Pattern Recognition 63 (2017) 182–192

Contents lists available at ScienceDirect

Pattern Recognition

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

Joint representation classification for collective face recognition☆

Liping Wang[a,*], Songcan Chen[b]

[a] Department of Mathematics, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[b] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

ARTICLE INFO

Keywords:
SRC
JRC
IQM
Practical IQM

ABSTRACT

In recent years, many representation based classifications have been proposed and widely used in face recognition. However, these methods code and classify testing images separately even for image-set of the same subject. This scheme utilizes only an individual representation rather than the collective one to classify such a set of images, doing so obviously ignores the correlation among the given set of images. In this paper, a joint representation classification (JRC) for collective face recognition is presented. JRC takes the correlation of multiple images as well as a single representation into account. Even for an image-set mixed with different subjects, JRC codes all the testing images over the base images simultaneously to facilitate recognition. To this end, the testing images are aligned into a matrix and the joint representation coding is formulated as a generalized $l_{2,q} - l_{2,p}$ matrix minimization problem. A unified algorithm, named by iterative quadratic method (IQM), and its practical implementation are developed specially to solve the induced optimization problem for any $q \in [1, 2]$ and $p \in (0, 2]$. Experimental results on three public databases show that the JRC with practical IQM not only saves much computational cost but also achieves better performance in collective face recognition than state-of-the-art methods.

1. Introduction

Recently, representation coding based classification and its variants have been developed for facial image recognition (FR) [1–5]. These schemes have achieved a great success in FR and have boosted the applications of image classification [6,7]. The main idea can be carried out by two steps: 1) coding a testing sample as a linear combination of all the training samples, then 2) classifying the testing sample to the subject with the most compact representation evaluated by coding errors. The equations frequently used for representation coding can be uniformly accommodated to the following framework

$$\min_x \| y - Ax \|_q^q + \lambda \| x \|_p^p, \quad 1 \le q \le 2, 0 < p \le 2, \tag{1}$$

where $A \in R^{m \times d}$ is the dictionary of coding atoms and $y \in R^m$ is a testing sample. $x \in R^d$ is the representation vector. With the solution x^* to (1), y is identified as follows

$$\text{identity}(y) = \arg \min_{1 \le i \le I} \{ \| y - A\hat{x}_i^* \|_2 \}, \tag{2}$$

where I denotes the number of classes. $\hat{x}_i^*$ is the recovered coefficient vector associated with class i which is extracted from x^* by keeping the ith class coefficients while the other entries are set to zero [1].

Different pairs of $q \in [1, 2]$ and $p \in (0, 2]$ result in different representation coding formulas. Sparse representation based classification (SRC) [1] is the most known one which uses the l_1-regularized least square problem ($q = 2, p = 1$ in (1)) to sparsely code the query image. The experimental results [1] exhibit the amazing recognition performance of SRC. But the authors of [2] argued that SRC over emphasized the importance of l_1-norm sparsity but ignored the effect of collaborative representation. Consequently, a collaborative representation based classification with l_2-regularized least square (CRC-RLS) was presented which is the special case of (1) with $q = p = 2$. Anyway, CRC-RLS's coding problem is easier to solve for its smoothness than that of SRC. Moreover, Wright et al. [3] ever used $l_1 -$ norm to measure the coding fidelity of y over A, which is another special case of (1) with $q = p = 1$. Compared with the works mentioned above, the representation and regularization measurements of (1) are extended to $\|\cdot\|_q (1 \le q \le 2)$ and $\|\cdot\|_p (0 < p \le 1)$ respectively. This modification provides possibility to adaptively choose the best formula for different applications. Moreover, the computational experiences [8–10] have showed that fractional norm $l_p (0 < p < 1)$ exhibits sparser pattern than l_1-norm. Then the unified generalization formula (1) is expected to achieve better performance.

On the other hand, Eq. (1) uses a coding vector to represent the

☆ The work is partially supported by the Chinese grants NSFC11471159, 61661136001, 11611130018 and Natural Science Foundation of Jiangsu Province (BK20141409).
* Corresponding author.
E-mail addresses: wlpmath@nuaa.edu.cn (L. Wang), s.chen@nuaa.edu.cn (S. Chen).

http://dx.doi.org/10.1016/j.patcog.2016.10.004
Received 19 May 2015; Received in revised form 19 June 2016; Accepted 5 October 2016
Available online 08 October 2016
0031-3203/

testing samples one by one. Given a collection of query images $y_1, y_2, \ldots, y_n \in R^m$, Eq. (1) codes each y_j independently by all the training samples A as

$$y_j \approx Ax_j, \quad 1 \le j \le n. \tag{3}$$

Then y_j is assigned by (2) based on its most compact coding coefficient x_j^*. Obviously the recognition of y_j depends on the single representation coding x_j^* individually but takes no account of the correlation with other testing samples (y_l, $l \neq j$). Even though different frontal faces take on different appearances, the pixel intensity values taken from facial images have direct correlation [11]. Similar images are located together while dissimilar images are spaced far apart which plays an important role in recognizing facial images [12]. In many applications, a great number of images for each known subject have been collected from video sequence or photo album. Face recognition has to be conducted with a set of probe images rather than a single one [13–15]. Collective face recognition or image-set based face recognition seems important and necessary.

Compared with regular face recognition, image-set based face recognition is much less studied. Few image-set based face recognition methods [13–15] ever explored the set-to-set classification after the testing images have been pre-separated according to different classes. To the best of our knowledge, collective face recognition for image-set mixed with different subjects has never been straightly concerned due to the challenging complexity. In this paper, we propose a joint representation coding based classification (JRC) for collective face recognition. To make sufficient use of correlation among the given set of images, we consider to jointly represent all the testing samples simultaneously over the training sample base. Here we employ matrix instead of vector as the coding variable to evaluate the distribution of feature space. The joint representation scheme is eventually formulated as a $l_{2,q} - l_{2,p}$ matrix minimization which covers the vector framework (1). To solve the matrix optimization problem with generalized measurements, a unified algorithm and its practical implementation are proposed and the convergence behavior is accordingly analyzed. Experiments on three public face datasets validate the improvement of JRC over the state-of-the-arts.

In short, the main contributions of this paper lie in:

(1) A joint representation coding based classification (JRC) is presented which implements collective images representation coding simultaneously. This approach is more economical and efficient in computational cost and CPU time. Moreover, JRC can handle collective face recognition but the testing images are not necessarily pre-separated according to classes which is different from the set-to-set approaches employed in [13–15].
(2) Joint coding technique takes account of the correlation hidden in the multiple testing face images. The generalized measurements $q \in [1, 2]$ and $p \in (0, 2]$ in representation coding Eq. (12) provide possible adaption to different applications. For example when $0 < p \le 1$, all the testing images are jointly represented by the training samples in sparse pattern. The recovered largest row coefficients are distinguished according to different subjects but jointly clustered with respect to images of the same subject.
(3) To solve the joint representation $l_{2,q} - l_{2,p}$ matrix optimization problem (12), a uniform algorithm is developed for any $q \in [1, 2]$ and $p \in (0, 2)$. The algorithm makes objective function strictly decrease until it converges to the optimal solution. To the best of our knowledge, it is an innovative approach to solve such a generalized $l_{2,q} - l_{2,p}$ matrix minimization problem.

This paper is organized as follows. In the second section, a joint representation based classification (JRC) will be established. The third section is dedicated to an iterative quadratic (IQM) algorithm for solving the joint matrix optimization problem induced by JRC. Some computational details are considered in the fourth section and a practical implementation is developed. The experimental results are reported in the fifth section while the convergence analysis on IQM is presented in Appendix B.

2. Joint representation classification for collective face recognition

2.1. Joint representation formulation

Suppose that we have I classes of subjects in the facial image dataset. $A_i \in R^{m\times d_i} (1 \le i \le I)$ denotes the i-th class, and each column of A_i is a sample of class i. Hence all the training samples are aligned by $A = [A_1, A_2, \ldots, A_I] \in R^{m\times d}$, where $d = \sum_{i=1}^{I} d_i$. Denote $Y = [y_1, y_2, \ldots, y_n] \in R^{m\times n}$ all the query images, we propose to jointly represent the image set simultaneously by

$$Y \approx AX, \tag{4}$$

where $X = [x_1, x_2, \ldots, x_n] \in R^{d\times n}$ stands for the joint coding matrix. As far as the columns are concerned, system (4) is an easy consequence of (3). To measure the fidelity of the joint coding system (4), we consider X in another sense. Let $A^i \in R^d$ and $Y^i \in R^n$ be the ith ($i = 1, 2, \ldots, m$) row vectors of matrix A and Y respectively, formula (4) is equivalent to

$$X^T (A^i)^T \approx (Y^i)^T \quad \text{for } i = 1, 2, \ldots, m. \tag{5}$$

It is noticed that A and Y array the sampled images column by column, hence their rows span the training and testing feature spaces respectively. In feature extraction view, the collective coding matrix X also plays approximation projecting role from the training feature space to the testing feature space. Traditional least square regression aims to minimize the error

$$\min_X \sum_{i=1}^{m} \| X^T (A^i)^T - (Y^i)^T \|_2^2 \quad \text{or} \min_X \sum_{i=1}^{m} \| A^i X - Y^i \|_2^2. \tag{6}$$

(6) can be easily reformulated as

$$\min_X \sum_{i=1}^{m} \| (AX - Y)^i \|_2^2 \tag{7}$$

where $(AX - Y)^i$ is the ith row vector of $AX - Y$. Actually we prefer a uniform generalization of (7) in the sense

$$\sum_{i=1}^{m} \| (AX - Y)^i \|_2^q \quad (1 \le q \le 2). \tag{8}$$

Under the assumption that joint representation and feature distribution share the similar pattern for all testing facial images, we use the following regularization

$$\sum_{i=1}^{d} \| X^i \|_2^p \quad (0 < p \le 2), \tag{9}$$

where X^i is the ith row vector of X for $i = 1, 2, \ldots, d$. Combining (8) and (9), we present the joint representation formulation as follows

$$\min_X \sum_{i=1}^{m} \| (AX - Y)^i \|_2^q + \lambda \sum_{i=1}^{d} \| X^i \|_2^p, \quad (1 \le q \le 2, 0 < p \le 2). \tag{10}$$

When the number of testing samples in Y is 1, collective representation Eq. (10) is reduced to the single coding Eq. (1). Compared with coding vector x, joint coding matrix X expands each coefficient entry to a row vector which naturally reflects the integral structure of dataset. The row vector norm $\| X^i \|_2$ gives the joint coefficient of all the testing images over the ith training samples. Then the joint row coefficient vector $\{\| X^i \|_2\}_{i=1}^{d}$ somewhat measures the correlation of different classes. To illustrate the joint pattern of JRC, we randomly choose 3 images of two classes in Georgia-Tech database for testing (see Fig. 1). In the left images, 600 (12 images each of 50 classes)

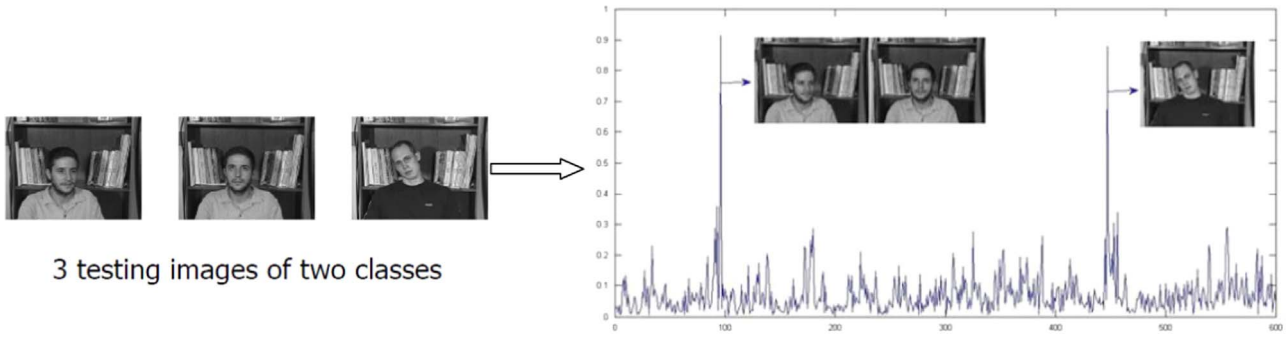


Fig. 1. Three images of two classes and the joint row coefficients by (10) (q=2,p=1).

samples are randomly chosen as training set. Denote X^* an optimal solution to matrix minimization (10) with respect to $q = 2, p = 1$. Under 1/8 downsampling ratio, the joint row coefficients $(\| (X^*)^1 \|_2, \| (X^*)^2 \|_2, \ldots, \| (X^*)^{600} \|_2)^T$ are recovered and plotted in Fig. 1 (more details about Georgia-Tech database and practical implementation can be seen in Section 5).

Fig. 1 distinctly shows the two largest row coefficients corresponding to 2 testing classes (subject 8 and 38 respectively). But for two images of the same class (subject8), the representation coefficients are jointly clustered. We also randomly test 11, 48, 77 and 96 images of 9, 33, 44 and 48 classes respectively, each subject contains $1 \sim 3$ pictures. The recovered row coefficients have the similar behavior as showed in Fig. 1 and the recognition accuracy are 100%. Joint sparse representation Eq. (10) with respect to q=2 and p=1 not only distinguishes different classes but also clusters the same subject which partly explains the wonderful performance of JRC for collective face recognition.

To simplify the formulation, we introduce the mixed matrix norm $l_{2,p}(p > 0)$ (taking $\| X \|_{2,p}$ for example)

$$\| X \|_{2,p} = \Big(\sum_{i=1}^{d} \| X^i \|_2^p \Big)^{\frac{1}{p}}, \quad X \in R^{d \times n}. \tag{11}$$

Then (10) is rewritten as

$$\min_X \| AX - Y \|_{2,q}^q + \lambda \| X \|_{2,p}^p, \quad (1 \le q \le 2, 0 < p \le 2). \tag{12}$$

Especially when $p \in (0, 1)$, $\|\cdot\|_{2,p}$ is not a valid matrix norm because it does not satisfy the triangular inequality of matrix norm axioms. Meanwhile the involved fractional matrix norm based minimization (12) is neither convex nor Lipschitz continuous which brings computational challenge. Designing an efficient algorithm solving such $l_{2,q} - l_{2,p}$ matrix minimization problem is very important. It is also the most challenging task of this paper.

2.2. Joint representation based classification

For fixed parameters q and p, suppose that X^* is a solution to optimization problem (12), that is

$$X^* = \arg \min_X \|AX - Y\|_{2,q}^q + \lambda \|X\|_{2,p}^p. \tag{13}$$

If X^* is partitioned to I blocks as follows

$$X^* = \begin{bmatrix} X_1^* \\ \vdots \\ X_i^* \\ \vdots \\ X_I^* \end{bmatrix} \tag{14}$$

where $X_i^* \in R^{d_i \times n} (1 \le i \le I)$. Let $\widehat{X}_i^*$ denote the coding matrix associated with class i, that is

$$\widehat{X}_i^* = \begin{bmatrix} 0 \\ \vdots \\ X_i^* \\ \vdots \\ 0 \end{bmatrix}, \tag{15}$$

then $A\widehat{X}_i^* = A_i X_i^* (1 \le i \le I)$. In this paper, we concern collective face recognition that the testing images have not been pre-separated according to classes. Hence for each testing image $y_j (j = 1, 2, \ldots, n)$, we have to independently classify y_j to the class with the most compact representation. By evaluating the error corresponding to each class

$$\| (Y - A\widehat{X}_i^*)_j \|_2 \quad i = 1, 2, \ldots, I \tag{16}$$

we pick out the index outputting the least error. The joint representation based classification for collective face recognition can be concluded as follows.

Algorithm 2.1. (JRC scheme for FR)

1. Start: Given $A \in R^{m \times d}$, $Y \in R^{m \times n}$ and select parameters $\lambda > 0$, $q \in [1, 2]$ and $p \in (0, 2]$.
2. Solve $l_{2,q} - l_{2,p}$-minimization problem(12) for coding matrix X^*.
3. For $j = 1: n$
   For i=1: I
   $e_i(y_j) = \| (Y - A_i X_i^*)_j \|_2$
   end
   Identity $(y_j) = \text{argmin}_{1 \le i \le I} \{e_i(y_j)\}$
   end

It is worth to point out that the JRC scheme can be easily extended for the presence of pixel distortion, occlusion or high noise in testing images. Modify (4) as

$$Y = AX + E \tag{17}$$

where $E \in R^{m \times n}$ is an error matrix. The nonzero entries of E locate the corruption or occlusion in Y. Substitute $\widehat{A} = [A, I] \in R^{m \times (d+m)}$ and $\widehat{X} = \begin{bmatrix} X \\ E \end{bmatrix} \in R^{(d+m) \times n}$ for A and X respectively, a stable joint representation coding can be formulated to

$$\min_{\widehat{X}} \|\widehat{A}\widehat{X} - Y\|_{2,q}^q + \lambda \|\widehat{X}\|_{2,p}^p, \quad (1 \le q \le 2, 0 < p \le 2). \tag{18}$$

Once a solution $\widehat{X}^* = \begin{bmatrix} X^* \\ E^* \end{bmatrix}$ to (18) is computed, setting $Y^* = Y - E^*$ recovers a clean image from corrupted subject. To identity the testing sample y_j, we slightly modify the error of y_j with each subject $e_i(y_j) = \| (Y - E^* - A_i X_i^*)_j \|_2$. Thus a robust JRC is an easy consequence of Algorithm 2.1. The corresponding algorithm and theoretical analysis can be similarly established. This paper will not concentrate on this subject.

We notice that the row joint coefficients $\| X^i \|_2 (1 \le i \le d)$ in JRC are nonnegative which implies that the nonnegative constraint on representation coding matrix X is necessary. For a set of images, especially for the images belonging to the same subject, the high

similarity and correlation might result in the dependence of column coefficient vectors in coding matrix X. Then low-rank constraint on collective representation matrix for correlation among given testing image set is another interesting topic on JRC.

3. An iterative quadratic method for JRC

Obviously, efficiently solving optimization problem (12) plays the most important role in scheme 2.1. The representation equations of SRC [1], CRC-RLS [2] and l_1-fidelity [3] are special cases of (12), the algorithms used in [1–3] can not be directly applied here. Such generally mixed matrix minimizations as (12) have been widely used in machine learning. Rakotomamonjy and his co-authors [16] proposed to use the mixed matrix norm $l_{q,p}(1 \le q < 2, 0 < p \le 1)$ in multi-kernel and multi-task learning. But the induced optimization problems in [16] have to be solved separately by different algorithms with respect to p=1 and $0 < p < 1$. For grouped feature selection, Suvrit [17] addressed a fast projection technique onto $l_{1,p}$-norm balls particularly for $p = 2, \infty$. But the derived method in [17] does not match problem (12). Similar joint sparse representation has been used for robust multimodal biometrics recognition in [18]. The authors of [18] employed the traditional alternating direction method of multipliers to solve the involved optimization problem. Nie et al. [19] applied $l_{2,0+}$-norm to semi-supervised robust dictionary learning, while the optimization algorithm has not displayed definite convergence analysis.

In this section, a unified method will be developed to solve the $l_{2,q} - l_{2,p}$ matrix minimization problem for any $1 \le q \le 2$ and $0 < p \le 2$. Especially when $p \in (0, 1)$, (12) is neither convex nor non-Lipschitz continuous which brings many computational difficulties. Actually the unconstrained $l_q - l_p$ minimization is strongly NP-hard for any $0 < p < 1$ and $q \ge 1$ [20]. Reweighed minimization algorithm [21–23] is an efficient algorithm for solving $l_2 - l_p (0 < p < 1)$ vector minimization problem which has been extended by Wang et al [24] to solve $l_{2,p}$-based matrix minimization for $0 < p \le 1$. Even the problem considered in [24] is the special case of (12) with $q = p \in (0, 1]$, the idea motivates us to develop an iteratively quadratic algorithm for the generalized $l_{2,q} - l_{2,p}$ matrix minimization for any $q \in [1, 2]$ and $p \in (0, 2)$. Moreover, the convergence analysis will be uniformly demonstrated in Appendix B.

After simply transformation, $\| X \|_{2,p}^p$ can be rewritten as (Proposition A.1 in Appendix A)

$$\| X \|_{2,p}^p = Tr(X^T H X), \tag{19}$$

where

$$H = \begin{cases} \mathrm{diag}\{\frac{1}{\|X^1\|_2^{2-p}}, \frac{1}{\|X^2\|_2^{2-p}}, \ldots, \frac{1}{\|X^d\|_2^{2-p}}\}, & p \in (0, 2); \\ I, & p = 2, \end{cases} \tag{20}$$

and $Tr(\cdot)$ stands for trace operation. If denote

$$G = \begin{cases} \mathrm{diag}\{\frac{1}{\|(AX - Y)^1\|_2^{2-q}}, & q \in [1, 2); \\ \frac{1}{\|(AX - Y)^2\|_2^{2-q}}, \ldots, \frac{1}{\|(AX - Y)^m\|_2^{2-q}}\}, \\ I, & q = 2, \end{cases} \tag{21}$$

we have $\| AX - Y \|_{2,q}^q = Tr((AX - Y)^T G (AX - Y))$. Hence the objective function of (12) can be reformulated to

$$J(X) := \|AX - Y\|_{2,q}^q + \lambda \|X\|_{2,p}^p = Tr((AX - Y)^T G (AX - Y)) + \lambda Tr(X^T H X). \tag{22}$$

It is well known that the *KKT* point of unconstrained optimization problem (12) is also the stationary point of $J(X)$. Compute the derivative of $J(X)$ with respect to matrix X and set it to zero, we get the *KKT* equation of problem (12) as follows,

$$\frac{\partial J(X)}{\partial X} = qA^T G(AX - Y) + \lambda p H X = 0. \tag{23}$$

Thus solving (12) is reduced to finding the solution to Eq. (23). If $A^T G A + \lambda \frac{p}{q} H$ is invertible, Eq. (23) can be solved by

$$X = (A^T G A + \lambda \frac{p}{q} H)^{-1} A^T G Y. \tag{24}$$

To find the iterative solution of fixed-point system (24), let us consider a closely related optimization problem

$$\min_X \hat{J}(X) := Tr((AX - Y)^T G (AX - Y)) + \lambda \frac{p}{q} Tr(X^T H X). \tag{25}$$

$\hat{J}(X)$ is almost equivalent to $J(X)$ in spite of a scaled factor $\frac{p}{q}$ in regularization parameter. If an iterative approximate solution X_k to (25) has been generated, G_k and H_k can be derived from X_k as definitions (20), (21). Then we can compute the next iterative matrix X_{k+1} by solving the following subproblem

$$\min_X Tr((AX - Y)^T G_k (AX - Y)) + \lambda \frac{p}{q} Tr(X^T H_k X). \tag{26}$$

Actually, (26) is a scaled quadratic approximation to $J(X)$ at the iterative point X_k. Let $M_k = A^T G_k A + \lambda \frac{p}{q} H_k$, since G_k and H_k are usually symmetric and positive definite, problem (26) is equivalent to the following quadratic optimization problem

$$\min_X Q_k(X) := \frac{1}{2} Tr(X^T M_k X) - Tr(Y^T G_k A X). \tag{27}$$

The minimizer to $Q_k(X)$ is also the solution to the linear system

$$M_k X = A^T G_k Y. \tag{28}$$

Based on the analysis and Eqs. (19)–(28), the mixed $l_{2,q} - l_{2,p}$ $(1 \le q \le 2, 0 < p \le 2)$ norm based optimization problem (12) can be iteratively solved by a sequence of quadratic approximate subproblems. Hence we name this approach *iterative quadratic method (IQM)* and the convergence analysis is established in Appendix B. $J(X)$ is proved to strictly descend w.r.t. iterations until a solution to (23) is found.

Define

$$\rho_k := \frac{J(X_k) - J(X_{k+1})}{J(X_k)},$$

then $\rho_k > 0$ can be used for stopping criterion in evaluating the approximate solution precision (see Remark B.2 in the Appendix B).

Algorithm 3.1. (IQM for Solving Problem (25))

1. Start: Given $A \in R^{m \times d}$, $Y \in R^{m \times n}$ and precision criterion $\epsilon > 0$. Select parameters $\lambda > 0$, $q \in [1, 2]$ and $p \in (0, 2]$.
2. Set k=1 and initialize $X_1 \in R^{d \times n}$.
3. For $k = 1, 2, \cdots$ until $\rho_k \le \epsilon$ do:

$H_k = \mathrm{diag}\{\frac{1}{\|X_k^i\|_2^{2-p}}\}_{i=1}^d \ (0 < p < 2) or H_k = I_d (p = 2)$;

$C_k = -Y$;

For $i = 1: I$

$B_i = A_i (X_k)_i$;

$C_k = B_i + C_k$;

end

$G_k = \mathrm{diag}\{\frac{1}{\|C_k^i\|_2^{2-q}}\}_{i=1}^m \ (1 \le q < 2) or G_k = I_m (q = 2)$;

$M_k = A^T G_k A + \lambda \frac{p}{q} H_k$;

$X_{k+1} = M_k^{-1} A^T G_k Y$.

It is noticed that each iteration has to compute the inverse of M_k in Algorithm 3.1 which is expensive and unstable. Here we suggest to employ the general Penrose inverse of M_k to update X_{k+1}. Moreover, the main computation $A_i X_i^*$ for classification is a by-product of B_i in computing the approximate solution X^*. Hence identifying the testing

images can be implemented with minor extra calculations.

Algorithm 3.1 is a unified method solving $l_{2,q}-l_{2,p}-$ minimizations for $q\in[1,2]$ and $p\in(0,2]$. This approach provides algorithmic support to adaptively choose better fidelity measurement and regularization in various applications. Especially, IQM provides a uniform algorithm for solving the existed representation based models: sparse representation ($q=2,\ p=1$), collaborative representation ($q=p=2$) and l_1-norm face recognition ($q=p=1$).

4. Practical implementation of JRC

In Algorithm 3.1, IQM has to update the matrix sequence by computing the inverse matrix of M_k. It is expensive in practical implementation especially for large scale problems. Reviewing the procedure of Algorithm 3.1, we notice that $X_{k+1}=M_k^{-1}A^TG_kY$ exactly solves the $k-$ th subproblem (27) which is unnecessary. It is observed that (27) is a quadratic positive definite subproblem. There are a lot of efficient algorithms to solve it approximately, such as conjugate gradient method, gradient methods with different stepsizes, etc. In this paper, we choose Barzilai and Borwein (BB) gradient method due to its simplicity and efficiency. BB gradient method was firstly presented in [25], afterwards extended and developed in many occasions and applications [25–30]. When applied to quadratic matrix optimization subproblem (27), the Barzilai and Borwein gradient method takes on

$$X_k^{(t+1)}=X_k^{(t)}-\alpha_k^{(t)}\nabla Q_k(X_k^{(t)}),\tag{29}$$

where the superscript (t) denotes the $t-$ th iteration solving (27). $\nabla Q_k(X_k^{(t)})$ is the gradient matrix of $Q_k(X)$ with respect to $X_k^{(t)}$

$$\nabla Q_k(X_k^{(t)})=M_kX_k^{(t)}-A^TG_kY.\tag{30}$$

The Barzilai and Borwein gradient method [25] chose the stepsize $\alpha_k^{(t)}$ such that $D_k^{(t)}=\alpha_k^{(t)}I$ has a certain quasi-Newton property

$$D_k^{(t)}=\arg\min_{D=\alpha I}\parallel S_k^{(t-1)}-DT_k^{(t-1)}\parallel_F\tag{31}$$

or

$$D_k^{(t)}=\arg\min_{D=\alpha I}\parallel D^{-1}S_k^{(t-1)}-T_k^{(t-1)}\parallel_F,\tag{32}$$

where $\|\cdot\|_F$ denotes Frobenius matrix norm and $S_k^{(t-1)}$, $T_k^{(t-1)}$ are determined by $X_k^{(t)}$ and $X_k^{(t-1)}$

$$S_k^{(t-1)}:=X_k^{(t)}-X_k^{(t-1)}T_k^{(t-1)}:=\nabla Q_k(X_k^{(t)})-\nabla Q_k(X_k^{(t-1)})=M_kS_k^{(t-1)}.\tag{33}$$

Solving (31) yields two BB stepsizes

$$\alpha_k^{(t)}=\frac{Tr((S_k^{(t-1)})^TT_k^{(t-1)})}{Tr((T_k^{(t-1)})^TT_k^{(t-1)})}\tag{34}$$

and

$$\alpha_k^{(t)}=\frac{Tr((S_k^{(t-1)})^TS_k^{(t-1)})}{Tr((S_k^{(t-1)})^TM_kS_k^{(t-1)})}.\tag{35}$$

Compared with the classical steepest descent method, BB gradient method often needs less computations but converges more rapidly [31]. For optimization problems higher than two dimensions, BB gradient method has theoretical difficulty due to its heavy nonmonotone behavior. But for strongly convex quadratic problem with any dimension, BB method is convergent at $R-$ linear rate [26,28]. BB gradient method has also been applied to matrix optimization problem [32] and exhibits desirable performance. Based on Eqs. (29)–(35), the last step in Algorithm 3.1, $X_{k+1}=M_k^{-1}A^TG_kY$, can be practically substituted by the BB gradient method as the $k-$ th inner loop.

Algorithm 4.1. (BB Gradient Method for Solving Subproblem (27))

1. Start: given the inner loop stopping criterion $\epsilon_2>0$
2. Initialize $X_k^{(1)}=X_k$ and $\nabla Q_k^{(1)}=M_kX_k^{(1)}-A^TG_kY$;
3. For $t=1,2,\cdots$ until $Tr(\nabla Q_k^{(t)})\le\epsilon_2$, output $X_{k+1}=X_k^{(t)}$, do:
   if $t=1$
   $\alpha_k^{(t)}=\frac{Tr((\nabla Q_k^{(t)})^T\nabla Q_k^{(t)})}{Tr((\nabla Q_k^{(t)})^TM_k\nabla Q_k^{(t)})}$;
   else
   $S_k^{(t-1)}=X_k^{(t)}-X_k^{(t-1)}$;
   $T_k^{(t-1)}=\nabla Q_k^{(t)}-\nabla Q_k^{(t-1)}$;
   $\alpha_k^{(t)}$ is computed as (35) or (35);
   end
   $X_k^{(t+1)}=X_k^{(t)}-\alpha_k^{(t)}\nabla Q_k^{(t)}$;
   $\nabla Q_k^{(t+1)}=M_kX_k^{(t+1)}-A^TG_kY$;

In the $k-$ th inner loop, Algorithm 4.1 chooses two initial matrices. One is the approximate solution X_k to the last subproblem and the other one is the Cauchy point from X_k [33]. The Cauchy stepsize $\alpha_k^{(1)}$ is the solution to the one-dimensional optimization problem

$$\min_{\alpha>0}\phi(\alpha):=Q_k(X_k-\alpha\nabla Q_k(X_k))\tag{36}$$

and the Cauchy point is $X_k+\alpha_k^{(1)}\nabla Q_k(X_k))$.

If M_k in Algorithm 3.1 is guaranteed to be positive definite (if not, H_k or G_k can be slightly perturbed), subproblem (27) is a strongly convex quadratic. BB gradient method with step length (35) or (35) will converge at R-linear rate.

For simplicity, we name the IQM with inexact Algorithm 4.1 practically iterative quadratic method (PIQM). Still denote $\{X_k\}$ the approximate matrix sequence generated by PIQM. BB inner loop makes the objective function value of subproblem (27) decline, that is $Q(X_{k+1})\le Q(X_k)$. Then $\{J(X_k)\}$ is always decreasing which is sufficient and necessary for $\{X_k\}$ uniformly converging to the stationary point of problem (12). The following conclusion can be easily derived.

Theorem 4.1. *Denote X^* the output point generated by PIQM, then X^* is an approximate stationary point of $J(X)$. Especially for $q,p\in[1,2]$, X^* is an approximate global minimizer of optimization problem (12). When p is fractional, X^* is one of KKT points.*

A practical version of iteratively quadratic method for joint representation classification can be concluded as follows.

Algorithm 4.2. (PIQM for JRC)

1. Start: load A, Y, choose $\lambda>0$, $q\in[1,2]$, $p\in(0,2]$ and precision levels $\varepsilon_1>0$, $\varepsilon_2>0$.
2. Apply PIQM to(12), output an approximate coding matrix $X^*:=X_{k+1}$.
3. Classify Y by X^*.

5. Experimental results

In this section, the joint representation based classification (JRC) with PIQM will be applied to collective face recognition. Three public datasets are used. Brief description is given as follows. .

AT&T database is formerly known “the ORL database of faces”. It consists of 400 frontal images for 40 individuals. For each subject, 10 pictures were taken at different times, with varying lighting conditions, multiple facial expression, adornments and rotations up to 20 degree. All the images are aligned with dimension 112×92. The database can be retrieved from http://www.cl.cam.ac.uk/Research/DTG/attarchive:-pub/data/att_faces.tar. Z as a 4.5 Mbyte compressed tar file. Typical pictures can be seen in Fig. 2.

Fig. 2. Typical images of AT & T database.

Fig. 3. Typical images of Georgia-Tech database.

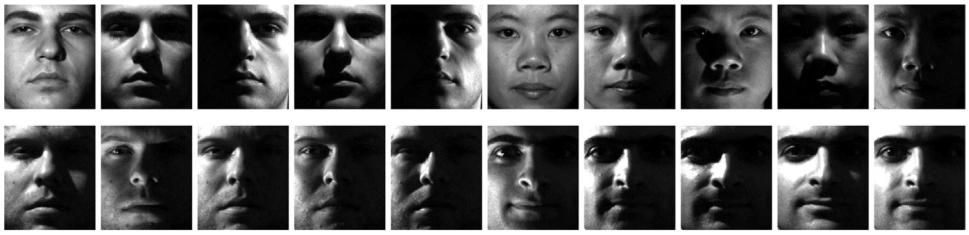


Fig. 4. Typical images of Extended Yale B database.

Georgia-Tech database contains 15 images each of 50 subjects. The images are taken in two or three sessions at different time with different facial expressions, scale and background. The average size of the faces in these images is 480×640 pixels. Georgia Tech face database and the annotation can be found in http://www.anefian.com/research/face_reco.htm. Typical pictures of four persons are shown in Fig. 3.

Extended Yale B database consists of 2414 frontal-face images of 38 subjects. Each subject has around 64 images. The images are cropped and normalized to 192×168 under various laboratory-controlled lighting conditions [34,35]. Fig. 4 displays typical pictures of 4 subjects.

Extensive experiments are conducted for different image sizes and different parameters. To demonstrate the collective efficiency of JRC, all the testing samples are mixed of all the classes. Four comparable schemes are implemented, JRC, SRC, CRC-RLS and traditional SVM classifier. JRC is practically carried out via PIQM while SRC is solved by $l_1 - l_s$ solver [37] and CRC-RLS employs the code from [2]. We realize SVM by the software LIBSVM [38] with linear kernel, the pseudo code can be found in http://www.csie.ntu.edu.tw/cjlin/libsvm/faq.html#f203. All the schemes are implemented by Matlab R2014a(win32) on a typical 4 GiB memory and 2.40 GHz PC.

Considering that JRC is a joint framework including SRC and CRC-RLS, we select six pairs of q, p in [1, 2] and (0, 2] respectively:

$q = p = 2$(corresponding to CRC−RLS), $q = 2$,

$p = 1$(corresponding to SRC),

and other four generalized cases

$q = 1.5\&p = 1$, $q = 1.5\&p = 0.5$, $q = 1\&p = 1$, $q = 1\&p = 0.5$.

The parameter λ in (12) is varied from 0.01 to 10 each 10 times. All the stopping precisions are set 10^{-3}.

All the images are re-sized as in [1,2] without other pre-processing. For AT & T database, the pictures are downsampled to 11×10. The downsampling ratios of Georgia-Tech database and Extended Yale B database are 1/8 and 1/16. For each subject, around 80% pictures are randomly selected for training and the left for testing. For example, 8

Table 1
The recognition accuracy (%) and running time (second) for AT & T database.

Methods	Accuracy	CPU time
SRC	97.5(±0.27)	67.2658
JRC(q=2,p=1)	97.5(±0.09)	0.1612
CRC-RLS	95.0(±0.49)	0.0872
JRC(q=p=2)	97.5(±0.27)	0.0073
SVM	95.0(±0.47)	0.0667
JRC(q=1.5,p=1)	97.5(±0.18)	0.3867
JRC(q=1.5,p=0.5)	95.0(±0.38)	1.8756
JRC(q=p=1)	97.5(±0.18)	0.1994
JRC(q=1,p=0.5)	97.5(±0.18)	0.1640

Table 2
The recognition accuracy (%) and CPU time (second) for Georgia-Tech database.

	Downsampling ratio 1/8		Downsampling ratio 1/16	
Methods	Accuracy	Time	Accuracy	Time
SRC	99.33(±0.17)	2843	97.33(±0.57)	1197
JRC(q=2,p=1)	99.33(±0.00)	2.41	97.33(±0.27)	1.07
CRC-RLS	98.00(±0.93)	1.95	96.67(±1.12)	0.66
JRC(q=p=2)	99.33(±0.35)	0.97	98.67(±0.46)	0.17
SVM	96.67(±0.62)	5.09	96.67(±0.96)	1.46
JRC(q=1.5,p=1)	99.33(±0.12)	4.89	98.67(±0.52)	3.86
JRC(q=1.5,p=0.5)	99.33(±0.46)	4.89	98.67(±0.52)	3.89
JRC(q=p=1)	99.33(±0.46)	5.54	99.33(±0.22)	1.11
JRC(q=1,p=0.5)	99.33(±0.51)	4.79	99.33(±0.22)	1.09

Table 3
The recognition accuracy (%) and CPU time (second) for Extended Yale B database.

	Down sampling ratio 1/8		Down sampling ratio 1/16	
Methods	Accuracy	Time	Accuracy	Time
SRC	96.76(±0.62)	4828	96.36(±0.64)	668.53
JRC(q=2,p=1)	96.96(±0.43)	22.67	76.11(±2.70)	164.71
CRC-RLS	96.76(±0.83)	2.02	95.55(±1.24)	1.9
JRC(q=p=2)	96.96(±0.30)	0.75	91.29(±0.98)	0.34
SVM	95.55(±0.68)	6.12	94.33(±0.77)	2.61
JRC(q=1.5,p=1)	96.96(±0.49)	22.04	87.05(±1.64)	22.03
JRC(q=1.5,p=0.5)	96.96(±0.66)	54.21	65.59(±3.57)	101.59
JRC(q=p=1)	96.96(±0.49)	27.08	90.49(±1.15)	20.51
JRC(q=1,p=0.5)	96.96(±0.49)	26.87	91.29(±0.96)	25.23

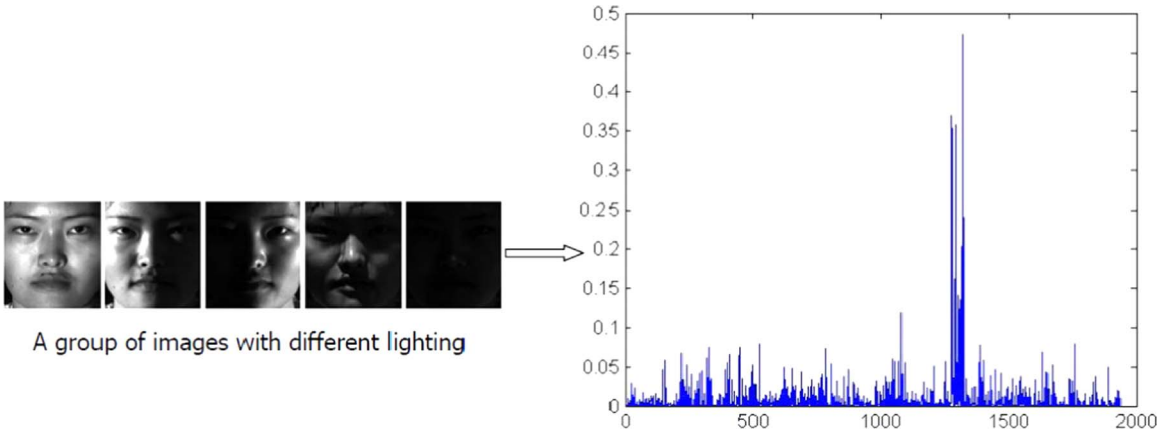


Fig. 5. A group of images under different lighting (left) and the recovered row coefficients by JRC (q=2, p=1) (right).

Table 4
The recognition accuracy (%) of different testing percents on GT database.

Testing percent (%)	30	40	50	60
SRC	97.6	97.33	87.53	81.11
JRC(q=2,p=1)	97.6	97.33	87.25	81.11
CRC-RLS	96.4	95.98	87	80.44
JRC(q=p=2)	97.6	97.33	87.79	81.33
SVM	96.1	95.77	87.53	81.11
JRC(q=1.5,p=1)	98	97.6	87.79	81.33
JRC(q=1.5,p=0.5)	98	97.6	87.79	81.33
JRC(q=p=1)	98	98	87.79	81.56
JRC(q=1,p=0.5)	98	98	87.79	82

pictures of each individual in AT & T database are randomly picked out for training while the left 2 are for testing. For statistical stability, we randomly generate 5 different training and testing dataset pairs for experiments. The recognition accuracy (mean ± deviation) and running time (average) are reported in Table 1–3.

Based on the experimental results on three databases, we draw the following conclusions:

- Jointly representing all the testing images simultaneously does accelerate face recognition. On all the databases, JRC (q=p=2) is the fastest one. The CPU time is thousand times less than that of SRC. For example, JRC (q=p=2) classifies 150 images in 0.17 s on Georgia-Tech database with downsampling ratio 1/16. And the accuracy rate is 98.67%, outperforming SRC (97.33%), CRC-RLS (96.67%) and SVM (96.67%). More details can be found in Table 1–3.
- JRC exhibits competitive performance in recognition accuracy. On AT & T database, the recognition rate of JRC is 97.5%, compared to 97.5% for SRC, 95% for CRC-RLS and SVM. On Georgia-Tech database, JRC achieves the best recognition rate (99.33%), consistently exceeds other classification schemes. On Yale B database with downsampling ratio 1/8, JRC also outperforms other methods in recognition accuracy.

Unfortunately, JRC does not keep the best achievement on downsampling ratio 1/16. The possible reason is that some pictures with strong contrast of lighting (see Fig. 4) aggravates the noise for other images in joint coding. The deviations of the recognition accuracy are consequently large. But for a single subject, joint representation can easily make up the weakness caused by lower ration downsampling and occlusion. We randomly select a subject in Extended Yale B database, five typical images are chosen with different lighting (normal, right-half dark, left-half dark, shallow dark and total dark). All the images are downsampled to 12×10 (1/16 downsampling ratio) and the joint row coefficients recovered by JRC($q = 2$, $p = 1$) are showed in Fig. 5. The largest group coefficients are corresponding to subject 26 which five images belong to. Even for the total dark sample, the row coefficient is correctly clustered. In this sense, JRC can efficiently handle collective face recognition with lower ration downsampling and lighting variation.

- In addition, we observe the effect of different testing-sample percent on recognition rate. Take Georgia-Tech database with 1/8 downsampling ration for example, the recognition accuracy rates (average of 5 times) according to randomly selected 30% – 60% testing samples are reported in Table 4.
- Different $q \in [1, 2]$ and $p \in (0, 2]$ for JRC indicate different feature pattern behind in the image set. Taking JRC ($q = 2$, $p = 1$) for example, the joint sparsity of representation and correlation of

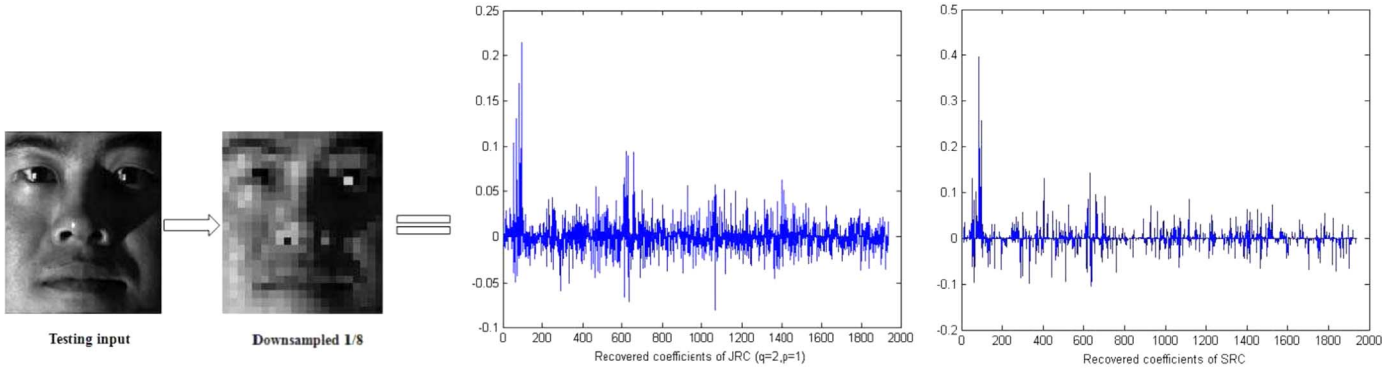
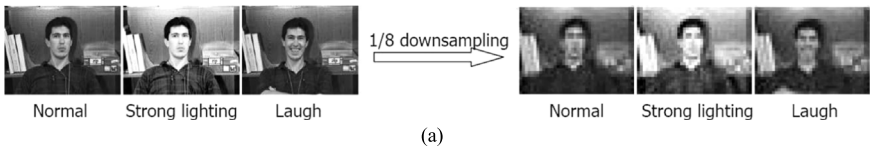


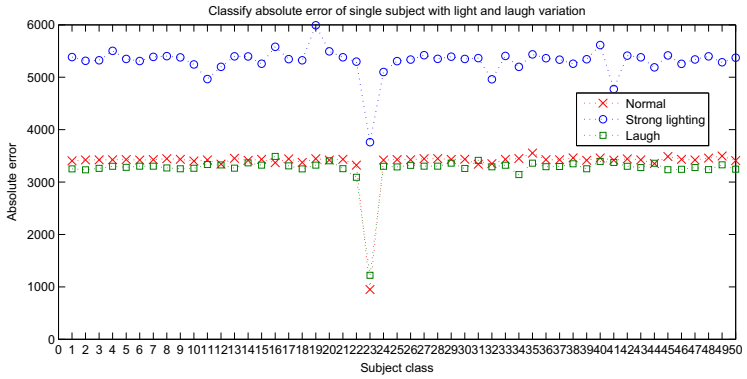
Fig. 6. Original image and the 1/8 downsampled image. The joint column coefficients recovered by JRC (q=2, p=1) and vector coefficients recovered by SRC.

multiple images are combined. The column representation coefficients reveal the joint effect of JRC ($q = 2$, $p = 1$), Fig. 6 gives an example from Yale B database. Compared to SRC, JRC($q = 2$, $p = 1$) concentrates a group sparsity but not a single one. Actually, the other testing samples (12 pictures) of the same subject also have the similar group representation pattern.

- Furthermore, we also conduct an experiment about the effect of lighting and facial expression on JRC. A normal and laughing images of the same subject are chosen from Georgia-Tech database. As a contrast, we get strong lighted picture from the normal one (Fig. 7(a)). All the images are downsampled to 60×80 (1/8 downsampling ration) and tested by joint sparse representation classification (JRC ($q = 2$, $p = 1$)). We plot the absolute errors of 3 testing samples over 50 training classes (Fig. 7(b)). All the testing images get the least error at subject 23 which gives the correct recognition. Fig. 7 (b) shows that the laugh image has almost the same classification errors to the normal one while the strong lighting affects the quantity of absolute errors. Actually, the recognition

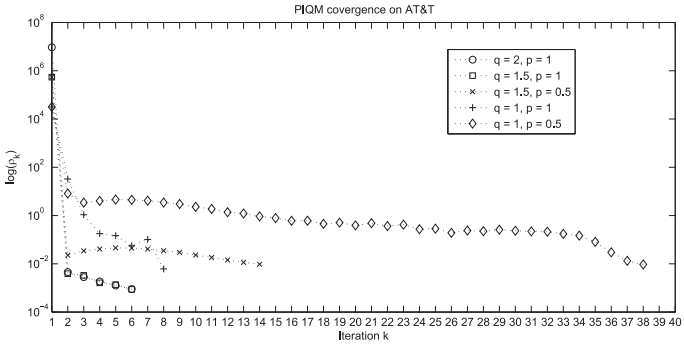


(a)

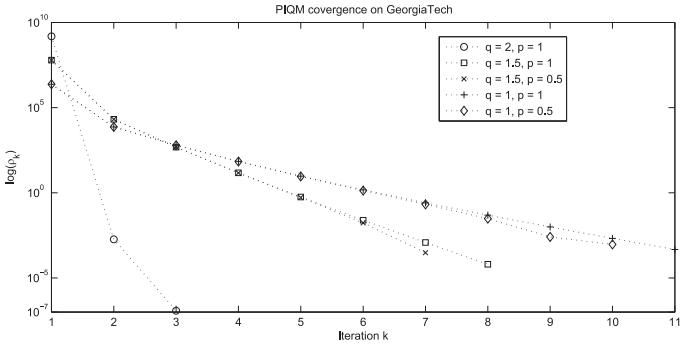


(b)

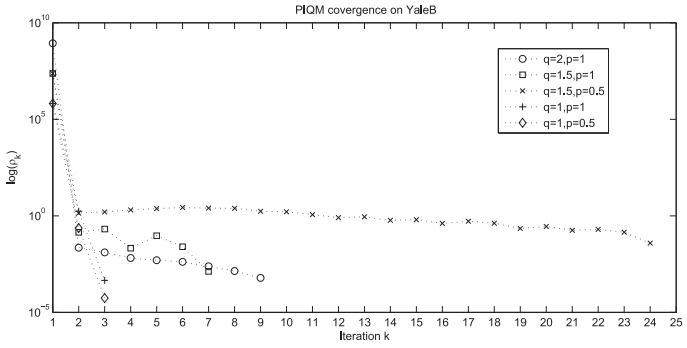
Fig. 7. (a) The original (left) and 1/8 downsampled (right) testing images w.r.t. normal, strong lighting and laugh cases. (b) The classification errors corresponding to 50 classes by JRC (q=2, p=1). The testing images belong to subject 23 of GT database.



(a)



(b)



(c)

Fig. 8. (a) PIQM on AT & T(b) PIQM on Georgia-Tech (c) PIQM on Yale B.

Table 5
The recognition accuracy (%) for different λ on Extended Yale B database with downsampling ratio 1/8.

λ =	0.01	0.1	1	10	100
CRC-RLS	28.34	66.82	95	96.76	96.76
JRC(q=p=2)	96.96	96.96	96.96	96.96	96.96

accuracy of JRC for a set of images of the same subject is not sensitive to different lighting which has been partly illustrated in Fig. 5.

- The convergence behavior of PIQM for JRC is displayed in Figs. 8. The x axis denotes the iteration and y-axis stands for the logarithm of ρ_k. PIQM converges within 40 steps on three databases for all jointly sparse models (five pairs of q and p). JRC (q=p=2) always converges in three iterations hence its plot is omitted here. Anyway, PIQM provides a uniform algorithm for varied JRC with respect to $q \in [1, 2]$ and $p \in (0, 2]$>.
- From Table 1–3, it is observed that CRC-RLS has a fairly good performance in recognition accuracy and CPU time. But CRC-RLS is heavily sensitive to the regularization parameter λ (see Table 5) because it has a smooth regularizer. By comparison, JRC (q=p=2) is more stable for its joint technique. Multiple images have complementary effect for recognition especially when the formulation is ill-posed.

Remark 5.1. In terms of the experiments, it is also an interesting subject to validate the performance of JRC on a true sequence of images form video which we have alluded to in the introduction. Due to the focus of this paper, it is not be covered experimentally. We will consider it in future efforts to explore this a bit further.

6. Conclusions

In this paper, a joint representation classification (JRC) for collective face recognition is proposed. By aligning all the testing images into a matrix, joint representation coding is reduced to a kind of generalized matrix optimization problem. Accordingly a unified algorithm and its practical implementation are developed to solve the mixed $l_{2,q} - l_{2,p}$ matrix minimizations for any $q \in [1, 2]$ and $p \in (0, 2]$. Experiment results on three datasets validate the collective performance of JRC. For a given set of images, JRC not only recognizes different subjects but also clusters multiple images belonging to the same one. Moreover, the effect of facial expression and lighting variation on JRC is analyzed. The joint representation based classification is confirmed to improve the performance in recognition rate and running time than state-of-the-art methods.

Acknowledgement

The authors would like to thank the anonymous reviewers very much for their helpful suggestions to improve this paper. The first author also thanks software engineer Aiwen Luo for his code support.

Appendix A. Appendix

Proposition A.1. *Given matrix $X \in R^{d\times n}$, we have $\|X\|_{2,p}^p = Tr(X^THX)$, where*

$$H = \begin{cases} \mathrm{diag}\{\frac{1}{\|X^1\|_2^{2-p}}, \frac{1}{\|X^2\|_2^{2-p}}, \ldots, \frac{1}{\|X^d\|_2^{2-p}}\}, & p \in (0, 2); \\ I, & p = 2, \end{cases} \tag{37}$$

and $Tr(\cdot)$ stands for trace operation.

Proof. Let X be partitioned according to rows, that is $X = \begin{bmatrix} X^1 \\ X^2 \\ \vdots \\ X^d \end{bmatrix}$, where $X^i \in R^n$ denotes the ith row of X. When p=2, $\| X \|_{2,2}$ is actually Frobenius norm. Hence

$$\|X\|_{2,2}^2 = \|X\|_F^2 = Tr(X^TX).$$

When $p \in (0, 2)$, we have

$$\|X\|_{2,p}^p = \sum_{i=1}^d \|X^i\|_2^p = \sum_{i=1}^d \frac{\|X^i\|_2^2}{\|X^i\|_2^{2-p}} = \sum_{i=1}^d \frac{Tr((X^i)^TX^i)}{\|X^i\|_2^{2-p}} = Tr(\sum_{i=1}^d \frac{(X^i)^TX^i}{\|X^i\|_2^{2-p}}) = Tr(X^THX), \tag{38}$$

where $H = \mathrm{diag}\{\frac{1}{\|X^1\|_2^{2-p}}, \frac{1}{\|X^2\|_2^{2-p}}, \ldots, \frac{1}{\|X^d\|_2^{2-p}}\}$. □

Appendix B. Appendix

In this part, we will demonstrate the theoretical convergence of Algorithm 3.1. The key point is that the objective function $J(X)$ strictly decreases with respect to iterations until the matrix sequence $\{X_k\}$ converges to a stationary point of $J(X)$.

Lemma B.1. *Let $\varphi(t) = t - at^{\frac{1}{a}}$, where $a \in (0, 1)$. Then for any $t > 0$, $\varphi(t) \le 1 - a$, and t=1 is the unique maximizer.*

Proof. Taking the derivative of $\varphi(t)$ and set to zero, that is

$$\varphi'(t) = 1 - t^{\frac{1}{a}-1} = 0$$

then $\varphi'(t) = 0$ has the unique solution t=1 for any $a \in (0, 1)$ which is just the maximizer of $\varphi(t)$ in $(0, +\infty)$. □

Based on Lemma B.1, for a given $a \in (0, 1)$,

$$t - at^{\frac{1}{a}} \le (1 - a) \tag{39}$$

holds in $(0, +\infty)$ and "=" is active if and only if t=1. Let a takes special values such as $a = \frac{q}{2}(q \in [1, 2))$ or $a = \frac{p}{2}$ $(p \in (0, 2))$, inequality (39) will result in the following formulas associated with $\|AX - Y\|_{2,q}^q$ and $\|X\|_{2,p}^p$.

Lemma B.2. *Given X_k and X_{k+1} in $R^{d\times n}$, the following inequalities hold,*

$$\|AX_{k+1} - Y\|_{2,q}^q - \frac{q}{2}\sum_{i=1}^{m}\frac{\|(AX_{k+1} - Y)^i\|_2^2}{\|(AX_k - Y)^i\|_2^{2-q}} \le (1 - \frac{q}{2})\|AX_k - Y\|_{2,q}^q \tag{40}$$

and

$$\|X_{k+1}\|_{2,p}^p - \frac{p}{2}\sum_{i=1}^{d}\frac{\|X_{k+1}^i\|_2^2}{\|X_k^i\|_2^{2-p}} \le (1 - \frac{p}{2})\|X_k\|_{2,p}^p \tag{41}$$

for any $q \in [1, 2)$ and $p \in (0, 2)$. Moreover, the equalities in Eqs. (40) and (41) *hold if and only if $\|(AX_{k+1} - Y)^i\|_2 = \|(AX_k - Y)^i\|_2$ for $i = 1, 2, \ldots, m$ and $\|X_{k+1}^i\|_2 = \|X_k^i\|_2$ for $i = 1, 2, \ldots, d$.*

Proof. Substituting $t_1 = \frac{\|(AX_{k+1} - Y)^i\|_2^q}{\|(AX_k - Y)^i\|_2^q}$ and setting $a_1 = \frac{q}{2}$ in Lemma B.1, we obtain

$$\frac{\|(AX_{k+1} - Y)^i\|_2^q}{\|(AX_k - Y)^i\|_2^q} - \frac{q}{2}\frac{\|(AX_{k+1} - Y)^i\|_2^2}{\|(AX_k - Y)^i\|_2^2} \le 1 - \frac{q}{2}. \tag{42}$$

Similarly taking $t_2 = \frac{\|X_{k+1}^i\|_2^p}{\|X_k^i\|_2^p}$ and $a_2 = \frac{p}{2}$ in $\varphi(t)$, we have

$$\frac{\|X_{k+1}^i\|_2^p}{\|X_k^i\|_2^p} - \frac{p}{2}\frac{\|X_{k+1}^i\|_2^2}{\|X_k^i\|_2^2} \le 1 - \frac{p}{2}. \tag{43}$$

Multiplying Eqs. (42) and (43) by $\|(AX_k - Y)^i\|_2^q$ and $\|X_k^i\|_2^p$ respectively, we have the following inequalities simultaneously

$$\|(AX_{k+1} - Y)^i\|_2^p - \frac{q}{2}\frac{\|(AX_{k+1} - Y)^i\|_2^2}{\|(AX_k - Y)^i\|_2^{2-q}} \le (1 - \frac{q}{2})\|(AX_k - Y)^i\|_2^q \tag{44}$$

for $i = 1, 2, \ldots, m$, and

$$\|X_{k+1}^i\|_2^p - \frac{p}{2}\frac{\|X_{k+1}^i\|_2^2}{\|X_k^i\|_2^{2-p}} \le (1 - \frac{p}{2})\|X_k^i\|_2^p, \quad i = 1, 2, \ldots, d. \tag{45}$$

Summing up i in formulas (44) and (45), we can derive (40) and (41).

Based on Lemma B.1, $t_1 = 1$ and $t_2 = 1$ are the unique minimizers for $\varphi(t)$ in $(0, +\infty)$ when $a_1 = \frac{q}{2}$ and $a_2 = \frac{p}{2}$ respectively. Namely, $\|(AX_{k+1} - Y)^i\|_2 = \|(AX_k - Y)^i\|_2$ and $\|X_{k+1}^i\|_2 = \|X_k^i\|_2$ are necessary and sufficient for equalities hold in (44) and (45) respectively. □.

Remark B.1. (40) and (41) are established nothing to do with Algorithm 3.1. The inequalities express the innate properties of mixed matrix norms $l_{2,q} - l_{2,p}$ for $q \in [1, 2)$ and $p \in (0, 2)$.

Theorem B.1. *Suppose that $\{X_k\}$ is the matrix sequence generated by* Algorithm 3.1. *Then $J(X_k)$ strictly decreases with respect tok for any $1 \le q \le 2$ and $0 < p \le 2$ until $\{X_k\}$ converges to a stationary point of $J(X)$.*

Proof. Based on the procedure of Algorithm 3.1, X_{k+1} is the solution to linear system (28), also the optimal matrix of problems (26) and (27). Thus we have

$$Q_k(X_{k+1}) \le Q_k(X_k). \tag{46}$$

For $q \in [1, 2)$ and $p \in (0, 2)$, (46) is equivalent to

$$q\sum_{i=1}^{m}\frac{\|(AX_{k+1} - Y)^i\|_2^2}{\|(AX_k - Y)^i\|_2^{2-q}} + \lambda p\sum_{i=1}^{d}\frac{\|X_{k+1}^i\|_2^2}{\|X_k^i\|_2^{2-p}} \le q\|AX_k - Y\|_{2,q}^q + \lambda p\|X_k\|_{2,p}^p, \tag{47}$$

It is noticed that $J(X_k) = \|AX_k - Y\|_{2,p}^p + \lambda\|X_k\|_{2,p}^p$. Adding inequalities (40) and $\lambda\cdot$ (41), the following formula will be derived

$$J(X_{k+1}) - (\frac{q}{2}\sum_{i=1}^{m}\frac{\|(AX_{k+1} - Y)^i\|_2^2}{\|(AX_k - Y)^i\|_2^{2-q}} + \lambda\frac{p}{2}\sum_{i=1}^{d}\frac{\|X_{k+1}^i\|_2^2}{\|X_k^i\|_2^{2-p}}) \le J(X_k) - (\frac{q}{2}\|AX_k - Y\|_{2,q}^q + \lambda\frac{p}{2}\|X_k\|_{2,p}^p). \tag{48}$$

Based on (47) and (48), $J(X_{k+1}) \le J(X_k)$ can be easily derived for $q \in [1, 2)$ and $p \in (0, 2)$. □

For q=2 or p=2, the inequalities is much easier to derive. Taking q=2 and $p \in (0, 2)$ for example, (46) is reduced to

$$\|AX_{k+1} - Y)\|_{2,2}^2 + \lambda\frac{p}{2}\sum_{i=1}^{d}\frac{\|X_{k+1}^i\|_2^2}{\|X_k^i\|_2^{2-p}} \le \|AX_k - Y\|_{2,2}^2 + \lambda\frac{p}{2}\|X_k\|_{2,p}^p, \tag{49}$$

Combining the formulas (49) and (41), we also obtain $J(X_{k+1}) \le J(X_k)$. In the case of $q = 2$, $p \in (0, 2)$ or $q = p = 2$, $J(X_{k+1}) \le J(X_k)$ can be deduced analogously.

Once $J(X_{k+1}) = J(X_k)$ happens for some k, the equalities in (47) and (48) (or (49)) hold. Hence the equalities in (40) and (41) are active. From Lemma B.2, we obtain $\|(AX_{k+1} - Y)^i\|_2 = \|(AX_k - Y)^i\|_2$ for $i = 1, 2, \ldots, m$ and $\|X_{k+1}^i\|_2 = \|X_k^i\|_2$ for $i = 1, 2, \ldots, d$. Thus $G_{k+1} = G_k$ and $H_{k+1} = H_k$ which implies that X_{k+1} is a solution to (24).

The objective function sequence $\{J(X_k)\}$ is decreasing and lower bounded. Hence $\{J(X_k)\}$ eventually converges to some minimum of problem (12). The descending quantity measures the convergence precision.

Remark B.2. The stopping criterion of Algorithm 3.1 can be chosen as $J(X_k) - J(X_{k+1}) \le \epsilon$ or $\rho_k := \frac{J(X_k) - J(X_{k+1})}{J(X_k)} \le \epsilon$ for some required precision $\epsilon > 0$.Theoretically, $X_k^i = 0$ or $C_k^i = 0$ likely occurs in some step k, then H_k and G_k can not be well updated for non-Frobenius norm case ($0 < p < 2$ and $1 \le q < 2$). We deal with it by perturbing with $\delta > 0$ such that $\{H_k\}_{ii} = \delta^{p-2} > 0$ and $\{G_k\}_{ii} = \delta^{q-2} > 0$. The descending of $\{J(X_k)\}$ is relaxed to

$$J(X_{k+1}) \le J(X_k) + (1 - \frac{p}{2})\delta^p \quad \text{or } J(X_{k+1}) \le J(X_k) + (1 - \frac{q}{2})\delta^q. \tag{50}$$

If the convergence precision ϵ is chosen fairly larger than perturbation δ ($\epsilon \gg \delta$), perturbed $J(X_k)$ can be still considered approximate decreasing. As a matter of fact, $X_k^i = 0$ and $C_k^i = 0$ never happen in practical implementation.

References

[1] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. PAMI 31 (2) (2009) 210–227.
[2] L. Zhang, M. Yang, X.C. in: Feng, Sparse representation or collaborative representation: which help face recognition?Proceedings of the 13th International Conference on Computer Vision, 2011, pp. 471–478.
[3] J. Wright, Y. Ma, Dense error correction via l_1 minimizatioin, IEEE Trans. Inf. Theory 56 (7) (2010) 3540–3560.
[4] S.H. Gao, I.W.H. Tsang, L.T. Chia, Kernel sparse representation for image classification and face recognition, in: ECCV, 2010.
[5] J.Z. Huang, X.L. Huang, D. Metaxas, Simultaneous Image Transformation and Sparse Representation Recovery, in: CVPR, 2008.
[6] A. Wagner, J. Wright, W. Xu, Y. Ma, Towards a Practical Face Recognition Sysstem: Robust Registration and Illumination by Sparse Representation, in: CVPR, 2009.
[7] Y. Peng, A. Wagner, J. Wright, W. Xu, Y. Ma, RASL: robust alignment by sparse and low-rank decomposition for linearly correlated image, IEEE Trans. PAMI 34 (11) (2012) 2233–2246.
[8] R. Chartrand, W.T. Yin, Iteratively reweighed algorithms for compressive sensing, in: Proceedings of the 33rd International Conference on Acoustics, Speech, and Signal Processing, 2008, pp. 3869–3872.
[9] R. Chartrand, Exact reconstructions of sparse signals via nonconvex minimization, IEEE Signal Process. Lett. 14 (10) (2007) 707–710.
[10] Z.B. Xu, H. Zhang, Y. Wang, X.Y. Chang, Y. Liang, $L_{\frac{1}{2}}$ regularizer, Sci. China: Ser. F 52 (6) (2010) 1159–1169.
[11] R. Brunelli, T. Poggio, Face recognition: features versue templates, IEEE Trans. Pattern Anal. Mach. Intell. 15 (10) (1993) 1042–1052.
[12] T. Heseltine, N. Pears, J. Austin, Z. Chen, Face recognition: a comparison of appearance-based approaches, Proc. 7th Digit. Image Comput.: Tech. Appl. (2003) 10–12.
[13] R.P. Wang, S.G. Shan, X.L. Chen, W. Gao, Manifold-manifold method distance with application to face recognition based on image set, IEEE Trans. Image Process. 21 (10) (2012) 4466–4479.
[14] P.F. Zhu, W.M. Zuo, L. Zhang, S.C.K. Shiu, D. Zhang, Image set-based collaborative representation for face recognition, IEEE Trans. Inf. Forensics Secur. 9 (7) (2014) 1120–1132.
[15] T. Kim, J. Kittler, R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 1005–1018.
[16] A. Rakotomamonjy, R. Flamary, G. Gasso, S. Canu, $l_p - l_q$ penalty for sparse linear and sparse multiple kernel multitask learning, IEEE Trans. Neural Netw. 22 (8) (2011) 1307–1320.
[17] S. Suvrit, Fast projection onto $l_{1,q}$-norm balls for grouped feature selection, in: Proceeding of Machine Learning and Knowledge Discovery in Databases, 2011, Athens, Greece.
[18] S. Sumit, M.P. Vishal, M.N. Nasser, C. Rama, Joint sparse representation for robust multimodal biometrics recognition, IEEE Trans. PAMI 36 (1) (2014) 113–126.
[19] H. Wang, F.P. Nie, W.D. Cai, H. Huang, Semi-supervised robust dictionary learning via efficient $l_{2,0+}$-norms minimizations, IEEE International Conference on Computer Vision, 2013, pp. 1145–1152.
[20] X.J. Chen, D.D. Ge, Z.Zh. Wang, Y.Y. Ye, Complexity of unconstrained $L_2 - L_p$ minimization, Math. Program. Ser. A 143 (2014) 371–383.
[21] X.J. Chen, Weijun Zhou, Convergence of the reweighted l_1 minimization algorithm for $l_2 - l_p$ minimization, Comput. Optim. Appl. 59 (2014) 47–61.
[22] Z.S. Lu, Iterative reweighted minimization methods for regularized unconstrained nonlinear programming, Math. Program. 147 (2014) 277–307.
[23] Emmanuel J. Candès, Michael B. Wakin, Stephen P. Boyd, Enhancing sparsity by reweighed l_1 minimization, J. Fourier Anal. Appl. 14 (5) (2008) 877–905.
[24] L.P. Wang, S.C. Chen, Y.P. Wang, A unified algorithm for mixed $l_{2,p}$-minimizations and its application in feature selection, Comput. Optim. Appl. 58 (2014) 409–421.
[25] J. Barzilai, J.M. Borwein, Two-point step size gradient methods, IMA J. Numer. Anal. 8 (1988) 141–148.
[26] M. Raydan, On the barzilai and borwein choice of steplength for the gradient method, IMA J. Numer. Anal. 13 (1993) 321–326.
[27] M. Raydan, The barzilai and borwein gradient method for the large scale unconstrained minimization problem, SIAM J. Optim. 7 (1997) 26–33.
[28] Y.H. Dai, L.Z. Liao, R-linear convergence of the barzilai and borwein gradient method, IMA J. Numer. Anal. 26 (2002) 1–10.
[29] Y.H. Dai, R. Fletcher, New algorithms for singly linearly constrained quadratic programs subject to lower and upper bounds, Math. Program. Ser. A 106 (2006) 403–421.
[30] Y.X. Yuan, A new stepsize for the steepest descent method, J. Comput. Math. 24 (2) (2006) 149–156.
[31] R. Fletcher, Low storage method for unconstrained optimization, Lect. Appl. Math. (AMS) 26 (1990) 165–179.
[32] B. Jiang, C.F. Cui, Y.H. Dai, Unconstrained optimization models for computing several extreme eigenpairs of real symmetric matrices, Pac. J. Optimzation 10 (1) (2014) 55–71.
[33] A. Cauchy, Méthode générale pour la résolution des systèms d'équations simultanées, Comp. Rend. Sci. Pari. 25 (1847) 141–148.
[34] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. PAMI 23 (6) (2001) 643–660.
[35] L. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Trans. PAMI 27 (5) (2005) 684–698.
[37] S.J. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, A interior-point method for large-scale l1-regularized least squares, IEEE J. Sel. Top Signal Process. 1 (4) (2007) 606–617.
[38] R.E. Fan, P.H. Chen, C.J. Lin, Working set selection using second order information for training SVM, J. Mach. Learn. Res. 6 (2005) 1889–1918.

Liping Wang received her PhD degree from Institute of Computational Mathematics and Scientific/Engineering Computing of the Chinese Academy of Sciences in 2004. Now she works as an associate professor in Department of Mathematics, Nanjing University of Aeronautics and Astronautics

Songcan Chen received his B.S. degree in mathematics from Hangzhou University (now merged into Zhejiang University) in 1983. In December. 1985, he completed his M.S. degree in computer applications at Shanghai Jiaotong University and then worked at the Nanjing University of Aeronautics and Astronautics (NUAA) in January 1986 as an assistant lecturer. There he received a Ph.D. degree, in 1997, in communication and information systems. Since 1998, as a full-time professor, he has been with the computer science and engineering department at NUAA. His research interests include pattern recognition, machine learning and neural computing. In these fields, he has authored or coauthored over 130 scientific journal papers.