A JOINT CLASSIFICATION APPROACH VIA SPARSE REPRESENTATION FOR FACE RECOGNITION

Yandong Wen, Youjun Xiang, Yuli Fu
School of Electronic & Information Engineering, South China University of Technology, P. R. China
wen.yandong@mail.scut.edu.cn, *yjxiang*@scut.edu.cn, *fuyuli*@scut.edu.cn

ABSTRACT

We consider the problem of automatically recognizing human faces in which sparse representation-based classification (SRC) offers a key. SRC includes two steps: seeking sparest solution and making decision by dictionary classifier (DC). Aiming at improving the performance of face recognition, this paper proposes a joint classification approach based on sparse representation. We initialize dictionary with part of the training samples and train a linear classifier (LC) with the remaining. Thus, the joint classifier (JC), which combines the DC and LC, can decide which subject the query image belongs to. To validate the joint classifier, a residual-based evaluating criterion is established to measure the classification reliability for two classifiers. Experimental results verify that the proposed joint classification strategy significantly improves recognition accuracy at the cost of affordable computational complexity.

Index Terms—Sparse Representation Classification, Face Recognition, Simplified Training, Residual-based criterion, Joint Classification

1. INTRODUCTION

Recent years have witnessed a significant progress in pattern recognition (PR) since sparse representation-based classification (SRC) [1] framework was reported. In SRC, a query sample is sparsely coded over a dictionary that is constructed by training samples including all object classes, and the label of query sample will be decided by finding the corresponding class with minimal reconstruction error. In general, SRC could be summarized as two steps: (1) Seeking sparest solution among the whole dictionary. (2) Making decision by dictionary classifier (DC). J. Wright et al.[1] had successfully applied SRC to face recognition (FR) and achieved promising results. It performs better than traditional FR techniques like nearest neighbor (NN)[2], nearest subspace (NS)[3] and support vector machine (SVM)[4].

SRC provides a new framework of classification for researchers and it also has been widely studied. In SRC framework, the dictionary is formed by all training images. Each column of the dictionary, also called atom, is stacked by a training image. Therefore, such huge size of dictionary makes it computationally expensive to find the sparsest solution, disabling SRC to work in real-time situations. Generally speaking, many feature extraction techniques can be used for getting a downsized dictionary which is constituted by samples with less features, such as Eigenfaces[5], Fisherfaces[6], random face[1] and so on. Duc-Son Pham et al proposed a joint representation and classification framework that included finding the most discriminative sparse vector and optimal classifier parameters [9]. They alternately updated dictionary, classifier and sparse coefficients to achieve optimization in each iteration. Qiang Zhang et al added discriminative ability in K-SVD[7] and proposed discriminative K-SVD (D-KSVD)[8]. Via dictionary learning, D-KSVD uses a relatively small-scale dictionary (fewer atoms) for classification without losing recognition accuracy and skips the step of manually selecting training set. In other words, a learned dictionary would achieve higher recognition accuracy compared to the original one (consist of raw atoms) with the same number of atoms. After D-KSVD is applied, the representation power of each atom is enhanced but its own label is discarded. In this way, DC is disabled and discriminative information is carried by sparse coefficients. The label of test sample could be obtained by a trained LC.

In spite of achieving better performance in [8], D-KSVD does exists certain drawbacks. It optimizes objective function simultaneously [8], but the convergence does not accelerate much. When the training set is changed, it costs lots of time to train a new dictionary. On the other hand, as the dimension of samples is reducing, the improvement of dictionary representation power has its own limit, which harms the performance of the subsequent classification.

Inspired by [9], we proposed an approach called sparse representation-based joint classification (SRJC), including a simplified approach to train LC for sparse vector, an evaluating criterion for DC and LC, and a joint classification strategy. Concretely, the LC is addressed in the scenario where DC performs poorly. In this way, the LC is trained with poor samples that are selected according to proposed evaluating criterion by a two-step algorithm, which significantly improves classification accuracy and greatly reduces training time.

The remainder of this paper is as follows. Section 2 briefly introduces SRC framework. In section 3, our pro-

posed method is presented and more details about LC will be discussed. In section 4, we present the experimental results which demonstrate improvement on Extend Yale B and AR databases, followed by concluding remarks in section 5.

2. SPARSE REPRESENTATION-BASED CLASSIFICATION

Given a $w \times h$ grayscale image, it was stacked into a vector $d_{i,j} \in \mathbb{R}^m$ ($m = w \cdot h$). We assume that training samples include k distinct classes. $D_i = [d_{i,1}, d_{i,2}, \ldots, d_{i,n_i}] \in \mathbb{R}^{m \times n_i}$ denotes the n_i training samples in i th class, where D_i is also called i th sub-dictionary and $d_{i,j}$ is a training sample from i th class. We arrange k sub-dictionaries as a dictionary with the form of $D = [D_1, D_2, \ldots, D_k] \in \mathbb{R}^{m \times n}$. In this way, a test sample $y \in \mathbb{R}^m$ belonging to r th class is represented as a linear combination of atoms in D . i.e. $y = \sum_{i=1}^{k} \sum_{j=1}^{n} x_{i,j} d_{i,j} = \sum_{i=1}^{k} D_i x_i$, where $x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,n_i}]^T \in \mathbb{R}^{n_i \times 1}$ is the coefficients corresponding to i th subject. In the ideal case, the coefficient vector should be $x = [0, \ldots, 0, x_{r,1}, x_{r,2}, \ldots, x_{r,n_r}, 0, \ldots, 0]^T \in \mathbb{R}^n$ whose entries are zero except those associated with the r th class. It means that y can be approximately represented by several atoms in the subject that y belongs to. Therefore, the key to object recognition is seeking a sparsest solution over D . We can formulate classification task as follow.

$$\hat{x}_0 = \arg\min_x \|x\|_0 \quad s.t.\ Dx = y \tag{1}$$

where $\|\cdot\|_0$ denotes the l_0 -norm, counting the number of non-zero entries of a vector. Since finding sparest solution of equation (1) is NP-hard[10], it is impractical to solve the problem as atoms increasing. Luckily, recent work in theory of compressed sensing[11-13] reveals that l_0 -minimization problem (1) is equal to l_1 -minimization problem when solution x is sparse enough.

$$\hat{x}_1 = \arg\min_x \|x\|_1 \quad s.t.\ Dx = y \tag{2}$$

This problem can be solved by Basis Pursuit (BP)[14] , fast iterative shrinkage-thresholding algorithm (FISTA) [15] and so on. Denoted by $\delta_i(x) = [0, \ldots, 0, x_{i,1}, x_{i,2}, \ldots, x_{i,n_i}, 0, \ldots, 0]^T$, the coefficient vector for class i is a new vector whose entries are zero except for those associated with class i . Next we calculate residual between y and $\hat{y}_i$, where $\hat{y}_i = D\delta_i(\hat{x})$. We classify test sample y by finding the class which leads to the minimal residual.

$$\min_i r_i(y) = \|y - \hat{y}_i\|_2 \tag{3}$$

The procedures of SRC are summarized as follows.

Table 1. The SRC algorithm

Step 1: Normalize each atom in dictionary, letting $\|d_i\|_2 = 1$.

Step 2: Given a test sample y , find sparse solution x by solving problem:

$$\hat{x}_1 = \arg\min_x \|x\|_1 \quad s.t.\ Dx = y$$

Step 3: Calculate residual $e_i(y) = \|y - \hat{y}_i\|_2$, where $i = 1, 2, \ldots, k$.

Step 4: Select a minimal residual and output the label.

$$identity(y) = \arg\min_i e_i$$

3. SPARSE REPRESENTATION-BASE JOINT CLASSIFICATION

3.1. Linear Classifier Model

A relevant work in [9] attempted to iteratively update dictionary and LC with feedback from the classification stage. Different from [9], the dictionary remains fixed during the training process so that the labels in original dictionary are preserved. That means we could use two classifiers to jointly make decision. On the other hand, keeping the dictionary fixed helps objective function achieve convergence much faster. The proposed model of LC is formulated as

$$W = \arg\min_{W,X} \|Y - DX\|_F + \lambda_1 \|X\|_1 + \lambda_2 \|H - WX\|_F + \lambda_3 \|W\|_F \tag{4}$$

where $\|\cdot\|_1$ denotes summation of l_1 -norm of all columns in a matrix. $X = \{x^i\}_{i=1}^u \in \mathbb{R}^{n \times u}$ is constituted by sparse vectors corresponding to training image set $Y = \{y^i\}_{i=1}^u \in \mathbb{R}^{m \times u}$. W denotes the linear classifier. Each column of H is a label vector of training image: $h_i = [0, 0, \ldots, 1, \ldots, 0, 0]^T \in \mathbb{R}^{k \times 1}$, where the position of non-zero entry indicates the subject. $\|Y - DX\|_F$ is the representation error, $\|H - WX\|_F$ the classification error, $\|X\|_1$ the sparse constraint and $\|W\|_F$ is the regularization penalty term. λ_1 , λ_2 and λ_3 are scalars keeping these terms in balance.

Equation (4) is a multivariable optimization problem which is a challenging work. The relevant researches are reported in [8,9,16]. These iterative algorithms alternately update one variable in each step while keeping others unchanged. However, it achieves convergence slowly and is sensitive to initialization. For simplicity, we propose a two-step algorithm to get an approximate solution, as shown in Table 2. Since our proposed algorithm does not execute iteratively, it reduces computational time.

Table 2. The algorithm of training linear classifier

Step 1: Initialize Y , D , H .

Step 2: Dropping the last two terms, problem (4) is equal to $\|Y - DX\|_F + \lambda_1 \|X\|_1$. We could obtain the sparse solutions X by BP or FISTA.

Step 3: Dropping the first and second terms, problem (4) is equal to $\|H - WX\|_F + (\lambda_3 / \lambda_2) \cdot \|W\|_F$. We could easily get to know $W = HX^T (X^T X + (\lambda_3 / \lambda_2)^2 \cdot I)^{-1}$, where X is obtained by **step 2**.

Step 4: Output X and W .

After a query sample being coded, its sparse vector carries discriminative information. As we mention before, an ideal i th test sample whose sparse vector x should be of the form $x = [0, \ldots, 0, x_{i,1}, x_{i,2}, \ldots, x_{i,n_i}, 0, \ldots, 0]^T$. Obviously these sparse vectors are linearly separable and could be classified by linear classifier W [8].

$$identity(y) = \arg\max_i z_i \tag{5}$$

where $z = W\hat{x} = [z_1, z_2, \ldots, z_k]^T \in \mathbb{R}^{k \times 1}$. Intuitively, z will be of the form $[0, 0, \ldots, 1, \ldots, 0, 0]^T$ in ideal case.

3.2. Criterion for evaluating the classification reliability

After test samples are classified, the classification reliabilities of them are different and could be measured. Residual distributions of two images from Extend Yale B database are shown in Fig.1. Both of two test samples are selected from subject 1 but classification in (a) is intuitively more clear than (b), in other words, more reliable. When the difference of two smallest residuals comes to zero, decision making would be meaningless.

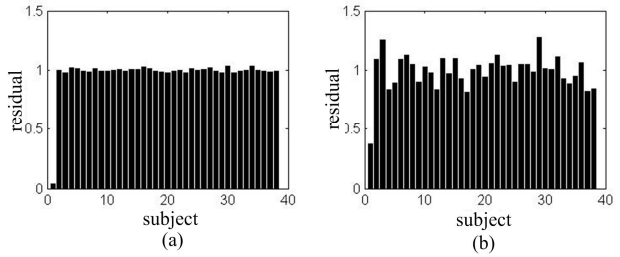

Fig. 1. Example in Extend Yale B database (a) residual of a test sample which could be well classified by DC (b) residual of a poor test sample which is difficultly classified by DC

In a multiple classifier system, it is also important to evaluate classification reliability of every classifier for the subsequent decision combination. Besides, the ability to detect and then reject an invalid test sample (outlier) is crucial for a recognition system. Sparsity Concentration Index (SCI) [1], described classification reliability from the perspective of sparse coefficient, performs well in detecting invalid sample within the framework of SRC. The SCI of a sparse vector x is formulated as

$$\mathrm{SCI}(x)=\frac{k\cdot\max_i\left\|\delta_i(x)\right\|_1/\left\|x\right\|_1-1}{k-1}\in[0,1] \tag{6}$$

In this paper, we proposed a criterion called residual-based index (RI) to measure classification reliability. The criterion is going to respectively evaluate classification reliability of DC and LC, which is precondition for making decision.

Assuming the coding residual $\left\|y-D\delta_i(\hat{x}_1)\right\|_2$ of each subject and each pixel in test sample follows Gaussian distribution[17]. The posterior probability that test sample y belongs to the i th subject could be given as

$$P(\vec{y}_i\mid y)=\frac{1}{\mu_1\cdot(2\pi)^{m/2}\cdot\left|C\right|^{1/2}}\exp\left\{-\mu_2\cdot(\vec{y}_i-\vec{y})^T\cdot C\cdot(\vec{y}_i-\vec{y})\right\} \tag{7}$$

where $\hat{y}_i=D\delta_i(\hat{x}_1)$, m denotes dimension of y and $|\bullet|$ denotes determinant of a matrix. C is the covariance matrix between y_i with y. For simplify, we set C to identity matrix, and drop constant $\sqrt{(2\pi)^m\cdot|C|}$. The RI of DC and LC are respectively given as

$$\begin{aligned}\mathrm{RI}(\vec{y}_i\mid y)&=\frac{1}{\mu_1}\exp\left\{-\mu_2\cdot\left\|\vec{y}_i-y\right\|_2^2\right\}\\ \mathrm{RI}(\vec{z}_i\mid z)&=\frac{1}{\mu_1}\exp\left\{-\mu_2\cdot\left\|\vec{z}_i-z\right\|_2^2\right\}\end{aligned} \tag{8}$$

RI is normalized by $\mu_1=1\big/\sum_i\exp\{-\mu_2\cdot\left\|\vec{y}_i-\vec{y}\right\|_2^2\}$ or $\mu_1=1\big/\sum_i\exp\{-\mu_2\cdot\left\|\vec{z}_i-z\right\|_2^2\}$, which depends on the classifier we use. μ_2 is set to 1. If $\max_i\mathrm{RI}_i\approx1/k$, all subjects almost share the same residual and decision making becomes meaningless. Higher value of $\max_i\mathrm{RI}_i$ indicates more reliable classification result. For example, two RIs of test sample from Fig.1 were calculated. In Table 3 we could measure the classification reliability and make a conclusion that the result in (a) is more accurate than (b).

Table 3. RIs of two test samples in Fig.1

Sample(a)	Sample(b)
0.0659	0.0480

3.3. Joint classification Strategy

Upon completion of training by proposed algorithm, we obtain a linear classifier W. How could we use both DC and LC to get better classification accuracy? In other words, is it possible to combine DC and LC and achieve a jointly classifying?

We choose a threshold $\tau\in(0,1)$ and regard a test sample as poor one if $\mathrm{RI}^{DC}(y)<\tau$. In this paper the LC is addressed in dealing with poor test samples. It means that poor test samples need to be jointly classified while the others maintain the result of DC. Another threshold $\theta\in(0,1)$ is chosen for reliability guarantee. When $\mathrm{RI}^{LC}(y)>\theta$, we can correct the result by LC. The strategy of joint classification could be summarized in Fig.2.

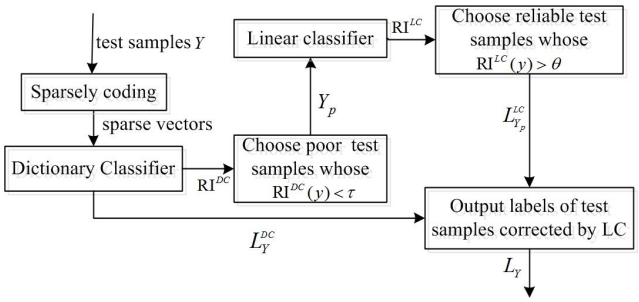

Fig. 2. Joint classification Strategy

4. EXPERIMENTAL VERIFICATION

In this section, we perform experiments on both Extend Yale B and AR database for face recognition and demonstrate that the joint classifier does improve recognition accuracy. We first present our approach in various feature dimensions and amount of atoms, compared to SRC[1] and D-KSVD[8]. And then we vary λ_2 and λ_3 in model to observe how the approach works under different parameters. We calculate sparse vector by BP solver which is released by SparseLab of Stanford.

4.1. Extend Yale B Database

The Extend Yale B database consists of 2414 face images of 38 classes. Each 192×168 image was cropped and captured

in various lighting condition[18]. For each subject, we randomly choose half of the images for dictionary constructing and training and the others for testing. We compute the recognition ratio with the down-sampling feature dimensions 30, 56, 120 and 504 just like [1]. Another experiment performs with 4, 8, 12, 16 atoms each subject of dictionary. In this section, λ_3/λ_2 is set to 1.

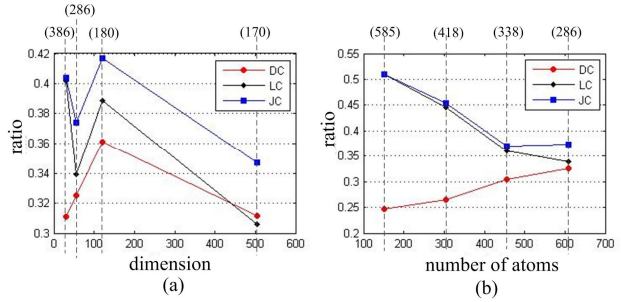

Fig. 3. Recognition rates of poor samples on Extend Yale B database. Results of DC, LC and JC are presented in (a)Different dimensions with 608-atoms dictionary. (b)Different numbers of atom at 120D. Note that the number in the bracket stands for the number of samples used to train LC.

Firstly, we run experiment with train set which are not using to construct original dictionary, and the number of unsuccessfully recognized samples is denoted by constant t. All training samples are ranked according to their RIs. The t training samples with lowest RIs are treated as poor sample set T_P which used for training LC. Then we set $\tau = \max \mathrm{RI}^{DC}(T_P)$ and $\theta = \min \mathrm{RI}^{LC}(T_P)$. LC outperforms DC since it is addressed in classifying poor samples, as shown in Fig.3(a)(b).

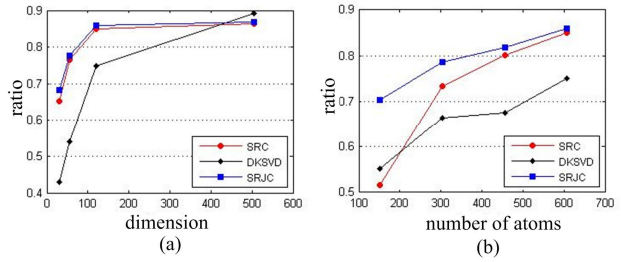

Fig. 4. Recognition rates on Extend Yale B database. SRJC compared with SRC and D-KSVD in (a) different dimensions with 608-atoms dictionary (b) different number of atoms at 120D.

Moreover, the joint classification of DC and LC comes up with better performance than any one of them, which means the proposed joint classification strategy works. In Fig.4(a), D-KSVD algorithm achieved best performance with 608 atoms(504D) dictionary, however, it cost a great deal of time in training compared with SRJC. In the practical scenario where dictionary is always consists of few atoms for less computation, DC performs quite poorly while the JC might greatly improve. As we can see in Fig.4(b), A 152 atoms(120D) dictionary achieves recognition rate of 51% using SRC while the recognition rate for SRJC is 70%.

4.2. AR Database

The AR database includes over 4,000 frontal images for 126 individuals. These images include more facial variations, including illumination change, expressions, and facial disguises comparing to the Extended Yale B database. For each subject, 14 various illumination images without any disguises were selected: 10 in them to form train set while using the others to test. In our experiment, we chose a subset of database that consists of 50 male subjects and 50 female subjects. We compute the recognition ratio with the down-sampling feature dimensions 32, 63, 126, 252, 420 and 630. Original dictionary includes 500 images consisting of 5 samples in each subject. In this section, λ_3/λ_2 is set to 0.01. Other setting is the same as experiment in Fig.3.

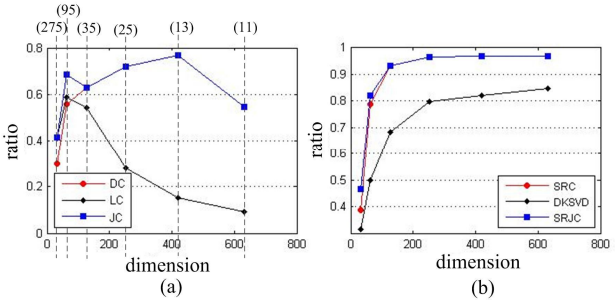

Fig. 5. Recognition rates on AR database. (a) Results of DC, LC and JC in different dimensions with 500-atoms dictionary. (b) SRJC compared with SRC and D-KSVD in different dimensions with 500-atoms dictionary. Note that the number in the bracket stands for the number of samples used to train LC.

When LC is trained by few poor samples, it works poorly and makes no contribution to JC. In other words, the better DC performs in training stage, the fewer poor samples are obtained which leads to an invalid LC and $\theta = \min \mathrm{RI}^{LC}(T_P) = 1$. Joint classification strategy rejects classification results whose $\mathrm{RI} < \theta$, which makes sure that JC would not be effected by the LC. Fig.5(a) shows that in the situation of 11, 13, 25 or 35 poor samples, the recognition ratio of JC would not decrease rapidly. Instead, its performance is close to DC because fewer RIs from LC could reach θ to correct original result.

4.3. Regularization parameter

Value of regularization parameter λ_3/λ_2 is set to avoid over-fitting or under-fitting. We also present experiments on two databases exploring the relationship between λ_3/λ_2 and recognition performance of JC in Fig.6.

When λ_3/λ_2 is set to 100 or larger, it comes to under-fitting and fail to classify. As the dimension of feature growing, a lower value of λ_3/λ_2 would help to avoid under-fitting. A fitted value of λ_3/λ_2 is the key to train LC which benefits subsequent joint classification.

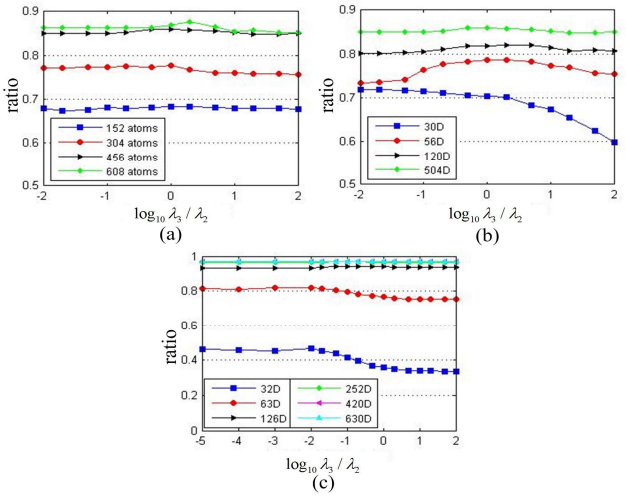

Fig. 6. Recognition rates of Extend Yale B database with various λ_3/λ_2 (a) experiment carried on different dimension with 608-atoms dictionary (b) experiment carried on different amount of atom with 120D. Recognition rates of AR database (c) experiment carried on different dimension with 500-atoms dictionary.

5. CONCLUSION

This paper proposed and elaborated a joint classification approach, namely SRJC, for face recognition. We apply a simplified training method to obtain LC. Compared to the standard SRC, LC improves the recognition accuracy by jointing working with DC. Moreover, an evaluating criterion for classification reliability is presented to validate the joint classification strategy. Experiments in Extend Yale B and AR database have verified that the JC outperforms each single classifier (LC and DC) in the FR task.

6. ACKNOWLEDGMENT

This work is supported by Guangdong and National ministry of education IAR project (Grant No. 2012B091100331), NSFC - Guangdong Union Project (Grant No. U0835003), Guangzhou Science and technology Project (Grant No. 2014J4100247) and NSFC (Grant No. 61471174).

REFERENCES

[1] Wright, John, et al. "Robust face recognition via sparse representation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 31.2 (2009): 210-227.

[2] Duda, Richard O., Peter E. Hart, and David G. Stork. Pattern classification. John Wiley & Sons, 2012.

[3] Ho, Jeffrey, et al. "Clustering appearances of objects under varying illumination conditions." Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. Vol. 1. IEEE, 2003.

[4] Guo, Guodong, Stan Z. Li, and Kap Luk Chan. "Face recognition by support vector machines." Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. IEEE, 2000.

[5] Turk, Matthew, and Alex Pentland. "Eigenfaces for recognition." Journal of cognitive neuroscience 3.1 (1991): 71-86.

[6] Belhumeur, Peter N., João P. Hespanha, and David Kriegman. "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection." Pattern Analysis and Machine Intelligence, IEEE Transactions on 19.7 (1997): 711-720.

[7] Aharon, Michal, Michael Elad, and Alfred Bruckstein. "-svd: An algorithm for designing overcomplete dictionaries for sparse representation." Signal Processing, IEEE Transactions on 54.11 (2006): 4311-4322.

[8] Zhang, Qiang, and Baoxin Li. "Discriminative K-SVD for dictionary learning in face recognition." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[9] Pham, Duc-Son, and Svetha Venkatesh. "Joint learning and dictionary construction for pattern recognition." Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008.

[10] Amaldi, Edoardo, and Viggo Kann. "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems." Theoretical Computer Science 209.1 (1998): 237-260.

[11] Donoho, David L. "For most large underdetermined systems of linear equations the minimal ℓ1- norm solution is also the sparsest solution." Communications on pure and applied mathematics 59.6 (2006): 797-829.

[12] Candes, Emmanuel J., and Terence Tao. "Near-optimal signal recovery from random projections: Universal encoding strategies?." Information Theory, IEEE Transactions on 52.12 (2006): 5406-5425.

[13] Candes, Emmanuel J., Justin K. Romberg, and Terence Tao. "Stable signal recovery from incomplete and inaccurate measurements." Communications on pure and applied mathematics 59.8 (2006): 1207-1223.

[14] Chen, Scott Shaobing, David L. Donoho, and Michael A. Saunders. "Atomic decomposition by basis pursuit." SIAM journal on scientific computing 20.1 (1998): 33-61.

[15] Beck, Amir, and Marc Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems." SIAM Journal on Imaging Sciences 2.1 (2009): 183-202.

[16] Z. Shi and Y. Lu, "An efficient initialization method for D-KSVD algorithm for image classification," in Image and Signal Processing (CISP), 2013 6th International Congress on, 2013, pp. 1029-1034.

[17] Yang, Meng, D. Zhang, and Jian Yang. "Robust sparse coding for face recognition." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[18] Georghiades, Athinodoros S., Peter N. Belhumeur, and David Kriegman. "From few to many: Illumination cone models for face recognition under variable lighting and pose." Pattern Analysis and Machine Intelligence, IEEE Transactions on 23.6 (2001): 643-660.