HSR: $L_{1/2}$-regularized sparse representation for fast face recognition using hierarchical feature selection

Bo Han¹ · Bo He¹ · Tingting Sun¹ · Tianhong Yan² · Mengmeng Ma¹ · Yue Shen¹ · Amaury Lendasse³,⁴

Received: 25 December 2014 / Accepted: 2 April 2015 / Published online: 17 April 2015
© The Natural Computing Applications Forum 2015

Abstract In this paper, we propose a novel method for fast face recognition called $L_{1/2}$-regularized sparse representation using hierarchical feature selection. By employing hierarchical feature selection, we can compress the scale and dimension of global dictionary, which directly contributes to the decrease of computational cost in sparse representation that our approach is strongly rooted in. It consists of Gabor wavelets and extreme learning machine auto-encoder (ELM-AE) hierarchically. For Gabor wavelets’ part, local features can be extracted at multiple scales and orientations to form Gabor-feature-based image, which in turn improves the recognition rate. Besides, in the presence of occluded face image, the scale of Gabor-feature-based global dictionary can be compressed accordingly because redundancies exist in Gabor-feature-based occlusion dictionary. For ELM-AE part, the dimension of Gabor-feature-based global dictionary can be compressed because high-dimensional face images can be rapidly represented by low-dimensional feature. By introducing $L_{1/2}$ regularization, our approach can produce sparser and more robust representation compared to $L_1$-regularized sparse representation-based classification (SRC), which also contributes to the decrease of the computational cost in sparse representation. In comparison with related work such as SRC and Gabor-feature-based SRC, experimental results on a variety of face databases demonstrate the great advantage of our method for computational cost. Moreover, we also achieve approximate or even better recognition rate.

Keywords Fast face recognition · Hierarchical feature selection · Gabor wavelets · ELM-AE · Sparse representation · $L_{1/2}$ regularization · HSR

1 Introduction

The technique of face recognition plays an important role in people’s life ranging from commercial to law enforcement applications, such as real-time surveillance, biometric personal identification and information security [1]. It is one of the most challenging topics in the interface of computer vision and cognitive science. Over past 40 years, extensive research on face recognition has been conducted by many psychophysicists, neuroscientists and engineers. In general views, the definition of face recognition can be formulated as that different faces in a static image can be identified using a database of stored faces. If the number of the face images in the database is fixed, the quality of face recognition will be mainly related to feature extraction and recognition modeling.

For feature extraction, there are roughly two kinds of popular face features including holistic features and local features. However, the classical methods using holistic features such as Eigenface [2], Fisherface [3] and
Randomface are hardly to reveal the essential structures of high-dimensional faces \cite{4}. Therefore, researchers recently prefer local-feature-based methods such as subspace learning \cite{5} or manifold representation \cite{6}. On one hand, high-dimensional images can be effectively projected into low-dimensional subspace or submanifold. On the other hand, compared to holistic-feature-based approaches, local-feature-based approaches are always less sensitive to variations of illumination, viewpoint and expression, which in turn improves the recognition rate.

For recognition modeling, lots of researchers usually evaluate the performance of model by recognition rate instead of computational cost. Recently, Wright et al. \cite{7} reported their work called the sparse representation-based classification (SRC). It can represent the testing image sparsely using training samples via $L_1$-norm minimization, which can be solved by balancing the minimum reconstructed error and the sparsest coefficients. Experimental results showed that the recognition rate of SRC is much higher than that of classical algorithms such as Nearest Neighbor, Nearest Subspace and Linear Support Vector Machine (SVM). However, there are three drawbacks behind the SRC. First, SRC is based on the holistic features, which cannot exactly capture the partial deformation of the face images. Second, $L_1$-regularized SRC usually runs slowly for high-dimensional face images. Third, in the presence of occluded face images, Wright et al. introduce an occlusion dictionary to sparsely code the occluded components in face images. However, the computational cost of SRC increases drastically because of large number of elements in the occlusion dictionary. Therefore, the computational cost of SRC limits its application in real-time area, which increasingly attracts researchers’ attention to solve this issue.

Recently, Yang and Zhang’s work \cite{8} claimed that if Gabor wavelets \cite{9} can be employed in feature extraction, it is possible to obtain a much more compact occlusion dictionary in the presence of occluded faces, which not only speeds up the computation but also improves the recognition rate. Although the GSRC provides us a good insight about how to reduce the computational cost of SRC in the presence of occluded faces, there is still an essence issue to be addressed. Namely, the computational cost of sparse representation is highly related to three aspects including the dimension of face images, the scale of occlusion dictionary and the speed of regularized optimization. If we want to reduce computational cost of SRC in a general condition, on the one hand, we should effectively project high-dimensional faces into low-dimensional features. On the other hand, we should find a sparser representation than $L_1$-regularized SRC.

Inspired by these observations, we propose a novel method for fast face recognition called $L_{1/2}$-regularized sparse representation using hierarchical feature selection (HSR). In order to decrease the computational cost of sparse representation, we employ hierarchical feature selection in the feature extraction, which consists of Gabor wavelets and extreme learning machine auto-encoder (ELM-AE) \cite{10}. To be more specific, Gabor wavelets could effectively extract local features at multiple scales and orientations \cite{11} forming Gabor-feature-based images, which can greatly improve the recognition rate. Moreover, in the presence of occluded faces, we can obtain a compact occlusion dictionary via sparse coding because of redundancies in Gabor-feature-based occlusion dictionary; thus, the scale of global dictionary can be decreased accordingly \cite{12}. In addition, high-dimensional face images can be effectively represented by low-dimensional features via ELM-AE; thus, the dimension of global dictionary can be decreased accordingly. Also the computational cost of ELM-AE is much less than Principal Component Analysis (PCA) used in SRC. In the recognition modeling, the main difference between our method HSR and SRC is that $L_1$-norm minimization is replaced by $L_{1/2}$-norm minimization \cite{13} because $L_{1/2}$-norm minimization can produce sparser representation, which directly decreases the computational cost of sparse representation. Although $L_{1/2}$-norm minimization belongs to non-convex optimization problems, it can be easily transformed into a series of weighted $L_1$-norm minimization, which is convenient for us to solve by existing methods. Moreover, $L_{1/2}$-norm minimization is more robust than $L_1$-norm minimization, which is more suitable to process occluded face images. In our experiments, the new method has been verified on representative face databases (Extended Yale B, AR and FERET) with different conditions such as lighting, pose, expression and occlusion. In comparison with related work such as SRC \cite{7} and GSRC \cite{8}, experimental results demonstrated that our method is slightly complicated in structure, but it shows the great advantage for computational cost. And we also achieve approximate or even better recognition rate. Therefore, our method has a great potential for the application of fast face recognition like real-time surveillance.

The rest of paper is organized as follows. In Sect. 2, we briefly discuss previous work on ELM and sparse representation based on $L_1$ regularization. In Sect. 3, we describe our new method including hierarchical feature selection and $L_{1/2}$-regularized sparse representation. In Sect. 4, we report experimental results on SRC, GSRC and our method HSR under representative face databases with different conditions. Also we present discussions on the performance of new method. Finally, in Sect. 5, we show conclusions on our current research and indicate two important directions for future work.
2 Previous works

2.1 The structure of the original ELM

Extreme learning machine (ELM) was proposed by Huang et al. [14–16] for faster learning speed and higher generalization performance. The essence of ELM is that the parameters of the hidden nodes can be generated randomly without manually tuning [17]. Specifically speaking, the input data \( x \) are mapped to \( L \)-dimensional hidden layer and the network output is given by Eq. (1).

\[
f_L(x) = \sum_{j=1}^{L} \beta_j h_j(x)
\]

where \( \beta_j = [\beta_{j1}, \beta_{j2}, \ldots, \beta_{jm}]^T \) are the output weights between the hidden nodes and the output nodes, \( h_j(x) = g_j(x^T x_j + b_j) \) is the output of hidden layer, \( x_j \) is the input weight, \( b_j \) is the input bias and \( g_j(x) \) is the activation function, they all correspond with the output of the \( j \)-th hidden node. The ELM algorithm can be summarized as follows. Given \( N \) training samples \( \{x_i, t_i\}_{i=1}^{N} \), where input data \( x_i = [x_{i1}, x_{i2}, \ldots, x_{iN}]^T \) and the target labels \( t_i = [t_{i1}, t_{i2}, \ldots, t_{iN}]^T \). The input data are mapped to \( L \)-dimensional hidden layer initially. The structure of ELM can be determined if the output weights \( \beta \) can be calculated, so the following learning problems can be formulated by Eq. (2).

\[
H\beta = T
\]

where \( T = [t_1, t_2, \ldots, t_N]^T \) is the target matrix, \( H = [h(x_1), \ldots, h(x_N)] \) is the hidden layer output matrix, and \( h(x_i) = [h_1(x_i), h_2(x_i), \ldots, h_L(x_i)]^T \). So the output weights \( \beta \) can be calculated by Eq. (3).

\[
\beta = H^\dagger T
\]

where \( H^\dagger [15, 18] \) denotes the Moore–Penrose generalized inverse of matrix \( H \).

To make the resultant solution more stable and have better generalization performance [10], a positive value \( \|C \) as a regularization term can be added to the diagonal of \( H^T H \) shown Eq. (4).

\[
\beta = \left( \frac{1}{C} + H^T H \right)^{-1} H^T T
\]

2.2 Sparse representation based on \( L_1 \) regularization

Given training samples, \( A_i = [s_{i1}, s_{i2}, \ldots, s_{in}] \in \mathbb{R}^{m \times n_i} \) from all the \( i \)-th training samples, where \( s_{ij} (j = 1, 2, \ldots, n_i) \) is an \( m \)-dimensional vector, which belongs to the \( j \)-th sample of the \( i \)-th class. Denote by \( y_0 \in \mathbb{R}^m \) a test sample from the same \( i \)-th class. Intuitively, \( y_0 \) can be approximately represented by the linear combination of the training samples within \( A_i \).

\[
y_0 = \alpha_{i1}s_{i1} + \alpha_{i2}s_{i2} + \cdots + \alpha_{in}s_{in} = \sum_{j=1}^{n_i} \alpha_{ij} s_{ij}
\]

Suppose that the test sample \( y_1 \) is initially unknown of the exact class, a new matrix \( A \) is defined to concatenate the entire training samples of all \( k \) classes:

\[
A = [A_1, A_2, \ldots, A_k] = [s_{11}, s_{12}, \ldots, s_{kn}]
\]

Then, the linear representation of \( y_1 \) can be naturally written as Eq. (7).

\[
y_1 = Ax
\]

According to sparse coding via \( L_1 \)-norm minimization, the sparse coefficients \( x \) can be calculated as Eq. (8).

\[
x = \arg \min_x \left\{ \|y_1 - Ax\|_2^2 + \lambda \|x\|_1 \right\}
\]

In the case of occluded data, we should express the test sample \( y_2 \) as a sum of sparse representation and error. Then, the previous model [7] can be modified as Eq. (9).

\[
y_2 = Ax + e_0 = [A, A_e] \begin{bmatrix} x \\ \omega \end{bmatrix} = Bo
\]

where \( B \in \mathbb{R}^{m \times (n_i + n_e)} \), and the \( e_0 \in \mathbb{R}^m \) is a noise term with bounded energy \( \|e_0\|_2 < \varepsilon \). According to sparse coding via \( L_1 \)-norm minimization, the sparse coefficients \( \omega \) can be calculated as Eq. (10).

\[
\omega = \arg \min_{\omega} \left\{ \|y_2 - B\omega\|_2^2 + \lambda \|\omega\|_1 \right\}
\]

Therefore, we can represent test sample \( \{y_1, y_2\} \) as sparse coefficients \( \{\hat{x}_1, \omega_1\} \), which can be employed to identify the class of test sample.

3 \( L_{1/2} \)-regularized sparse representation using hierarchical feature selection

3.1 Framework

In this section, we briefly introduce the new method called \( L_{1/2} \)-regularized sparse representation using hierarchical feature selection (HSR). Our method will tackle with two critical issues in face recognition. First, how can we reduce computational cost of recognition modeling while keeping the recognition rate? Second, how can we ensure the robustness of our method to occluded faces? Our approach roughly consists of feature extraction and recognition.
modeling, which solves above problems accordingly. And also we provide the convincing reasons for the choice of recognition model and parameters. The structure of HSR is shown in Fig. 1.

For feature extraction, by employing hierarchical feature selection, we can compress the scale and dimension of global dictionary, which directly contributes to the decrease of computational cost in sparse representation. It consists of Gabor wavelets and ELM-AE. For Gabor wavelets’ part, according to theories of visual neuroscience, the mechanism of retina cells in human eyes can be simply simulated by Gabor wavelets, which could effectively extract local feature at multiple scales and orientations. And the local-feature-based methods are always less sensitive to variations of illumination, viewpoint and expression. Therefore, Gabor-feature-based images can improve recognition rate to some extent. Moreover, for the occluded images, the enormous scale of occlusion dictionary is the principle factor to affect the computational cost of sparse representation. Because of redundancies exist in Gabor-feature-based occlusion dictionary, the scale of Gabor-feature-based global dictionary can be compressed. For ELM-AE part, we hope to modify the basic ELM to represent input training and testing images meaningfully. Namely, the output weight of ELM-AE is responsible of learning the features from the input data via singular values. According to ELM theory, ELM-AE is a universal approximator that has a strong ability to achieve compressed, sparse and equal dimension representation. So it is reasonable to believe that images in a higher-dimensional input space can be effectively projected into a lower-dimensional feature space via ELM-AE. Thus, the dimension of Gabor-feature-based global dictionary can be compressed.

For recognition modeling, our approach is strongly rooted in the framework of sparse representation, which has showed its excellent performance on recognition rate especially for occluded faces. The testing image can be sparsely represented by the linear combination of the training samples, and our target is to balance the reconstructed error and the sparest coefficients via different kinds of regularization. In our approach, we choose $L_{1/2}$-norm minimization instead of $L_1$-norm minimization used in SRC or another regularized parameters because of two reasons. First, $L_1$ regularization locates between $L_0$ regularization and $L_\infty$ regularization; so $L_1$ regularization has sparse property, and it can be solved easily. Naturally thinking, $L_{1/2}$ regularization locates between $L_0$ regularization and $L_1$ regularization, so we expect that $L_{1/2}$ regularization has sparser property than $L_1$ regularization. Actually, the geometry property of $L_{1/2}$ and $L_1$ regularization has obviously proved our expectation. Second, Xu’s experiments [13] demonstrated that the performance of sparse representation using $L_{1/2}$ regularization is stronger than that using other $L_p$ regularization ($0 < p < 1/2$ or $1/2 < p < 1$). One might argue that $L_{1/2}$-norm minimization belongs to non-convex optimization problems, which means it is hard to solve. However, we can transform it into a series of weighted $L_1$-norm minimization, which is convenient for us to solve by existing methods. Moreover, according to Xu’s experiments, $L_{1/2}$-norm minimization is more robust than $L_1$-norm minimization, which is more suitable to process occluded faces.

3.2 Hierarchical feature selection

In sparse representation, the compression of global dictionary normally comes from the reduction of dimension
and scale (the number of elements), which directly contributes to the decrease of computational cost. For hierarchical feature selection, we employ Gabor wavelets and ELM-AE hierarchically. By employing Gabor wavelets, we can initially represent original images by Gabor-feature-based images, which can improve the recognition rate. By using ELM-AE, high-dimensional images can be rapidly represented by low-dimensional features; thus, we can compress the dimension of Gabor-feature-based global dictionary and testing images. Figure 2 depicts the process of hierarchical feature selection in detail. Each color block is regarded as a feature of an image case, and each column represents a case.

3.2.1 Gabor-feature-based image representation and occlusion dictionary

The motivation that we choose Gabor wavelets for image representation is mainly due to their biological relevance and computational properties. In this section, we will formulate how to represent original image via Gabor wavelets below. Then, in the presence of occluded images, we briefly introduce how to compress the scale of Gabor-feature-based occlusion dictionary via sparse coding.

Gabor wavelets [8] usually demonstrate good characteristics of spatial locality and orientation selectivity. Moreover, in the space and frequency domains, they are optimally localized. It can also be defined with the orientation \( \mu \) and scale \( m \) as follows:

\[
\psi_{\mu,m}(z) = \frac{|k_{\mu}|}{\sigma^2} e^{-||k_{\mu}||^2/(2\sigma^2)} \left[ e^{i k_{\mu} z} - e^{-\sigma^2/2} \right]
\]  

(11)

where the pixel of an image is \( z = (x, y) \), the wave vector is defined as \( k_{\mu} = k_{0} e^{i \phi_{\mu}} \) with \( k_{0} = k_{\max}/f \) and \( \phi_{\mu} = \pi \mu/8 \). \( k_{\max} \) is the maximum frequency, and \( f \) is the spacing factor between kernels in the frequency domain. Besides, \( \sigma \) determines the ratio of the Gaussian window width to wavelength. In most cases, Gabor wavelets have five different scales and eight orientations. As Liu and Wechsler’s work [19], the real part of Gabor wavelets can be shown in Fig. 3a.

Here we should also note that when the parameters of Gabor wavelets are as \( \sigma = 2\pi, k_{\max} = \pi/2, f = \sqrt{2} \), the Gabor wavelets demonstrate the excellent characteristics of spatial frequency, orientation selectivity and spatial locality.

According to above discussion, we can naturally represent high-dimensional images via Gabor wavelets. The Gabor-feature-based local representation is equal to the convolution of the input images with each Gabor wavelet. For example, the convolution of image \( \text{img}(z) \) with a Gabor wavelet is defined as below.

\[
G_{\mu,v}(z) = \text{img}(z) * \psi_{\mu,v}(z)
\]  

(12)

These convolution results show different scales, localities and orientations corresponding to the Gabor wavelets. As Liu and Wechsler’s work [19], the convolution results are all complex number. To contain a Gabor-feature-based image, we should first normalize all convolution results and then concatenate them to form an augmented feature vector \( \chi \).

\[
\chi = \left( G_{0,0}^T G_{0,1}^T \ldots G_{4,7}^T \right)^T
\]  

(13)

where \( \chi \) is an image based on Gabor-feature, which not only improves recognition rate but also bears to image local deformation to some degree. \( \chi \) can be shown in Fig. 3b.

To make a further step, we can derive Gabor-feature-based image representation in two situations. First, without any occlusion, the linear representation of \( y_1 \) can be
rewritten as \( y_1 = A z \). By employing Gabor wavelets, we can derive the Eq. (14).

\[
\chi(y_1) = X(A_1)x_1 + X(A_2)x_2 + \cdots + X(A_k)x_k = X(A)z
\]

(14)

where \( X(A) = [X(A_1), X(A_2), \ldots, X(A_k)] \) and \( X(A_i) = [\chi(s_{i1}), \ldots, \chi(s_{in})] \).

In the presence of occluded testing image \( y_2 \), the Eq. (14) should be modified as Ma’s work indicated as Eq. (15).

\[
\chi(y_2) = [X(A), X(A_e)] \begin{bmatrix} x \\ x_e \end{bmatrix} = X(B) \omega
\]

(15)

where \( X(A_e) \) is the Gabor-feature-based occlusion dictionary, and \( x_e \) is the representation coefficient vector of the input Gabor-feature vector \( \chi(y_2) \) over \( X(A_e) \).

The occlusion dictionary \( A_e \) in SRC is normally selected as the identity matrix \( I \) [7], and SRC has a large number of elements in occlusion dictionary, which definitely increases the computational cost of optimization. So we should compress it from two aspects including the dimension and scale. We hope to compress the scale of Gabor-feature-based occlusion dictionary because redundancies exist in it. For example, if the size of image is 83 \( \times \) 60, and the original occlusion dictionary is \( A_e \in \mathbb{R}^{4980 \times 4980} \). Then, Gabor-feature-based occlusion matrix \( X(A_e) \in \mathbb{R}^{199200 \times 4980} \), where we set \( \mu = \{0, \ldots, 7\} \) and \( v = \{0, \ldots, 4\} \). Figure 4 represents the eigenvalues of \( X(A_e) \), and we can see that a few eigenvectors (140) of \( X(A_e) \) have significant eigenvalues, which implies that \( X(A_e) \) is obviously redundant. Therefore, We suppose \( z = X(A_e) = [z_1, \ldots, z_n] \in \mathbb{R}^{m_v \times n_v} \) Gabor-feature-based occlusion dictionary; then, the scale-compressed occlusion dictionary is denoted by \( \Gamma = [d_1, \ldots, d_p] \in \mathbb{R}^{m_v \times p} (p \ll n_v) \), and we can represent \( Z \) by \( \Gamma \) via sparse coding. So our objective function is defined as Eq. (16).

\[
J_{F,A} = \arg \min_{Z} \{ \frac{1}{2} A_z^T + \frac{1}{2} A_t \} \quad \text{s.t.} \quad d_i d_j = 1 \forall j
\]

(16)

It is easy to solve this optimization problem by optimizing \( \Gamma \) and \( A \) alternatively. Therefore, the compression of scale is easily achieved. The Eq. (15) can be modified by Eq. (17).

\[
\chi(y_2) = [X(A), \Gamma] \begin{bmatrix} x \\ x_{F} \end{bmatrix}
\]

(17)

3.2.2 ELM-AE for high-dimensional image representation

The motivation that we choose ELM-AE for image representation is due to its representation ability and computational cost. We will mainly introduce ELM-AE for high-dimensional image representation below. And then we will briefly verify the performance of ELM-AE.

For auto-encoder [20], the output data \( \hat{x} \) are similar to the input data \( x \). Some interesting structure can be obtained when constraints are placing on the networks. Based on the above concept, the ELM-AE was first proposed by Cambria et al. [10], and the main objective of ELM-AE is to represent the input data meaningfully and rapidly. In terms of the number of the input nodes and hidden nodes, there are three different representation including compressed representation, sparse representation and equal dimension representation. For face recognition task, we hope to represent input training and testing images by compressed representation. The training structure of ELM-AE can be seen as Fig. 5a.

To be more specific, we first modify the basic ELM [15, 21] to conduct unsupervised learning (\( f = x \)), and random weights and biases of the hidden nodes are chosen to
orthogonal because orthogonalization will make the generalization of ELM-AE better.

The orthogonal random weight and bias can be calculated by Eq. (18).

\[ a^T a = I, \quad b^T b = 1 \]  \hspace{1cm} (18)

where \( a = [a_1, \ldots, a_L] \) is the orthogonal random weight and \( b = [b_1, \ldots, b_L] \) is the orthogonal random bias between the input nodes and hidden nodes. Then, we calculate the hidden layer output matrix \( H \) as the original ELM does.

\[ H = g(ax + b) \]  \hspace{1cm} (19)

After training process, we first hypothesize that the output weights of ELM can be treated as coding parameters of auto-encoder, which is responsible for learning the low-dimensional features from the high-dimensional data. And the output weight \( \beta \) as Eq. (20) modified by Eq. (4).

\[ \beta = \left( I + H^T H \right)^{-1} H^T X \]  \hspace{1cm} (20)

where \( H = [h_1, \ldots, h_N] \) is the output of hidden layer and \( X = [x_1, \ldots, x_N] \) is the input data.

Therefore, the trained ELM-AE will be employed to conduct high-dimensional image representation, for example, the original image is initially represented via Gabor wavelets, and the dimension of Gabor-feature-based image will naturally increase. Then, we represent Gabor-feature-based image \( \chi \) into low-dimensional features \( \chi_r \) via ELM-AE by Eq. (21).

\[ \chi_r = \chi \beta \]  \hspace{1cm} (21)
After hierarchical feature selection, the image representation \( y \) can be visualized in Fig. 5b, which will be sent into the framework of sparse representation for further processing.

It shows that ELM-AE itself can speed up the process of feature extraction. Moreover, in HSR, the ELM-AE can effectively reduce the dimension of global dictionary and testing images, which greatly relieve the computational burden undertaken by optimization methods.

### 3.3 \( L_{1/2} \)-regularized sparse representation

In this section, we first introduce generic framework of sparse representation and then compare different regularized parameters such as \( L_{1/2} \), \( L_1 \) and \( L_2 \). Finally, we decide to employ \( L_{1/2} \) because it can produce sparser and more robust representation compared to \( L_1 \), which speeds up the face recognition.

#### 3.3.1 Generic framework of sparse representation for face recognition

Supposing that the number of well-aligned training face images of each class is fixed, we collect training images together forming a large training dictionary and each column is normalized via \( L_2 \)-norm. One classical assumption is that a new image of \( i \)th class can be well represented as a linear combination of all the \( i \)th class training samples. So the linear representation of the test image \( y_1 \) can be written as Eq. (22).

\[
y_1 = Ax \in \mathbb{R}^m
\]

where \( x \) is a vector of sparse coefficients. In practical situations, the coefficient vector is often complicated when the test image \( y_2 \) is occluded partially. The linear model should be modified as Eq. (23).

\[
y_2 = Ax + e_0 = [A, A_c] \begin{bmatrix} x \\ a_x \end{bmatrix} = B \omega
\]

where \( e_0 \in \mathbb{R}^m \) is a vector of error, \( B = [A, A_c] \) and \( \omega = [x, a_x]^T \); thus, face recognition with occlusion can be represented as the coefficients. In order to introduce more general framework of sparse representation, we choose general loss function \( L \) for balanced constructed error and an uncertain norm for regularized parameters. Therefore, the coefficients can be represented as Eq. (24).

\[
\hat{\omega}_k = \arg \min_\omega \{ L(y_2, B \omega) + \lambda \omega_k \}
\]

where \( || \cdot ||_k \) denotes \( k \)-norm that represents the uncertain parameter. Therefore, we can represent test sample \( y_2 \) using sparse coefficients \( \hat{\omega}_k \). The general framework can almost explain all special cases. For example, if general loss function \( L \) can be denoted as square loss, then the framework can be converted into AIC [22] and BIC [23] criteria when \( k = 0 \); the framework can be converted into Lasso algorithm [24] when \( k = 1 \); and the framework can be converted into ridge regression [25] when \( k = 2 \). In our approach, we choose traditional square loss function, and the next section will discuss the choice of regularized parameters.

#### 3.3.2 Regularized parameters: \( L_0 \), \( L_{1/2} \) and \( L_1 \)

The motivation why we choose \( L_{1/2} \)-norm minimization is due to two aspects.

First, although the coefficients can be obtained via \( L_0 \)-norm minimization, solving \( L_0 \)-norm minimization is a NP-hard problem. Therefore, Tibshirani introduced the Lasso algorithm (\( L_1 \)-norm minimization) to obtain relative sparse coefficients, which is much easier for us to solve because \( L_1 \)-norm minimization belongs to the convex problem [26]. What’s more, they proved that \( L_0 \) regularization is equal to \( L_1 \) regularization on the certain constraint condition.

However, in the practical application, we find that \( L_1 \)-norm minimization usually cannot produce the sparsest solution, so a question is raised that whether we can introduce a new regularized parameter, which provides a sparser solution than \( L_1 \) regularization. Fortunately, Xu’s experimental results proved that \( L_{1/2} \) regularization can produce sparser representation compared with \( L_1 \) regularization, which is also proved by their geometry property. From the Fig. 6, the bound of \( L_p \) regularization has different shape, and the solution of \( L_p \)-norm minimization is equal to the intersection of the bound and the loss function. For example, we can clearly see that solution via \( L_2 \) is not sparse at all and the solution via \( L_{1/2} \) is sparser than that via \( L_1 \) because the bound of \( L_{1/2} \) is easier to intersect with loss function at coordinates.

Although \( L_{1/2} \)-norm minimization belongs to non-convex optimization problems, we can transform it into a series of weighted \( L_1 \)-norm minimization, which is also convenient for us to solve by existing methods. Moreover, according to Xu’s experiments, \( L_{1/2} \)-norm minimization is more robust than \( L_1 \)-norm minimization, which is more suitable to process occlusion in face images.

**Fig. 6** Possibility of sparse solution via \( L_1 \), \( L_2 \) and \( L_{1/2} \)
Second, although we initially want to explore other possibilities such as $L_p$-norm minimization ($0 < p < 1/2$ or $1/2 < p < 1$), Xu’s experiments [13] clearly demonstrated that the performance of sparse representation using $L_{1/2}$ regularization is stronger than that using other $L_p$ regularization ($0 < p < 1/2$ or $1/2 < p < 1$). Therefore, $L_{1/2}$ regularization can completely replace $L_p$ regularization ($0 < p < 1$). $L_{1/2}$-norm algorithm can be fulfilled as follows.

**Algorithm 1** The $L_{1/2}$-norm algorithm (Input: $\beta^0$; Output: $\beta^t$)

1. Let $t = 0$, set the maximum number of iterations $K$ ($K$ is condition of the algorithm’s termination), initialize $\beta^0 = (1,1, ..., 1)^T$.
2. Calculate the equation $\beta^{t+1} = \arg\min \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{i=1}^{p} |\beta_i|/\sqrt{|\beta_i^2|} \right\}$, and then let $t = t + 1$.
3. When $t < K$, goto 2. And when $t = K$, output $\beta^t$.

**Algorithm 2** The HSR algorithm (Input: $y_2$; Output: identity($y_2$))

1. Images representation: $y_2$ is the input of HSR algorithm (occluded face), which is a special case of normal face. The columns of $A$ and $A_e$ are normalized to have unit $L_2$-norm.

   $$y_2 = Aa + e_0 = [A, A_e] [\alpha^T] = B\omega$$

   Where $y_2 \in \mathbb{R}^{n \times 1}$, $B \in \mathbb{R}^{m \times (n_A + n_{A_e})}$.

2. Gabor wavelets: extract local feature to enhance recognition rate and compress the scale of Gabor feature based occlusion dictionary via sparse coding.

   $$\chi(y_2) = [\chi(A), \chi(A_e)] [\alpha^T] = B_x \omega$$

   $$\chi(y_2) = [\chi(A), I] [\alpha^T] = B_r \omega_r$$

   Where $\chi(y_2) \in \mathbb{R}^m \times 1$, $B_x \in \mathbb{R}^{m \times (n_A + n_{A_e})}$, $B_r \in \mathbb{R}^{m \times (n_A + n_r)}$, and $n_r \ll n_{A_e}$.

3. ELM-EA: compress the dimension of global dictionary and testing image

   $$\chi_r(y_2) = [\chi_r(A), I_r] [\alpha^T] = B_r^c \omega_r$$

   Where $\chi_r(y_2) \in \mathbb{R}^{m_r \times 1}$, $B_r^c \in \mathbb{R}^{m_r \times (n_A + n_r)}$, and $m_r \ll m_x$.

4. Solve the $L_{1/2}$-norm minimization problem: get a sparser and more robust representation.

   $$\delta_{1/2} = \arg\min_\omega \left\{ \|\chi_r(y_2) - B_r^c \omega_r\|^2_2 + \lambda \|\omega_r\|_{1/2}^{1/2} \right\}$$

   Where

   $$\delta_{1/2} = [\delta_{1/2}, \delta_{1/2}]$$

   $$B_r^c = [\chi_r(A), I_r]$$

   And $\lambda$ is a positive scalar number that balances the reconstructed error and sparse coefficients.

5. Compute the residuals

   $$r_i(\chi_r(y_2)) = \|\chi_r(y_2) - I_r \delta_{1/2} - \chi_r(A) \delta_i(\delta_{1/2})\|^2_2$$

   for $i = 1, 2, ..., k$.

   Where $\delta_i(\cdot): \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the characteristic function which selects the coefficients associated with the $i^{th}$ class.

6. Output that identity($y_2$) = identity($\chi_r(y_2)$) = arg $\min_i r_i(\chi_r(y_2))$
Overall, we naturally introduce $L_{1/2}$-regularized sparse representation for fast face recognition using hierarchical feature selection (HSR). Our method has solved two critical issues raised in the Sect. 3.1. First, we employ Gabor wavelets and ELM-AE hierarchically in order to reduce the dimension and scale of global dictionary, which accordingly reduces the computational cost of sparse representation. Second, $L_{1/2}$-norm minimization can produce sparser and more robust representation than $L_1$-norm minimization, which not only reduces the computational cost of optimization but also is more suitable to process occlusion in face images. The algorithm of HSR and its explanation are given below.

4 Experimental results

In this section, we present some experimental results on available benchmark databases to compare the performance of the proposed algorithm HSR with GSRC and SRC. We do not compare HSR to deep learning algorithms because the computational cost of deep learning is very expensive [27]. To evaluate the performance of HSR comprehensively, this section is divided into two detailed sections. In Sect. 4.1, we first tested our method on the face datasets without occlusion. And then in Sect. 4.2, we tested the new method on the face datasets against occlusion using two different frameworks (no partition and partition). All the simulations for the HSR, GSRC and SRC algorithms are carried out in MATLAB 7.8 environment running in an Intel Xeon E5-1650 3.20 GHz CPU. In the experiments of Gabor wavelets, the parameters are set as $k_{\text{max}} = \pi/2$, $f = \sqrt{2}$, $\sigma = \pi$, eight orientations $\mu = \{0, \ldots, 7\}$ and five different scales $\upsilon = \{0, \ldots, 4\}$ by our experience. And the parameters are fixed for all the experiments below. The activation function of ELM is set to ‘sig’ representing the sigmoidal function, the parameter $C$ is set to 100 and the number of the hidden neurons is equal to the compressive feature space dimension. In addition, all face images provided in the databases are cropped and aligned by the location of eyes. The face images from the databases are further normalized to zero mean and unit variance.

4.1 Face recognition without occlusion

We compared the performance of the HSR with two classical algorithms SRC and GSRC on three typical facial image databases: Extended Yale B [28], AR [29] and FERET [30]. For the Extended Yale B and AR databases, we compared the performance of HSR, SRC and GSRC versus feature dimension. Moreover, we compared the performance of different compression methods (ELM-AE and PCA) and different regularized parameters ($L_1$ and $L_{1/2}$) on two databases. For the FERET, we compared the performance of HSR, SRC and GSRC versus pose angle.

1. Extended Yale B Database: The database consists of 2414 frontal face images of 38 individuals. The images are normalized to 9684 under various laboratory-controlled lighting conditions. We randomly selected half of the images for training (i.e., 32 images per subject), and the other half for testing. Choosing the training set randomly assures that our results will be independent of any special choice. Figure 7 shows some samples from the same object class, and it is obvious that only illumination is added to these images. The dimension of the Gabor-feature-based image is 199200 ($83 \times 60 \times 40$) through a set of Gabor wavelets, which includes five different scales and eight orientations. They can capture abundant local features to form Gabor-feature-based image, which will take a lot of time to process this high-dimensional image. To compress the feature space, we applied ELM-AE (a part of HSR) and PCA (a part of SRC and GSRC), respectively, with the feature dimensions 30, 56, 120, 224 and 504 on the Gabor-feature-based images. Then, we computed the recognition rate and the computational cost. In addition, the computational cost of sparse representation is equal to the testing time because there is no training process. In our experiments, we set $\lambda = 0.001$ [7] in HSR, GSRC and SRC by our experience. It shows the recognition rates in Fig. 8a and computational cost in Fig. 8b of HSR comparing with GSRC and SRC versus the feature dimension. It is turned out that with the increase of feature dimension, the recognition rate becomes higher and the computational cost becomes more. HSR achieves a maximum recognition rate of 98.52 % with 504D feature space. In contrast, the maximum recognition rate of GSRC is 98.44 % and SRC is 97.50 %. The computational cost of SRC and new method is similar, which much less than that of GSRC. According to a specific dimension 504D, the computational cost of compression by PCA is 36.2022s, while ELM-AE is 1.3011s.

![Fig. 7 Samples from the same object of Extended Yale B dataset](image-url)
2. AR Database: The AR database consists of 4000 frontal images from 126 individuals. We chose a subset consisting of 50 male subjects and 50 female subjects. For each subject, 14 images are selected, which includes only illumination changes and expressions. Figure 9 shows several samples from the same object class with the variation of expression and illumination. We selected seven images from Session 1 for training and seven images from Session 2 for testing. The images were cropped and converted to gray scale with the size is $83 \times 960$. The dimension of the Gabor-feature vector is 12000 after a set of Gabor wavelets. Then, we continued to reduce the feature space with five dimensions: 30, 54, 130, 311 and 540. We also set $\lambda = 0.001$ in HSR, GSRC and SRC, like on the Extended Yale B database. It shows the recognition rates in Fig. 10a and the computational cost in Fig. 10b of HSR comparing with GSRC and SRC versus the feature dimension. On this database, the maximum recognition rate of HSR, GSRC and SRC is 95.86, 95.86 and 93.57 %, respectively. The computational cost of HSR is much less than that of the SRC and GSRC versus feature dimension especially for computational cost.

For briefly verifying the compression performance of ELM-AE, we selected PCA as a control group for high-dimensional image representation. Only the computational cost of the compression component was taken into consideration in our experiments. We compared the computational cost of ELM-AE and PCA on the Extended Yale B and the AR database before the process of recognition modeling. The testing sets of the Extended Yale B and the AR are compressed to a certain dimension (405D for the Extended Yale B and 450D for the AR), whose computational costs are list on the follow Table 1.

We also conducted a quantitative experiment using different compression methods (ELM-AE and PCA) and different regularized parameters ($L_1$-norm and $L_{1/2}$-norm) on two databases (Table 2). For one thing, when testing sets are compressed to a certain dimension (405D for the Extended Yale B and 450D for the AR), we demonstrated that $L_{1/2}$-norm minimization is superior to the $L_1$-norm minimization for computational cost while keeping the approximate recognition rate. For another thing, ELM-AE and PCA are used to compress the Gabor-feature-based images in different mechanisms. So under the same regularized parameter, the computational cost of methods using ELM-AE is less.

3. FERET pose database: This database includes 1400 images from 200 subjects (7 images per subject).
Among the 1400 images, 600 images are the frontal face with illumination and facial expressions and the others are the face variation with different pose angles. The images marked with ‘ba’, ‘bd’, ‘be’, ‘bf’, ‘bg’, ‘bj’ and ‘bk’ stand for the different illumination, facial expressions and pose angles (Fig. 11). In our experiments, the images of this database were already cropped to the size of 80x80. In order to examine the robustness of HSR comparing with the other original algorithms, we tested the recognition rates and computational cost with respect to the variable pose angle. Then, in the first test, we used images marked with ‘ba’ and ‘bj’ for training, and images marked with ‘bk’ for testing. In another four tests, images marked with ‘ba’, ‘bj’ and ‘bk’ were used as training set, and the rest of images were, respectively, used as testing set. After feature extraction, the dimension was fixed on 350D in above three methods. We set the parameters $k = 0.005$ for HSR and GSRC and $k = 0.05$ for SRC, which will conduct the best results. The results showed a growing trend of the recognition rate with less pose angle variability in Fig. 12a. When the pose angle becomes larger, the recognition rate of HSR is almost 40% higher than the nearest competitor but still poor. Besides, the computational cost of HSR and GSRC is much less than that of SRC in Fig. 12b. The above experiment illustrated the performance of HSR is also much better than that of the SRC and GSRC versus pose angle.

4. Parameter values

All the parameters for the experiments are fixed to the optimal values by our trials and experience from other papers. In this section, we list some experiments on different parameters based on the AR database to demonstrate our choice for parameters. First, we evaluate the activation function of the ELM-AE with the 300D dimension of dictionary and $k = 0.001$. Second, we evaluate the regularization parameter $\lambda$ in the case of other fixed parameters. We assume that the dimension of dictionary is set to 300D and 540D.

From the Table 3, sigmoid function performs the best recognition rate in reasonable computational cost. From the Table 4, recognition rate is inversely proportional to regularization parameter, so is the computational cost. In other words, the less the value of $\lambda$ is, the better recognition rate can be achieved. However, computational cost increases at the same time. We should compromise the performance between recognition rate and computational cost. Therefore, the middle value $\lambda = 0.001$ performs a good recognition rate in reasonable computational cost.

4.2 Face recognition with occlusion

In this section, we also compared the performance of the HSR with SRC and GSRC on a subset of AR dataset, which includes occluded images. The chosen subset consists of 1300 images from 100 subjects (50 male and
In this subset, 700 images of unoccluded frontal face with expression and illumination variation were used for training set. Besides, the rest data were split into two separate test sets of equal size. The first test set contains 300 images, on which all the 100 subjects are wearing sunglasses. The second test set also contains 300 images, and all the subjects wear scarves instead. Sunglass occludes about 20% of the image and scarf occludes about 40% of the image intuitively (Fig. 13).

The parameters of HSR and GSRC were set to $\lambda = 0.0005$ and SRC used $\lambda = 0.005$, which will conduct the best results. The images of this dataset were resized to $83 \times 60$; then, the size of global dictionary is 4980 x 5680 in the original SRC. In the case of the proposed HSR and GSRC, the dimension of Gabor-feature-based image is 199200 ($83 \times 60 \times 40$) and then decreases to 5600D by ELM-AE. Meanwhile, the scale of Gabor-feature-based occlusion dictionary is compressed to 100 by sparse coding. As a result, after hierarchical feature selection, the size of global dictionary is 5600 x 800. Table 5 has shown the experimental results on two testing sets implemented by SRC, GSRC and HSR. Apparently, SRC performs the worst recognition rate and the highest computational cost; in other words, the holistic features used in SRC are not suitable for the occluded images and the scale of global dictionary decides the computational cost to some degree. Besides, it is clearly seen that the computational cost of HSR is much less than that of GSRC in two datasets, while the recognition rate of HSR is higher than that of GSRC in AR scarves.
We quoted the approach in Wright et al. [7] to partition the whole image into blocks and processed each block independently, assuming the occlusion part is contiguous. In these blocks, some of them are assumed to be completely occluded and some of them may be partially occluded. We calculated the performance of each block using the HSR, which naturally determined the performance of the whole image by voting. In our experiments, the image is divided into 8 \((4 \times 2)\) blocks and rescaled to the size of a small \((21 \times 30)\) for AR database pixel patch. In each block, after hierarchical feature selection, the dimension of Gabor-feature-based image is 800, and the scale of Gabor-feature-based occlusion dictionary is fixed on 20. Thus, the global dictionary in SRC is \(630 \times 1330\), while the global dictionary of HSR and GSRC is \(800 \times 720\). Table 6 illustrates the recognition rate and computational cost with the partition approach. The HSR with the partition achieves \(99.33\%\) in the case of sunglasses testing set and \(99.00\%\) in the case of scarves testing set with the least computational cost.

For comparing difference between partitioned and non-partitioned approaches, we have visualized all results from Tables 5 and 6 into Figs. 14 and 15.

We can clearly see that using partitioning method, the recognition rates generally increase, while the computational costs generally decrease except for GSRC. We believe that the number of subblocks which make wrong classifications is normally less than the number of subblocks that are correctly classified, which can ensure the final recognition rate. What’s more, because of partitioning, the dimension and scale of occlusion dictionary are accordingly decreased, which in turn reduce the computational cost of sparse representation.

### Table 5 Performance of non-partitioning methods (SRC, GSRC and HSR)

| Testing set methods | AR sunglasses | AR scarves |
|---------------------|---------------|------------|
|                     | SRC           | GSRC       | HSR        | SRC           | GSRC       | HSR        |
| Rec. rate (%)       | 79.69         | 94.33      | 85.67      | 49.00         | 90.67      | 93.67      |
| Time (s)            | 13,402.00     | 4788.30    | 1826.90    | 137,610.00    | 4782.00    | 1935.10    |

### Table 6 Performance of partitioning methods [SRC (p), GSRC (p) and HSR (p)]

| Test set methods   | AR sunglasses | AR scarves |
|--------------------|---------------|------------|
|                    | SRC (p)       | GSRC (p)   | HSR (p)    | SRC (p)       | GSRC (p)   | HSR (p)    |
| Rec. rate (%)      | 96.00         | 99.67      | 99.33      | 93.67         | 98.67      | 99.00      |
| Time (s)           | 4686.80       | 12,310.00  | 1610.60    | 4693.40       | 12,259.00  | 1617.00    |
5 Conclusions

In this paper, we proposed a novel method for fast face recognition called $L_{1/2}$-regularized sparse representation using hierarchical feature selection (HSR). By employing hierarchical feature selection, we can extract the local features from image, which improves recognition rate because local features are less sensitive to the facial variation. More importantly, the global dictionary can be easily compressed in the dimension and scale by hierarchical feature selection, which speeds up the computation of sparse representation. To be more specific, it is feasible to compress the scale of Gabor-feature-based occlusion dictionary via sparse coding. And high-dimensional images and global dictionary can be rapidly compressed into low-dimensional feature space via ELM-AE. By introducing $L_{1/2}$-regularized sparse representation, our method can produce sparser representation than $L_1$-regularized SRC, which in turn speeds up the face recognition. Besides, our method can also produce more robust representation than $L_1$-regularized SRC, which is more suitable to identify occluded faces such as AR sunglasses and scarves. We evaluated our method on a variety of face databases. Experimental results have demonstrated the great advantage of our method for computational cost in comparison with SRC and GSRC. Besides, we also achieve approximate or even better recognition rate. Therefore, our method has a great potential for the application of fast face recognition like real-time surveillance. Our future work will focus on two aspects. First, we will extend ELM-AE into Multi-Layer ELM-AE, which may extract more representative features in order to improve the recognition rate. Second, we will optimize the $L_{1/2}$ regularization algorithm in order to reduce the computational cost further.

Acknowledgments The authors would like to thank Dr Jun Zhou at Griffith University for helpful and excellent discussions and comments. This work is partially supported by Natural Science Foundation of China (41176076, 31202036 and 51075377).

References

1. Zhao WY, Chellappar, Phillips PJ, Rosenfeld A (2003) Face recognition: a literature survey. ACM Comput Surv 35:399–459
2. Turk M, Pentland A (1991) Eigenfaces for recognition. J. Cogn Neurosci 3:71–86
3. Belhumeur PN, Hespahna JP, Kriengman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE PAMI 19:711–720
4. Barkan O et al (2013) Fast high dimensional vector multiplication face recognition. In: IEEE international conference on computer vision (ICCV), 2013
5. Jiang Xudong (2011) Linear subspace learning-based dimensionality reduction. Sig Process Mag IEEE 28(2):16–26
6. Freifeld O, Black MJ (2012) Lie bodies: a manifold representation of 3D human shape. Computer vision—ECCV 2012. Springer Berlin Heidelberg, pp 1–14
7. Wright J, Yang AY, Ganesh A et al (2009) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31(2):210–227
8. Yang M, Zhang L (2010) Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary. Computer vision—ECCV 2010. Springer Berlin Heidelberg, pp 448–461
9. Sharma BP, Rajesh R, Rajesh M (2013) Face recognition using Gabor wavelet for image processing applications. In: Proc. of Int. Conf. on Emerging Trends in Engineering and Technology. http://searchdl.org/public/book_series/AETS/3/98.pdf
10. Cambria E et al (2013) Extreme learning machines. IEEE Intell Syst 28(6):30–59
11. Gupta HA, Raju A, Alwan A (2013) The effect of non-linear dimension reduction on Gabor filter bank feature space. J Acoust Soc Am 134(5):4069
12. Cao Juwen, Lin Zhiping (2014) Bayesian signal detection with compressed measurements. Inf Sci 289:241–253
13. Xu ZB (2010) Data modeling: visual psychology approach and L1/2 regularization theory. In: Proceedings of international congress of Mathematicians, pp 3151–3184
14. Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing 70(1):489–501
15. Huang GB, Zhu QY, Siew CK (2004) Extreme learning machine: a new learning scheme of feedforward neural networks. In:
Proceedings of IEEE international joint conference on neural networks, vol 2, pp 985–990

16. Cao J, Lianglin X (2014) Protein sequence classification with improved extreme learning machine algorithms. BioMed Res Int 2014:1–12. Article ID 103054

17. Huang GB, Zhou H, Ding X (2012) Extreme learning machine for regression and multiclass classification. IEEE Trans Syst Man Cybern Part B Cybern 42(2):513–529

18. Noble B (1975) Methods for computing the Moore–Penrose generalized inverse, and related matters. In: Nashed MZ (ed) Generalized Inverses and Applications. Academic, New York, pp 245–301. http://www.univie.ac.at/nuhag-php/bibtex/open_files/no76_B.%20Noble0001.pdf

19. Liu C, Wechsler H (2002) Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE IP 11:467–476

20. Baldi P (2012) Autoencoders, unsupervised learning, and deep architectures. ICML Unsupervised and Transfer Learning, JMLR: Workshop and Conference Proceedings, vol 27, pp 37–50. http://www.mtome.com/Publications/CiML/CiML-v7-book.pdf#page=53

21. Huang GB, Chen L (2007) Convex incremental extreme learning machine. Neurocomputing 70(16):3056–3062

22. Akaike H (1987) Factor analysis and AIC. Psychometrika 52(3):317–332

23. Burnham KP, Anderson DR (2004) Multimodel inference understanding AIC and BIC in model selection. Sociol Methods Res 33(2):261–304

24. Zhao P, Yu B (2006) On model selection consistency of Lasso. J Mach Learn Res 7:2541–2563

25. Ogutu JO, Torben S-S, Hans-Peter P (2012) Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In: BMC proceedings, vol 6, no Suppl 2, BioMed Central Ltd

26. Kim J, Koh K, Kusting M et al (2007) A method for large-scale 11-regularized least squares. IEEE J Sel Top Signal Process 1:606–617

27. Bengio Y (2009) Learning deep architectures for AI. Found Trends® Mach Learn 2(1):1–127

28. Lee K, Ho J, Kriegman D (2005) Acquiring linear subspaces for face recognition under variable lighting. IEEE PAMI 27:684–698

29. Martinez A, Benavente R (1998) The AR face database. CVC Technical Report no. 24

30. Phillips PJ, Moon H, Rizvi SA, Rauss P (2000) The FERET evaluation methodology for face recognition algorithms. IEEE PAMI 22:1090–1104