable in } r \text{ on } \mathbb{R} \setminus \{0\}) \end{array}$

P2) $\begin{array}{l} \rho'(r; s, \delta) \geq 0 \quad \forall\, r \in \mathbb{R}_+ \setminus \{0\} \\ (\rho \text{ increasing in } r \text{ on } \mathbb{R}_+ \setminus \{0\}) \end{array}$

P3) $\left\{ \begin{array}{ll} \rho''(r; s, \delta)(t) \geq 0 & \forall\, r \in [-r^*, r^*] \quad (r^* \text{ inflection point}) \\ \rho''(r; s, \delta)(t) < 0 & \forall\, r \in (-\infty, -r^*) \text{ or } (r^*, \infty) \end{array} \right.$

P4) $\lim_{r \to \infty} \rho(r; s, \delta) = \frac{\delta}{2}$

P5) $\lim_{s \to 0} \rho(r; s, \delta) = \frac{r^2}{4}$

The properties in Prop. 4 are derived from simple investigations of the first and second derivatives calculated respectively as

$$\rho'(r; s, \delta) = \frac{r e^{s(\delta - r^2)}}{1 + e^{s(\delta - r^2)}}, \tag{16}$$

$$\rho''(r; s, \delta) = \frac{e^{s(\delta - r^2)}\big(1 - 2r^2 s + e^{s(\delta - r^2)}\big)}{\big(1 + e^{s(\delta - r^2)}\big)^2}. \tag{17}$$

We note that $\rho(r; s, \delta)$ is even and increasing for positive arguments, and $\rho(0; s, \delta) = 0$. Property P5) indicates that for sufficiently small values of s the loss function ρ is convex.

Following [38] the data fidelity term considered in the proposed model (3) is the separable loss function:

$$H(r) = \sum_{i=1}^{M} \rho(r_i). \tag{18}$$

The function ρ resembles a quadratic for r near 0, but flattens in a line for larger arguments. Thus, $H(r)$ approximates the ℓ_2-norm for moderately sized r, but the influence of outlier features is suppressed.

The functional $J(\alpha; \mu, \lambda)$ defined in (3) is lower semi-continuous and bounded from below by zero, since both F_μ and H have a global minimum at zero. However, due to Prop. 1 and Prop. 4, $J(\alpha; \mu, \lambda)$ is not coercive and, in general, non-convex.

4 Numerical solution via graduated optimization

In this section we present an iterative algorithm for computing an approximate solution of the unconstrained non-convex optimization problem (3) with H defined in (18) and F_μ in (6).

Following a graduated optimization-based strategy, at first a convex relaxation of the original non-convex problem (3) is generated by imposing a parameter μ_0 sufficiently large and a value of s in (15) sufficiently small, giving rise to the following functional

$$J(\alpha; \mu_0, \lambda) = F_{\mu_0}(\alpha) + \frac{1}{\lambda} H(A\alpha - b). \tag{19}$$

The function (19) for any $\lambda > 0$ value is convex due to the fact that for $\mu \to \infty$, ψ_μ in F_μ defined in (6), tends to a line, and the function ρ in the loss function H defined in (18) tends to $r^2/4$ as $s \to 0$ - see P5) in Prop. 4 - and consequently both F_μ and H are convex functions.

The iterative method is initialized by calculating a minimizer of $J(\alpha; \mu_0, \lambda_0)$ in (19), with any $\lambda_0 > 0$. Afterward, at each iteration $k \geq 1$, an intermediate solution α^{k+1} is calculated by solving the following minimization problem

$$\alpha^{k+1} = \arg \min_{\alpha \in \mathbb{R}^N} \{J(\alpha; \mu_k, \lambda_k)\}, \tag{20}$$

starting from the previously obtained α^k, and setting s in ρ as a suitable fixed value which induces the non-convexity of H. At each iteration the parameters μ_k and λ_k are monotonically decreased so that $\mu_k < \mu_{k-1}, \lambda_k < \lambda_{k-1}$, thus producing a penalty function F_μ which approaches the ℓ_0-quasinorm.

The problem (20) is in general a challenging non-convex non-separable optimization problem. An approximate solution is obtained at each iteration by applying the Majorization-Minimization (MM) strategy which carries out successive convex minimizations [26].

In the *majorization step*, we generate a tangent majorant of the function (surrogate functional) $J(\alpha; \mu_k, \lambda_k)$ defined as

$$\tilde{J}(\alpha, \alpha^k; \mu_k, \lambda_k) := \tilde{F}_{\mu_k}(\alpha; \alpha^k) + \frac{1}{\lambda_k} \tilde{H}(r; r^k), \tag{21}$$

where $r^k = A\alpha^k - b$, and the functions $\tilde{F}$ and $\tilde{H}$ are computed as specified in Sections 4.1 and 4.2, respectively. Then, in the *minimization step*, the following convex nonsmooth minimization problem is solved

$$\alpha^{k+1} = \arg \min_{\alpha \in \mathbb{R}^N} \left\{ \tilde{J}(\alpha, \alpha^k; \mu_k, \lambda_k) \right\} \tag{22}$$

by applying the forward-backward splitting method [9] accelerated by a strategy proposed in [6] as described in Section 4.3.

The proposed method advances iteratively by gradually reducing the parameters μ and λ, using the solution of the current iteration as initial point for the next iteration, approaching the solution of the original problem (3). The proposed

procedure is then applied to the solution of the Group-Sparse Representation FR (GSR-FR) problem; its algorithmic implementation is provided in Section 5. Numerical details on this iterative graduated optimization procedure will be given in the following section, and some convergence results will be provided in Section 6.

4.1 The majorant $\tilde{F}_\mu$ of F_μ

We propose to majorize F_μ by a convex linear tangent majorant $\tilde{F}_\mu$ near α^k, defined as follows

$$
\begin{aligned}
\tilde{F}_\mu(\alpha; \alpha^k) &= F_\mu(\alpha^k) + \sum_{i=1}^N \frac{\partial F_\mu}{\partial \alpha_i}(\alpha^k)\left(\alpha_i - \alpha_i^k\right) \\
&= \sum_{i=1}^N \frac{\partial F_\mu}{\partial |\alpha_i|}(\alpha^k) \cdot |\alpha_i| + \text{terms independent of } \alpha, \\
&= \|W\alpha\|_1 + \text{terms independent of } \alpha, \quad (23)
\end{aligned}
$$

where α_i is the ith component of $\alpha \in \mathbb{R}^N$, and

$$
W = \text{diag}(w_i), \quad w_i = \frac{\partial F_\mu}{\partial |\alpha_i|}(\alpha^k) = \frac{\partial \psi_\mu}{\partial |\alpha_i|}(\|\delta_\ell(\alpha^k)\|_1), \quad (24)
$$

and ℓ represents the class to which α_i belongs. Therefore, $\tilde{F}_\mu(\cdot; \alpha^k)$ is a tangent majorant of F_μ at α^k, i.e. it satisfies

$$
\tilde{F}_\mu(\alpha; \alpha^k) \geq F_\mu(\alpha) \quad \forall \alpha, \quad \text{and} \quad \tilde{F}_\mu(\alpha^k; \alpha^k) = F_\mu(\alpha^k). \quad (25)
$$

4.2 The majorant $\tilde{H}$ of H

For the loss function H defined in (18) we propose to employ a quadratic tangent majorant, since the function itself behaves like a quadratic function for small arguments.

As the function H is separable, each coordinate $\rho(r_i)$ is majorized separately by adding a quadratic term to the first order Taylor approximant of $\rho(r_i)$ near r_i^k, thus obtaining

$$
\tilde{\rho}(r_i; r_i^k) = \rho(r_i^k) + \rho'(r_i^k)(r_i - r_i^k) + \frac{1}{2}v_i(r_i - r_i^k)^2. \quad (26)
$$

The constant v_i is chosen so that $\tilde{\rho}(r_i; r_i^k)$ has its minimum at $r_i = 0$, where also $\rho(0) = 0$. By differentiating (26) we have

$$
\tilde{\rho}'(r_i; r_i^k) = \rho'(r_i^k) + v_i(r_i - r_i^k). \quad (27)
$$

By imposing the condition $\tilde{\rho}'(0; r_i^k) = 0$ we obtain $v_i = \frac{\rho'(r_i^k)}{r_i^k}$, which in turns leads to

$$
\tilde{\rho}(r_i; r_i^k) = \frac{1}{2}v_i r_i^2 + \text{terms independent of } r_i. \quad (28)
$$

Replacing component-wise the majorants in (26) in the definition (18), we obtain the following majorant $\tilde{H}$ for H

$$
\begin{aligned}
\tilde{H}(r; r^k) &= \sum_{i=1}^M \tilde{\rho}(r_i; r_i^k) \\
&= \frac{1}{2}\|V^{1/2} r\|_2^2 + \text{terms independent of } r, \quad (29)
\end{aligned}
$$

where the diagonal matrix V is defined as

$$
V = \text{diag}(v_i), \quad v_i = \frac{\rho'(r_i^k)}{r_i^k}. \quad (30)
$$

For all r^k, the function $\tilde{H}(\cdot; r^k)$ is a quadratic tangent majorant of H at r^k, i.e. it satisfies

$$
\tilde{H}(r; r^k) \geq H(r), \quad \forall r \quad \text{and} \quad \tilde{H}(r^k; r^k) = H(r^k). \quad (31)
$$

4.3 Forward-Backward (FB) inner Iteration

The coefficients α^{k+1} are computed at each iteration by solving the minimization problem in (21)-(22) using the forward-backward strategy [9] combined with the variant proposed in [6] of the *Fast Iterative Soft Thresholding Algorithm* (FISTA) [2] to achieve a convergent acceleration. By replacing in (21) the majorants $\tilde{F}_\mu$, $\tilde{H}$ given in (23) and (29), and omitting the terms independent of α, the convex optimization problem (22) reads as

$$
\begin{aligned}
\tilde{J}(\alpha, \alpha^k; \mu_k, \lambda_k) &= \|W^k \alpha\|_1 + \frac{1}{2\lambda_k}\|(V^k)^{1/2}(A\alpha - b)\|_2^2 \\
&= \|W^k \alpha\|_1 + \frac{1}{2\lambda_k}\|D^k \alpha - y^k\|_2^2, \quad (32)
\end{aligned}
$$

where $D^k = (V^k)^{1/2} A$ and $y^k = (V^k)^{1/2} b$.

Problem (32) is solvable and it can be cast in the form of minimizing $f + g$, where f is a continuous convex function but nonsmooth, due to the ℓ_1-norm, and g is a differentiable convex function with Lipschitz constant $L = \|(D^k)^T(D^k)\|_2$. The FB splitting strategy then reduces to a soft thresholding with an explicitly given closed-form solution

$$
\tilde{\alpha}^{k+1} = S_{\lambda_k \beta W^k}(\alpha^k + \beta(D^k)^T(y^k - D^k \alpha^k)).
$$

The parameter $\beta > 0$ represents the step size, and $S_t(\nu)$ is a point-wise soft-thresholding function which, for given vectors t and ν, applies soft thresholding with parameter t_i to the element ν_i of ν, namely

$$
[S_t(\nu)]_i = \text{sign}(\nu_i) \max(0, |\nu_i| - t_i), \quad \forall i.
$$

The FB algorithm with acceleration approximates α^{k+1} by iterating over j, assuming an initial α_0^k and $\tilde{\alpha}_0^{k+1}$, for all $j \geq 1$

$$
\begin{aligned}
\tilde{\alpha}_j^{k+1} &= S_{\lambda_k \beta W^k}(\alpha_{j-1}^k + \beta(D^k)^T(y^k - D^k \alpha_{j-1}^k)) \quad (33) \\
\alpha_j^{k+1} &= \tilde{\alpha}_j^{k+1} + \tau_j(\tilde{\alpha}_j^{k+1} - \tilde{\alpha}_{j-1}^{k+1}), \quad (34)
\end{aligned}
$$

where the weights τ_j in (34) used for convergence acceleration are computed as in [6], namely

$$
\tau_j = \frac{t_j - 1}{t_{j+1}}, \qquad t_j = \frac{j + a - 1}{a} \qquad a > 2 \quad \forall j. \quad (35)
$$

We stop the FB iterations (33)-(34) as soon as the relative error sequences

$$
err_j := \left| \frac{\tilde{J}(\alpha_j^{k+1}, \alpha^k; \mu_k, \lambda_k) - \tilde{J}(\alpha_{j-1}^{k+1}, \alpha^k; \mu_k, \lambda_k)}{\tilde{J}(\alpha_j^{k+1}, \alpha^k; \mu_k, \lambda_k)} \right| \quad (36)
$$

drop below a prescribed threshold.

5 Algorithm GSR-FR

To summarize previous results, in Algorithm GSR-FR we report the main computational steps of the overall proposed MM iterative approach.

The GSR-FR Algorithm is initialized by setting α^0 to be the minimizer of (19), and stopped as soon as the relative change

$$
\frac{\|\alpha^{k+1} - \alpha^k\|_2}{\|\alpha^{k+1}\|_2}, \quad (37)
$$

drops below a prescribed threshold.

5.1 Choice of Parameters

The algorithm GSR-FR requires the choice of some parameters that depend on the FR problem. This subsection describes how they have been chosen.

The values of the parameters λ and μ have to be initialized and then reduced in the iterative graduated optimization process. For efficiency we would like to start with relatively small values, and reduce them rapidly. The choices described here have been found to work well in the face recognition setting.

The λ parameter is initialized as $\lambda_0 = c_{\lambda_{in}} \|A^T V b\|_\infty$, where the value of $c_{\lambda_{in}}$ ranges in the interval $[5 \cdot 10^{-8}, 10^{-5}]$, depending on the number M of features used. The lower the value of M, the larger the factor $c_{\lambda_{in}}$ required. The updating of λ requires $c_\lambda = \frac{F_{\mu_k}(\alpha^k)}{F_{\mu_{k+1}}(\alpha^k)} 0.99$ in order to satisfy condition (39) in Prop. 6.

The parameter μ, which controls the penalty function (6), is reduced at each step k by a factor $c_\mu = 0.3$.

Algorithm 1 GSR-FR

---

inputs: $A \in \mathbb{R}^{M \times N}$, normalized training faces; $b \in \mathbb{R}^M$, test face

outputs: κ class of b; SCI concentration index

parameters: λ_0, μ_0

Solve optimization problem (3):
    Compute α^0 by minimizing (19)
    OUTER LOOP
    repeat
      Majorization step:
        generate surrogate $\tilde{J}(\alpha, \alpha^k; \lambda_k, \mu_k)$ in (32)
          using (24) and (30)
      Minimization step:
      INNER LOOP (FB)
      $\alpha_0^{k+1} = \alpha^k$
      repeat
        Compute $\tilde{\alpha}_{j+1}^{k+1}$ by solving (33)
        Update α_{j+1}^{k+1} by (34)
      until $err_{j+1} \leq \gamma \cdot \lambda_k$
      EXIT INNER LOOP
      $\alpha^{k+1} = \alpha_{j+1}^{k+1}$ (result of inner loop)
      Parameter reduction:
      $\lambda_{k+1} = c_\lambda \cdot \lambda_k$
      $\mu_{k+1} = c_\mu \cdot \mu_k$
    until (37) satisfied
    $\alpha^* = \alpha^{k+1}$

Classify and Validate:
    Compute the representative class κ by (4)
    Compute the $SCI(\alpha^*)$ for the class κ by (5)

The inner loop FB approximately solves the optimization problem (32). For larger λ_k, it is not necessary to solve the inner loop to high accuracy, since the result α^k simply serves as a starting value for the next step. As λ_k gets close to 0, we require more accuracy. Nevertheless, the number of iterations required in the inner loop remains small thanks to the warm starting strategy applied. In the stopping criteria of the inner loop, the parameter γ ranges in the interval $[5 \cdot 10^{-8}, 10^{-5}]$, depending on the number M of features used. The lower the value of M, the smaller γ value is required.

The loss function (18) depends on the two parameters d and s. We use the choice advocated in [38]. We set d to a certain percentile of the square of the residual (80th percentile without occlusion, 60th percentile with occlusion), and s is directly derived by $s = 8/d$.

5.2 Complexity analysis

Assuming that the dimensionality M of face features is fixed, and N is the number of sample images, the complexity of the GSR-FR Algorithm basically depends on the update of the weighting matrices V and W, defined in (30) and (24) respectively, and on the solution of the FB inner loop by solving (33)-(34).

In the outer loop, the weighting matrix updates cost $O(M)$ for V and $O(N)$ for W. Since the FB inner loop relies on a soft thresholding of a vector, the most computationally expensive part is the update of this vector, which implies the evaluation of the matrix-matrix-vector product $(D^k)^T(D^k)\alpha^k$. This is reduced to applying A, $V_k^{1/2}$ and A^T in sequence to α^k, thus the cost is $2MN + M$.

Summarizing, the overall complexity is $O(k_1 k_2 MN)$ where k_1 and k_2 are the number of outer and inner iterations respectively. Experimentally the outer loop usually requires a few iterations k_1, while the number of iterations k_2 of the inner loop can vary from less than ten to several hundred, depending on both the number of features considered and on the percentage of outliers in the dictionary.

Finally, we remark that in order to guarantee the convergence of the FB inner loop, the assumptions in Prop.5, that will be discussed in Section 6, require the evaluation of the step-size β at each outer loop, which in turn involves the estimation of the dominant eigenvalue of the matrix $(D^k)^T(D^k)$. In our experience, it is sufficient to fix β as a suitable reduction of an initial estimation of $\|(D^1)^T(D^1)\|_2$.

6 Convergence results

The convergence of Majorization-Minimization-based algorithms for solving non-convex regularized optimization problems is a challenging issue. In [7] a convergence result is provided under the assumption of coercivity, which is not guaranteed in our model. In this section we analyze the convergence of the MM approach for the face recognition problem, whose main steps are given in Algorithm GSR-FR.

In order to guarantee that the proposed iterative approach does not break down, we should require that at each iteration k the solution α^{k+1} of the problem (32) exists and is unique. For this aim, we recall in Prop. 5 the result proved in [6], on the convergence of the sequence $\{\alpha_j\}_{j \in \mathbb{N}}$ computed by (33)-(34) to a minimum of problem (32) using (35) for the acceleration strategy, and then we prove in Prop. 6 the descent property for $J(\alpha^k; \mu^k, \lambda^k)$.

Proposition 5. *The accelerated FB scheme (33)-(34) converges to a minimizer of the convex functional $\tilde{J}(\alpha, \alpha^k; \mu_k, \lambda_k)$ in (32), if the sequence τ_j is computed as in (35) and the step size satisfies $0 < \beta < 1/L$, where $L = \lambda_{max}((D^k)^T(D^k))$.*

Proposition 6. *Assume that the conditions of Prop.5 hold. Let μ_k, λ_k and μ_{k+1}, λ_{k+1} be values of the penalization parameters at two successive iterative steps such that*

$$\lambda_{k+1} < \lambda_k \quad and \quad \mu_{k+1} < \mu_k,$$

and α^k, α^{k+1} be the corresponding solutions of the optimization problems (32). Then the objective function $J(\alpha; \mu_k, \lambda_k)$ in (3) satisfies

$$J(\alpha^{k+1}; \mu_{k+1}, \lambda_{k+1}) < J(\alpha^k; \mu_k, \lambda_k), \qquad (38)$$

provided that

$$\lambda_{k+1} < \frac{F_{\mu_k}(\alpha^k)}{F_{\mu_{k+1}}(\alpha^k)} \lambda_k. \qquad (39)$$

Proof.

$$
\begin{aligned}
J(\alpha^k; \mu_k, \lambda_k) &= F_{\mu_k}(\alpha^k) + \frac{1}{\lambda_k} H(r^k) \\
&> F_{\mu_{k+1}}(\alpha^k) + \frac{1}{\lambda_{k+1}} H(r^k) \quad \text{by (39)} \\
&= \tilde{F}_{\mu_{k+1}}(\alpha^k; \alpha^k) + \frac{1}{\lambda_{k+1}} \tilde{H}(r^k; r^k) \quad \text{by (25),(31)} \\
&= \tilde{J}(\alpha^k, \alpha^k; \mu_{k+1}, \lambda_{k+1}) \quad \text{by (21)} \\
&\geq \tilde{J}(\alpha^{k+1}, \alpha^k; \mu_{k+1}, \lambda_{k+1}) \quad \text{by (32)} \\
&= \tilde{F}_{\mu_{k+1}}(\alpha^{k+1}; \alpha^k) + \frac{1}{\lambda_{k+1}} \tilde{H}(r^{k+1}; r^k) \\
&\geq F_{\mu_{k+1}}(\alpha^{k+1}) + \frac{1}{\lambda_{k+1}} H(r^{k+1}) \quad \text{by (25),(31)} \\
&= J(\alpha^{k+1}; \mu_{k+1}, \lambda_{k+1}). \qquad (40)
\end{aligned}
$$

$\square$

Condition (39) is naturally satisfied by observing the behavior of ψ_μ for decreasing values of μ, as shown in Fig.1.

Definition 7. *A convex (not necessarily differentiable) function $f(\alpha)$ is said to be σ-strongly convex if and only if there exists a constant $\sigma > 0$, called the* modulus of strong convexity *of $f(\alpha)$, such that the function $f(\alpha) - \frac{\sigma}{2}\|\alpha\|_2^2$ is convex.*

The following important result on strongly convex functions holds, [31].

Lemma 8. *Let $f(\alpha) : \mathbb{R}^n \to \mathbb{R}$ be σ-strongly convex and α^* be a minimizer of function $f(\alpha)$. Then the following inequality holds:*

$$\frac{\sigma}{2}\|\alpha - \alpha^*\|_2^2 \leq f(\alpha) - f(\alpha^*) \quad \forall \alpha \in \mathbb{R}^n. \qquad (41)$$

Proposition 9. *The surrogate functional $\tilde{J}(\alpha, \alpha^k, \mu_k, \lambda_k)$ in (32) is σ-strongly convex with $\sigma = \frac{1}{\lambda^k}$.*

The proof of Prop. 9 follows straightforward by observing that the sum of a convex term and a σ-strongly convex term is σ-strongly convex (see [31]). The first term in $\tilde{J}(\alpha, \alpha^k, \mu_k, \lambda_k)$ is convex, due to the ℓ_1-norm, and the second term is σ-strongly convex with $\sigma = \frac{1}{\lambda^k}$.

Proposition 10. *Let $\{\alpha^k\}_{k=1}^{\infty}$ denote the sequence of iterates generated by the proposed procedure applied to the solution of the non-convex non-smooth optimization problems of the form (3). For any initial guess α^0 the following statements hold:*

s1) the sequence $\{J(\alpha^k; \mu_k, \lambda_k)\}_{k=0}^{\infty}$ is monotonically non-increasing and convergent;

s2) the sequence $\{\alpha^k\}_{k=0}^{\infty}$ is of finite length, in the sense that: $\sum_{k=0}^{+\infty} \|\alpha^{k+1} - \alpha^k\|_2^2 < +\infty$, which implies $\lim_{k \to \infty} \|\alpha^{k+1} - \alpha^k\|_2^2 = 0$.

Figure 4: Experiment 1: training images from the AR database (first row) and EY database (second row).

Proof. Proof of s1) follows from Prop. 6 which guarantees the monotonicity of $J(\alpha^k; \mu_k, \lambda_k)$, and by noting that $\{J(\alpha^k; \mu_k, \lambda_k)\}_{k=0}^{\infty}$ is continuous and bounded from below by zero, hence convergent.

Since $\tilde{J}(\alpha; \alpha^k, \mu_k, \lambda_k)$ is σ-strongly convex (see Prop.9), we apply inequality (41) in Lemma 8 where $f(\alpha)$ is replaced by $\tilde{J}(\alpha, \alpha^k, \mu_k, \lambda_k)$ and α^* by α^{k+1}, $\forall \alpha \in \mathbb{R}^n$, $\forall k \geq 0$:

$$
\begin{aligned}
\frac{\sigma}{2}\|\alpha - \alpha^{k+1}\|_2^2 &\leq \tilde{J}(\alpha; \alpha^k, \mu_k, \lambda_k) - \tilde{J}(\alpha^{k+1}; \alpha^k, \mu_k, \lambda_k) \\
&\leq J(\alpha^k; \mu_k, \lambda_k) - J(\alpha^{k+1}; \mu_k, \lambda_k),
\end{aligned}
$$
$$(42)$$

where (42) comes from $\tilde{J}(\alpha; \alpha^k, \mu_k, \lambda_k) = J(\alpha^k; \mu_k, \lambda_k)$ and $\tilde{J}(\alpha^{k+1}; \alpha^k, \mu_k, \lambda_k) \geq J(\alpha^{k+1}; \mu_k, \lambda_k)$. Summing the inequalities (42) over k yields:

$$\sum_{k=0}^{+\infty} \|\alpha^{k+1} - \alpha^k\|_2^2 \leq \frac{2}{\sigma} \sum_{k=0}^{+\infty} [J(\alpha^k; \mu_k, \lambda_k) - J(\alpha^{k+1}; \mu_k, \lambda_k)] = \qquad (43)$$

$$= \frac{2}{\sigma}(J(\alpha^0; \mu_k, \lambda_k) - J^*). \qquad (44)$$

where J^* denotes the finite limit of the convergent sequence $\{J(\alpha^k; \mu_k, \lambda_k)\}_{k=0}^{\infty}$. Since $0 < \sigma < +\infty$ and the sequence $\{J(\alpha^k; \mu_k, \lambda_k)\}_{k=0}^{\infty}$ is monotonically non-increasing, then the right-hand of (44) is a finite non-negative number and the series on the left-hand of (43) is convergent, thus proving s2). $\square$

An analysis of the convergence behaviour of the sequence $\{\alpha^k\}$ generated by (20) to the critical point of $J(\alpha; \mu, \lambda)$ is beyond the scope of this paper.

7 Numerical Experiments

In this section, we investigate the performance of the proposed GSR-FR Algorithm, by presenting experiments on three widely used benchmark face databases: AR face database [27], Extended Yale B [23], and PubFig [48].

The AR database consists of over 4,000 frontal images for 126 individuals (70 men and 56 women) of size $768 \times 576 = 442,368$. For each individual, 26 pictures were taken in two separate sessions (2 different days). These images include facial variations, illumination changes, expressions, and facial disguises, such as sunglasses and scarves.

The Extended Yale B database (EY) consists of 2,414 cropped frontal face images of 38 individuals. For each subject, there are approximately 64 face images of size $192 \times 168 = 32,256$, which are captured under various laboratory-controlled lighting conditions.

The PubFig is a large, real-world face dataset consisting of 58,797 images of 200 people in web photos and, unlike most other existing face datasets, these images are taken in completely uncontrolled situations with non-cooperative subjects [48].

We compare our method with recently proposed sparse-representation-based classical methods, including the original SRC [36], the RRC_L_1 method [41, 44], the half quadratic with multiplicative form HQ_M proposed in [30], and two of the state-of-the-art methods for face recognition, named FLR-IRC, proposed in [19, 20], and the RMR method in [37]. We solve the ℓ_1-minimization problem for the SRC method by using the $l_1\_ls()$ Matlab routine available from Mathworks database. For all the other methods we used the solvers provided by the authors of the corresponding papers.

All the experimental results were obtained by running a Matlab implementation of the proposed algorithm on a PC with Intel Core i7 3.40 GHZ and 16 GB of RAM under Windows 10.

In the remaining section we investigate the following FR examples in a closed-set protocol:

- cases with non-contiguous variations, such as variations in illumination or in facial expressions (Experiment 1).

- cases with contiguous variations, such as random block-image occlusions of different sizes and facial disguises, such as scarves and sunglasses (Experiment 2).

Finally, we test the algorithm under a FR open-set protocol (Experiment 3).

7.1 Experiment 1: Illumination and Expression Changes

The aim of this experiment is validate the performance of the proposed GSR-FR Algorithm in FR problems with variations such as illumination and expression changes. In Figure 4 some training images from database AR (first row) and EY (second row) are shown, respectively.

Setup for AR Database
We chose a subset of 50 males and 50 females with only illumination and expression changes. For each subject we selected the first seven images from Session 1 for training, which leads to a dictionary $A \in \mathbb{R}^{M \times N}$ with $N = (50 + 50) \times 7$ atoms. For testing, we used the associated seven images from Session 2, which corresponds to different vectors b in (1). Each image is resized and cropped to 60×43 pixels. For feature extraction by downsampling, we further resized the images to sizes 7×5, 10×7, 15×10, 20×14 and 30×21. For PCA we used feature spaces of dimensions 30, 54, 120 and 300.

Setup for EY Database
We randomly selected half of the images for training (about 32 images per subject), and the other half for testing, thus composing a dictionary $A \in \mathbb{R}^{M \times N}$ with $N = 32 \times 38$ atoms. Each image is cropped and resized to size 54×48. For feature extraction by downsampling we did a further reduction to sizes 6×6, 9×8, 13×12, and 27×24. For PCA we used feature spaces of dimensions 30, 84, 150, 304.

The recognition rates of this experiment have been reported in Table 1 for feature extraction by downsampling and in Table 2 for feature extraction by PCA. In both Tables the second column represents the number M of features considered for the dictionary $A \in \mathbb{R}^{M \times N}$ and the test $b \in \mathbb{R}^M$.
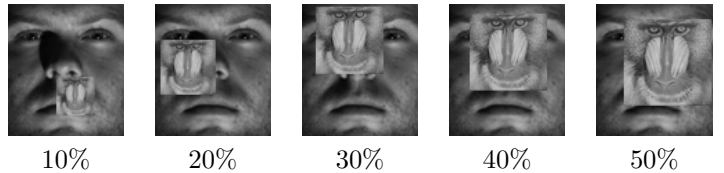

| 10% | 20% | 30% | 40% | 50% |

Figure 5: Experiment 2: Block-occluded test image in EY database, with varying occlusions from 10% to 50%.
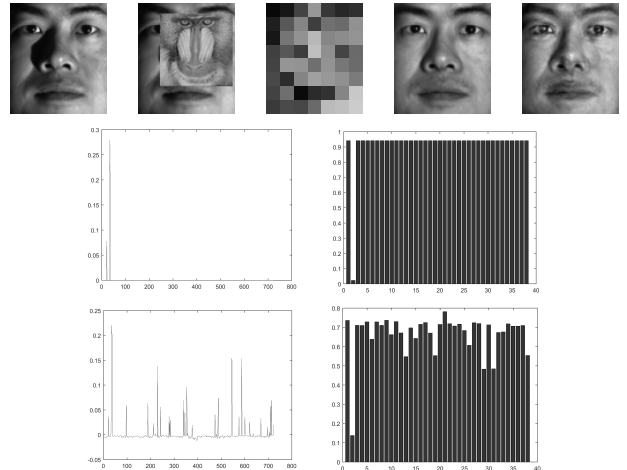

Figure 6: Experiment 2: recognition of subject 14 from EY database under 50% block image occlusion.

The results in the Tables for the RRC_L_1 method applied to images in the Extended Yale B database, are slightly different from those given in [41] because of the random selection of the training and testing faces.

For both the considered databases, we observe that the proposed GSR-FR Algorithm performs better than all the other considered methods, especially in case of highly reduced data, which correspond to lower M values. However, the improvement in performance affects the computational timings, as illustrated in Table 3 for the tests shown in Table 1.

7.2 Experiment 2: Occlusions and Disguises

In this subsection we test the robustness of GSR-FR Algorithm to different kinds of contiguous occlusions, such as block occlusion and real disguises.

7.2.1 Face recognition under block-image corruption

For training we used the faces from subset 1 and 2 of the EY database, characterized by normal-to-moderate lighting conditions, and, for testing, the faces from subset 3, with more extreme lighting conditions. On each test image b of dimension 96×48, a randomly located square block has been replaced with an unrelated image (the baboon image in SRC [36]), producing varying occlusions from 10% to 50% of the original testing image (see Fig.5 for an example of occluded test images). For the feature extraction we then downsampled all the images to sizes 6×5, 8×7, 12×10 and 24×21.

Figures 6 and 7 show two examples of face recognition under 50% block-image occlusions. In the first row of Figures 6 and 7 we report from left to right: the uncorrupted original test images, the test images with an occlusion of 50%, the

Table 1: Experiment 1: FR rates under changes in illumination and pose, feature extraction by downsampling

DB	M	**GSR-FR**	RRC$\_$L$_1$	SRC	FLR-IRC	HQ_M	RMR
AR	7×5	**63.8%**	54.7%	55.8 %	47.9 %	49.8 %	44.6 %
	10×7	**81.3%**	78.1%	78.2 %	81.3 %	71.9 %	69.8 %
	15×10	**91.4%**	88.4%	87.3 %	91.3 %	88.0 %	85.0 %
	20×14	**95.0%**	92.0%	91.3 %	94.5 %	94.0%	85.7 %
	30×21	**96.7%**	95.7%	92.4 %	96.4 %	96.6 %	95.8 %
EY	6×6	**81.5%**	78.3%	81.0 %	26.2 %	60.0 %	55.2 %
	9×8	**92.8%**	90.0%	90.8 %	80.4 %	83.6 %	76.5 %
	13×12	**97.2%**	95.6%	94.8%	96.7 %	92.0 %	90.2 %
	27×24	**99.1%**	98.7%	98.2%	98.8 %	96.8 %	96.8 %

Table 2: Experiment 1: FR rates under changes in illumination and pose, feature extraction by PCA

DB	M	**GSR-FR**	RRC$\_$L$_1$	SRC
AR	30	**75.3%**	70.8%	73.5%
	54	**89.1%**	87.6%	83.3%
	120	**95.3%**	95.1%	90.1%
	300	**97.3%**	96.3%	93.3%
EY	30	**93.6%**	88.5%	90.5 %
	84	**97.8%**	97.6%	95.6 %
	150	**99.0%**	98.4%	97.2 %
	300	**99.2%**	98.9%	98.5%

Figure 8: Experiment 2: training images for class 1 in AR database (first row); test images in AR database with sunglasses and scarves (second row).
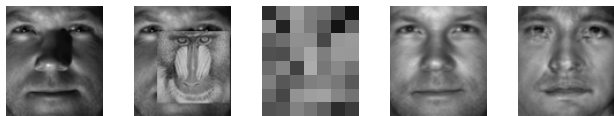

Figure 7: Experiment 2: recognition of subject 35 from EY database under 50% block image occlusion.

downsampled 8×7 test images which represent the input of both reconstruction algorithms, and the reconstructed images obtained by applying the GSR-FR and FLR-IRC algorithms.

In the second row of Figures 6 we plot the coefficient vectors (α^*) associated with the reconstructions illustrated in the first row of Figures 6, together with the residuals per class obtained by solving the minimization problem in (4) for both the considered methods.

The reconstruction obtained by the GSR-FR Algorithm in Fig. 6 is of significantly better quality, this is justified by a visually inspection of the plots in the second row, which show that the solution α^* obtained by the GSR-FR Algorithm is very sparse, with significant values in the correct class; and theoretically motivated by the group-sparsity induced in the solution in the non-convex functional minimized by the proposed GSR-FR Algorithm. In contrast, the coefficients of the FLR-IRC algorithm present significant values corresponding to the exact class, but also some spurious coefficients in other classes which lead to a corrupted reconstruction (see Fig. 6, right image).

In the case illustrated in Fig. 7 the FLR-IRC algorithm fails to recognize the test image in the training database and it produces a reconstructed image which combines several faces belonging to different classes into the resulting image face (see Fig. 7, last image in the first row).

Table 4 shows the FR rates of the several considered algo-

rithms on different downsampling dimensions (labeled as M in the Table) in terms of varying percentages of occlusion ranging from 10% to 50%. The GSR-FR Algorithm outperforms the compared algorithms for highly reduced data and high occlusion percentages. For each test the best recognition rate results are marked in boldface.

The FLR-IRC algorithm relies on the nuclear norm of the residual image, therefore the test setting used in Experiment 1, which presents non-contiguous occlusions, is not optimal for this method since the residual image is not modelled as a low-rank matrix. This partially motivates the better performance of the proposed GSR-FR with respect to the FLR-IRC method. On the other hand, the test setting of Experiment 2 contains contiguous block occlusions which represent the optimal context for FLR-IRC method. However, the better recognition rates of the FLR-IRC method are maintained only if the features considered are represented either by the entire image or by a downsampling of large dimensions of it (not less than 24×21 pixels). Moreover, unlike our proposal, the FLR-IRC method is suitable only for features extracted from the values of the original image, and thus the use of PCA or similar features for FLR-IRC would not produce good performance.

Table 5 shows the performance in terms of execution times of GSR-FR, RRC$\_$L$_1$, HQ_M, RMR and FLR-IRC methods. Unfortunately, the main drawback of the GSR-FR Algorithm is the computational timing, since it is slower than the compared methods at least in its current naive implementation.

7.2.2 Face recognition under real face disguise

We considered a subset of the AR database to test the performance of our method on face recognition under disguises. We chose a subset of 799 images of various facial expressions without occlusion for training, which leads to a dictionary

Table 3: Experiment 1: Time (secs) under changes in illumination and pose, feature extraction by downsampling

DB	M	**GSR-FR**	FLR-IRC	RRC_L_1	SRC	HQ_M	RMR
AR	7×5	0.32	0.01	0.07	0.74	0.14	0.023
	10×7	0.39	0.06	0.08	0.80	0.17	0.024
	15×10	0.47	0.08	0.11	0.82	0.19	0.027
	20×14	0.60	0.12	0.24	0.90	0.27	0.27
	30×21	0.80	0.14	0.44	0.90	0.36	0.47
EY	6×6	0.66	0.21	0.31	1.75	0.08	0.07
	9×8	0.87	0.29	0.42	2.15	0.55	0.08
	13×12	0.93	0.31	0.68	2.18	1.05	0.09
	27×24	1.13	0.44	1.33	2.24	1.873	1.22

Table 4: Experiment 2: FR rates with image-block occlusion

ALG	M	Percentage occluded			
		10%	30%	40%	50%
SRC	6×5	36.3%	22.6%	17.4%	14.3%
	8×7	51.0%	31.7%	20.7%	17.4%
	12×10	75.4%	48.6%	30.3%	20.9%
	24×21	96.5%	72.3%	54.3%	35.2%
RRC_L_1	6×5	50.9%	24.4%	19.6%	13.8%
	8×7	67.3%	32.4%	19.6%	16.4%
	12×10	93.3%	50.7%	31.8%	26.4%
	24×21	99.8%	94.7%	80.4%	61.3%
HQ_M	6×5	**58.2%**	35.8%	25.1%	17.9%
	8×7	81.8%	60.9%	43.2%	31.5%
	12×10	95.6%	80.9%	64.9%	43.2%
	24×21	98.2%	93.1%	87.1%	66.7%
FLR-IRC	6×5	29.56%	20.0%	12.22%	9.11%
	8×7	**85.6%**	48.4%	33.62%	19.3%
	12×10	**99.1%**	**86.0%**	62.9%	36.2%
	24×21	100%	**100%**	99.8%	92.4%
GSR-FR	6×5	54.7%	**40.7%**	**31.8%**	**23.6%**
	8×7	75.1%	**61.8%**	**46.9%**	**32.4%**
	12×10	95.3%	82.9%	**73.8%**	**52.7%**
	24×21	**100.0%**	94.7%	89.3%	81.1%
RMR	6×5	45.8%	24.9%	22.9%	13.3%
	8×7	67.3%	38.9%	26.4%	22.0%
	12×10	89.6%	56.9%	51.3%	35.3%
	24×21	97.6%	76.2%	59.6%	46.0%

Table 5: Experiment 2: time in seconds under block occlusion

M	Percentage Occluded 10-50 %				
	RRC_L_1	FLR-IRC	GSR-FR	HQ_M	RMR
6×5	1.3	**0.15**	0.60	0.1	0.024
8×7	1.5	**0.20**	0.63	0.25	0.025
12×10	2.4	**0.21**	0.64	0.33	0.027
24×21	4.6	**0.26**	0.67	0.52	0.38

Figure 9: Experiment 3: Facial images from PubFig83+LFW database. A few distractors are shown in the second row.

$A \in \mathbb{R}^{M \times N}$ with $N = 799$. Fig.8(first row) shows an example of eight selected training images of the first subject (class 1) in the AR database.

Then we formed two separate testing sets of 200 images. The images in the first set were from the neutral expression with sunglasses (the 8th image in each session) which cover roughly 20% of the face (see Fig.8(second row)), while the images in the second set were from the neutral expression with scarves (the 11th image in each session) which cover roughly 40% of the face (see Fig.8(second row)). All the images were resized to 9×6, 13×10, 27×20 and 42×30, respectively.

The recognition rates in case of real face disguises are shown in Table 6. For both the considered disguises, sunglasses and scarves, the proposed GSR-FR algorithm outperforms the other methods. In particular, in the case of sunglasses GSR-FR achieved a recognition rate between 0.5% and 9.2% higher than RRC_L_1, between 1% and 29% higher than FLR-IRC, between 4% and 8% higher than HQ_M, and between 3.3% and 30.5% higher than RMR. In the case of scarves, GSR-FR Algorithm achieved a 7.5% to 29% higher recognition rate than RRC_L_1, between 18% and 30.5% higher than FLR-IRC, between 37.5% and 42.5% higher than HQ_M, and between 23.5% and 25% higher than RMR.

7.3 Experiment 3: Open-Set Face Recognition

In this experiment we test the effectiveness of the proposed GSR-FR algorithm in an open-set protocol. For this aim we introduced in the algorithm a decision threshold $\theta \in [0, 1]$ on the Sparsity Concentration Index defined in (5). We analyzed the performance on the following two test cases:

1. **EY+EY**: the training set includes only a half of the subjects in the database EY, namely 608 face images of 19 subjects, while the test set contains all the same subjects of the training set (590 faces), but under different illumina-

11

Table 6: Experiment 2: face recognition rates with disguise

M	Sunglasses					
	SRC	RRC_L$_1$	FLR-IRC	**GSR-FR**	HQ_M	RMR
9×6	46.5%	61.3%	41.0%	**70.5%**	70.8 %	41.0 %
13×10	72.0%	92.0%	76.4%	**93.5%**	85.5%	56.5 %
27×20	83.0%	96.0%	98.0%	**99.5%**	94.0%	75.0 %
42×30	89.0%	99.0%	98.5%	**99.5%**	95.5%	96.2 %

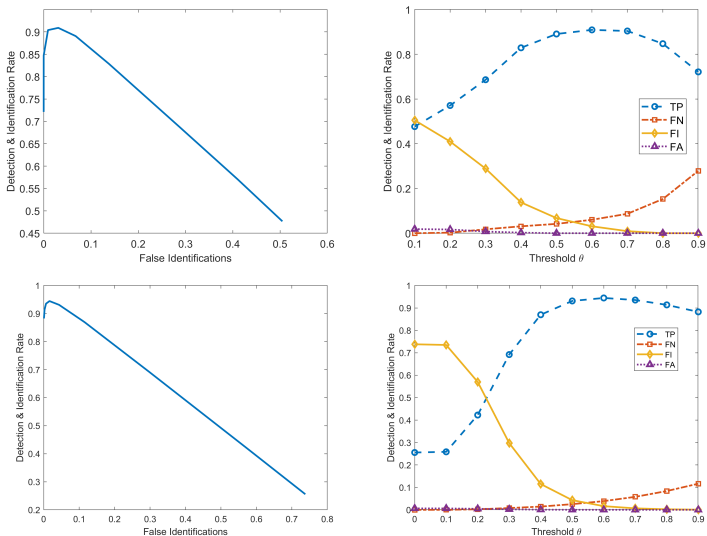| M	Scarves					
	SRC	RRC_L$_1$	FLR-IRC	**GSR-FR**	HQ_M	RMR
9×6	10.0%	32.0%	30.5 %	**61.0%**	18.5 %	36.0 %
13×10	16.0%	69.0%	58.0%	**84.5%**	35.5 %	57.0 %
27×20	21.5%	81.5%	79.0%	**94.0%**	41.5 %	66.0 %
42×30	37.0%	89.5%	79.0%	**97.0%**	60.5 %	73.5 %

Figure 10: Experiment 3: Face Recognition Evaluation for the dataset **EY+EY** (first row), and **PubFig83+LFW** (second row): (left) Detection & Identification Rate (DIR) curve; (right) Detection rate with respect the decision thresholding θ.

tions and poses, plus 608 extra face images of the other 19 subjects in the EY database which are not in the training set (the unknown identities or distractor set). We considered feature extraction by downsampling reducing the images to size 27×24.

2. **PubFig83+LFW**: this enriched dataset proposed in [3] is the combination of PubFig [48] and the LFW [50] datasets to form a new benchmark dataset for open face identification. PubFig83+LFW divides the 13,002 faces of 83 individuals from PubFig database into 2/3 training set (8720 faces) and 1/3 testing set (4282 faces) and sets 12,066 faces of 5503 individuals from LFW as distractor set. All images are aligned by eye position and cropped to the dimension 256×256. Figure 9 illustrates some facial images from PubFig. We considered feature extraction by DNN ResNet, introduced in [32], which reduces each face image to a vector of dimension 128.

The GSR-FR algorithm has been evaluated by using the Detection and Identification Rate (DIR) curve [15] which plots the identification rates with respect to the false identifications, where an unknown subject is recognized as an identity of the gallery (training dataset), see Fig.10(first row, left) for case 1, and Fig.10(second row, left) for case 2.

The algorithm achieved its good identification accuracies up to 92% (case 1) and 94% (case 2), at low false identification rates.

For a more detailed analysis on the optimal thresholding θ value, we also report in Fig.10(right) the identification rate with respect to increasing θ values. In particular, we plot the identification rate (True Positive - TP -), the false rejected rate of known identities (False Negative - FN -), the false identification rate (- FI -) and the false acceptance rate of known identities (False Acceptance - FA -).

For $\theta = 0$ our algorithm is forced in any case to classify the test subject as one of the subject in the dictionary, that is why for $\theta = 0$ the TP+FN corresponds to the percentage of subjects in the gallery, while the FI decreases for increasing θ values up to an optimal θ value which is between 0.6 and 0.7 which leads to very good performance in terms of maximum identification rate and minimum false identification rate for both the test cases. Both FI and FA rates contribute to false acceptance in an open-set identification system. However, in a real scenario, e.g. in surveillance systems, is much more dangerous to accept an unknown identity rather than to misdetect a known identity in the gallery. The proposed algorithm turns out to be robust to both these false detections.

8 Conclusion

We proposed, analyzed and tested a new variational model for solving the sparse representation problem which, relying on properly designed non-convex penalty and loss functions, can take full advantage of sparsity and clustering intrinsic properties of the face recognition problem. In sparsity-based face recognition, the functional to be minimized consists typically of two functions to choose – a regularizer on the coefficients $F(\cdot)$ and an error penalty $H(\cdot)$. Recently, most of the effort has been addressed to induce sparsity in the error term $H(\cdot)$ which makes the model robust to noise and occlusions without the need for augmenting the training matrix. For this aim, many variants of the separable nonconvex function $H(\cdot)$ introduced in [38] have been proposed, which, however, relax the penalty $F(\cdot)$ to a convex ℓ_1-norm term. Based on the assumption that in the training matrix A the samples are sorted such that samples from the same class are contiguous, we proposed to built a group-sparsity penalty $F_\mu(\cdot)$ as a superposition of penalties

for group of contiguous coefficients. This leads to a non-convex structured sparsity-inducing regularizer and a numerical challenging non-convex non-smooth optimization problem. The novelty in this work is not only the choice of $F(\cdot)$, but also the robust graduated optimization scheme based on majorization-minimization strategy combined with FB accelerated splitting methods, which in particular reduces the solution to a sequence of convex optimization sub-problems. The experimental results show an improvement in terms of recognition rates especially when using feature extraction with a low number of features, and in the presence of occlusions. The results of this paper can be extended to a number of models involving sparse representations where the training data are grouped in classes and in each class they live in a low-dimensional subspace of a high-dimensional ambient space, thus reducing significantly the problem dimension.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen, *Face Description with Local Binary Patterns: Application to Face Recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2006), 28 , pp. 2037–2041.

[2] A. Beck and M. Teboulle, *Fast Gradient-based Algorithms for Constrained Total Variation Image Denoising and Deblurring Problems*, IEEE Transactions on Image Processing, (2009), 18 , pp. 2419–2434.

[3] B.C. Becker, E.G. Ortiz, Evaluating open-universe face identification on the web, IEEE Conf Comput Vis Pattern Recogn Workshops, (2013), pp. 904–911.

[4] E.J. Candès, M.B. Wakin, and S.P. Boyd, *Enhancing Sparsity by Reweighted l_1 minimization*, J. Fourier Anal. Appl. (2008), 14, pp. 877–905.

[5] Z. Cao, Q. Yin, X. Tang, and J. Sun, *Face Recognition with Learning-Based Descriptor*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2010), pp. 2707–2714.

[6] A. Chambolle, and C. Dossal, *On the Convergence of the Iterates of the Fast Iterative Shrinkage/Thresholding Algorithm*, Journal of Optimization Theory and Applications, (2015), 166(3), pp. 968–982.

[7] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbo, *A Majorize-Minimize Subspace Approach for $\ell_2-\ell_0$ Image Regularization*, SIAM J. Imaging Sciences, (2013), 6(1), pp. 563–591.

[8] J. Chorowski and J. Zurada, *Obtaining Full Regularization Paths for Robust Sparse Coding with Applications to Face Recognition*, 11th International Conference on Machine Learning and Applications, (2012), 1, 2012, pp. 356–361.

[9] P.L. Combettes and V.R. Wajs, *Signal Recovery by Proximal Forward-Backward Splitting*, Multiscale Model. Simul., (2005), 4, pp.1168–1200.

[10] E. Elhamifar and R. Vidal, *Robust Classification using Structured Sparse Representation*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2011), pp. 1873–1879.

[11] Z. Fan, M. Ni, Q. Zgu, E. Liu, *Weighted Sparse Representation for Face Recognition*, Neurocomputing, (2015), 151, pp. 304–309.

[12] Z. Fan, D. Zhang, X. Wang, Q. Zhu, Y. Wang, *Virtual Dictionary Based Kernel Sparse Representation for Face Recognition*, Pattern Recognition, (2018), 76, pp. 1–13.

[13] G. Gasso, A. Rakotomamonjy, and S. Canu, *Recovering Sparse Signals with a Certain Family of Nonconvex Penalties and DC Programming*, IEEE Trans. on Signal Processing, (2009), 57, pp. 4686–4698.

[14] S. Gao, I.W.-H. Tsang, L.-T. Chia, *Sparse Representation with Kernels*, IEEE Trans. on Image Processing, (2013), 22(2), pp. 423–434.

[15] P. J. Phillips, P. Grother, and R. Micheals. Handbook of Face Recognition, chapter Evaluation Methods in Face Recognition. Springer, 2nd edition, 2011.

[16] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, *Face Recognition using Laplacianfaces*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2005), 27(3), pp. 328–340.

[17] R. He, W.S. Zheng, and B.G. Hu, *Maximum Correntropy Criterion for Robust Face Recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence,(2011), 33, pp. 1561–1576.

[18] B. Heisele, P. Ho, and T. Poggio, *Face Recognition with Support Vector Machines: Global versus Component-based Approach*, Proc. 8th International Conference on Computer Vision, (2001), pp. 688–694.

[19] M. Iliadis, H. Wang, R. Molina and A. K. Katsaggelos *Robust and Low-Rank Representation for Fast Face Identification With Occlusions*, IEEE Transactions on Image Processing, (2017), 26(5), pp. 2203–2218.

[20] M. Iliadis, H. Wang, R. Molina and A. K. Katsaggelos, `https://github.com/miliadis/FIRC`, 2017.

[21] X. Jiang, *Asymmetric Principal Component and discriminant Analyses for Pattern Classification*,IEEE Transactions on Pattern Analysis and Machine Intelligence,(2009), 31, pp. 931–937.

[22] X. Jiang, *Linear Subspace Learning-based Dimensionality Reduction*, IEEE Signal Processing Magazine, (2011), 28, pp. 16–26.

[23] K.C. Lee, J. Ho, and D. Kriegman, *Acquiring Linear Subspaces for Face Recognition under Variable Lighting*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2005), 27, pp. 684–698.

[24] A. Lanza, S. Morigi, F. Sgallari, *Constrained TVp L2 Model for Image Restoration*, Journal of Scientific Computing, (2016), 68(1), pp. 64–91.

[25] A. Lanza, S. Morigi, and F. Sgallari, *Convex Image Denoising via Non-convex Regularization with Parameter Selection*, Journal of Mathematical Imaging and Vision, (2016), 56(2), pp. 195–220.

[26] A. Lanza, S. Morigi, I. Selesnick, and F. Sgallari, *Nonconvex nonsmooth optimization via convex–nonconvex majorization–minimization*, Numerische Mathematik, (2017), 136(2), pp. 343–381.

[27] A. Martinez and R. Benavente, *The AR Face Database*, Tech. Rep., Centre de Visió per Computador, Universitat Autónoma de Barcelona, 1998.

[28] L.B. Montefusco, D. Lazzaro, and S. Papi, *A Fast Algorithm for Nonconvex Approaches to Sparse Recovery Problems*, Signal Processing, (2013), 93, pp. 2636–2647.

[29] Qian J, Luo L, Yang J, et al. *Robust Nuclear Norm Regularized Regression for Face Recognition with Occlusion*, Pattern Recognition, (2015), 48, pp. 3145–3159.

[30] R. He, W. Zheng, T. Tan and Z. Sun, *Half-Quadratic-Based Iterative Minimization for Robust Sparse Representation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2014), 36(2) , pp. 261–275.

[31] Shai Shalev-Shwartz, *Online Learning and Online Convex Optimization*, Foundations and Trends in Machine Learning, (2011), 4(2), pp. 107–194.

[32] K.He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, (2016) IEEE Conference on Computer Vision and Pattern Recognition.

[33] X. Tang and G. Feng, *Weighted Group Sparse Representation Based on Robust Regression for Face Recognition*, Biometric Recognition, 7th Chinese Conference, CCBR 2012, Guangzhou, China, December 4-5, 2012. Proceedings, Lecture Notes in Computer Science, Springer, Berlin, (2012) 7701, pp. 42–49.

[34] X. Tang, G. Feng, J. Cai, *Weighted Group Sparse Representation for Undersampled Face Recognition*, Neurocomputing (2014), 145, pp. 402–415.

[35] Z. Lu, B. Xu, N. Liu, and Q. Liao, *Face Recognition via Weighted Sparse Representation using Metric Learning*, Proceedings of IEEE International Conference on Multimedia and Expo (ICME) (2017), pp. 391–396.

[36] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, *Robust Face Recognition via Sparse Representation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2009), 31 , pp. 210–227.

[37] J. Xie, J. Yang, J. Qian, Y. Tai, and H. M. Zhang, *Robust Nuclear Norm-Based Matrix Regression with Applications to Robust Face Recognition*, in IEEE Transactions on Image Processing, (2017), 26(5), pp. 2286–2295.

[38] M. Yang, D. Zhang, J. Yang, and D. Zhang, *Robust Sparse Coding for Face Recognition*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2011), pp. 625–632.

[39] M. Yang, P. Zhu, F. Liu, and L. Shen, *Joint Representation and Pattern Learning for Robust Face Recognition*, Neurocomputing, (2015), 168, pp. 70–80.

[40] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, *Nuclear Norm Based Matrix Regression with Applications to Face Recognition with Occlusion and Illumination Changes*, in IEEE Trans on Pattern Analysis and Machine Intelligence, (2017), 39 (1), pp. 156–171.

[41] M. Yang, L. Zhang, J. Yang, and D. Zhang, *Regularized Robust Coding for Face Recognition*, IEEE Transactions on Image Processing, (2013), 22, pp. 1753–1766.

[42] B. Zhang, Y. Gao, S. Zhao, and J. Liu, *Local Derivative Pattern Versus Local Binary Pattern: Face Recognition with High-Order Local Pattern Descriptor*, IEEE Transactions on Image Processing, (2010), 19, pp. 533–544.

[43] L. Zhang, W.-D. Zhou, P.-C. Chang, J. Liu, Z. Yan, T. Wang, F.-Z. Li, *Kernel Sparse Representation-based Classifier*, in IEEE Transactions on Signal Processing, (2012), 60, pp. 1684–1695

[44] L. Zhang, Matlab code for RRC_L_1, available at `http://www4.comp.polyu.edu.hk/~cslzhang` (2013).

[45] Z. Zhang and B. Rao, *Extension of SBL Algorithms for the Recovery of Block Sparse Signals with intra-block Correlation*, IEEE Transactions on Signal Processing, (2013), 61, pp. 2009–2015.

[46] H. Zhang, J. Yang, J. Qian, and W.Luo, *Nonconvex Relaxation Based Matrix Regression for Face Recognition with Structural Noise and Mixed Noise*, Neurocomputing, (2017), 269, pp. 188–198.

[47] C.-H. Zheng, Y.F. Hou, and J. Zhang, *Improved Sparse Representation with Low-Rank Repesentation for Robust Face Recognition* Neurocomputing, (2016), 198, pp. 114–124.

[48] N. Kumar, A. C. Berg, P. N. Belhumeur, and S.K. Nayar, Attribute and Simile Classifiers for Face Verification, International Conference on Computer Vision (ICCV), 2009.

[49] N. Pinto, Z. Stone, T. Zickler, and D. D. Cox, Scaling Up Biologically-Inspired Computer Vision: A Case Study in Unconstrained Face Recognition on Facebook, Proc. Workshop on Biologically Consistent Vision (in conjunction with CVPR), 2011.

[50] E. Learned-Miller, G.B. Huang, A. RoyChowdhury, H. Li, and G. Hua, Labeled Faces in the Wild: A Survey, In Advances in Face Detection and Facial Image Analysis, edited by Michal Kawulok, M. Emre Celebi, and Bogdan Smolka, Springer, pp. 189–248, 2016.