Sparse Representation with Regularization Term for Face Recognition

Jian Ji(✉), Huafeng Ji, and Mengqi Bai

Department of Computer Science & Technology, Xidian University, Xi'an 710071, China
jji@xidian.edu.cn

Abstract. In recent years there has been a growing interest in the study of sparse representation based classification (SRC) which has obtained great success in face recognition. However, SRC is overly dependent on the size of training samples while overlooking the correlation information that is critical to the real-world face recognition problems. Besides, some method considers the correlation information but overlooks the discriminating ability of sparsity. In this paper, we propose a new method called trace norm sparse representation based classification (TSRC) which introduces a regularization term in the SRC model and considers both sparsity and correlation. The TSRC method can benefits from both ℓ_1-norm and ℓ_2-norm, which is flexible and can obtain satisfactory results. Experimental results on 2 face databases clearly show that the proposed TSRC method outperforms many state-of-the-art face recognition methods.

Keywords: Face recognition · Trace norm · Sparse representation based classification · Sparsity and correlation

1 Introduction

Face recognition, as one of the most successful applications of image analysis and understanding, has recently received significant attention and adequate development, especially during the past decades.Nevertheless, due to the different interference of different conditions cause corruption and errors of different degrees, for example, various facial expression, pose and illumination conditions, the face image processing effect is not so ideal. Furthermore, when the feature space is not sufficient sample database and high dimension, the existence of these problems will meet more challenges in face recognition.

The conventional method of face recognition(sparse PCA [1], 2DPCA [2]) selects a limited subset or model from training samples, instead of the entire training set for image detection or signal classification and representation. So when the train sample space is small, the performances are not very good. These methods based on feature space, such as NN (Nearest Neighbor) and the support vector machine (SVM), when the image between different classes is very similar, will have a low recognition effect.

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 10–20, 2015.
DOI: 10.1007/978-3-662-48570-5_2

Therefore, face recognition methods based sparse representation classification emerge as the times requirement [3-5].Sparse representation method is based on the hypothesis that the testing images are approximation in a low dimension subspace which is obtained by the training samples, and then can be represented by a small number of training samples. Sparse representation based classification [3] (SRC) seek sparse representation of a query image in an over-complete dictionary, and then obtain recognition performance through comparing the minimal sparse error to identity the query image class. SRC can be seen as a generalization of NN and NFS, but it can get better recognition performance [3]. SRC overemphasize sparsity of data while ignoring the correlation between the dictionaries, which often results in lack of information. Thus, when the training samples are highly correlated, SRC will produce unstable results. Some of the literature has shown the importance of correlation structure [6-8]. Zhang proposed the CRC method which made full use of the correlation data for face recognition and used the ℓ_2-norm model [9].

Only when the training sample is large, the SRC method shows a good recognition performance and it can't use the correlation data to obtain useful information. While the CRC method can get a good result by the correlation, but when the correlations of training samples are limited, it may not perform well. We propose a new face recognition method called the trace norm sparse representation classification (TSRC) which applies the trace norm as the regularization term into the dictionary. The trace norm can benefit from ℓ_1-norm and ℓ_2-norm, in other words it can take advantage of sparsity as well as data correlation. After we proved the feasibility of the regularization term and answered the minimization problem of the trace norm, we draw a lot of experiments in different face image databases, and compare the face recognition performance between different methods including SRC [3], SVM [10], NN, NFS [11] and LSRC [6].

2 Backgrounds of Sparse Representation

2.1 Generalized Sparse Representation

Denote the data set of training samples labeled with the i-th class as $A_i = [v_{i,1}, v_{i,2}, ..., v_{i,n_i}]$, Any new test sample y from the same class can be linearly expressed as:

$$y = \alpha_{i,1} v_{i,1} + \alpha_{i,2} v_{i,2} + \alpha_{i,n_i} v_{i,n_i} \tag{1}$$

where $\alpha_{\mathrm{i},j}$ are some scalars.

We define a new matrix A for the entire training set as the concatenation of the n training samples of all k object classes:

$$A = [A_1, A_2, ..., A_k] = [v_{1,1}, v_{1,2}, ..., v_{k,n_k}] \tag{2}$$

Then, the linear representation of y can be rewritten in terms of all training samples as:

$$y = Ax_0 \tag{3}$$

where $x_0 = [0,...,0,\alpha_{i,1},\alpha_{i,2},...,\alpha_{i,n_i},0,...,0]^T$ is a coefficient vector whose entries are zero except those associated with the i-th class.

The purpose of sparse representation is to estimate the main information of the test sample using non-zero coefficient as little as possible. In other words, we need to find the x_0 which has less non-zero coefficient and can be a good estimation of y with A.

2.2 Classification Based on Sparse Representation (SRC)

J. Wright et al. introduced the Sparse Representation based Classification (SRC) method which had applied to face recognition and pattern recognition [3]. The model is as follows,

$$(\ell^0):\hat{x}_1 = \arg\ \min||x||_0 \quad s.t \quad Ax = y \tag{4}$$

$||x||_0$ is the ℓ_0-norm, defined as the number of non-zero entries in the vector x. The problem of ℓ_0-norm can be proved is NP hard, even if making approximately calculation is also very difficult [12]. Some paper reveals that if the solution x_0 sought is sparse enough, the ℓ_0-minimization problem (4) is equal to the solution to the following ℓ_1-minimization problem:

$$(\ell^1):\hat{x}_1 = \arg\ \min||x||_1 \quad s.t \quad Ax = y \tag{5}$$

$||x||_1$ is the ℓ_1-norm, defined as $||x||_1 = \sum_i |x_i|$, that is the sum of the absolute values of all the entries. The model (3) can be explicitly modified to the flowing form account for small possible dense noise:

$$y = Ax_0 + z \tag{6}$$

where $||z||_2 < \varepsilon$, z is noise item , the sparse solution x_0 can be obtained by the following ℓ_1-minimization problem:

$$\hat{x}_1 = \arg\min||x||_1 \quad s.t \quad ||Ax - y||_2 \le \varepsilon \tag{7}$$

$||x||_2$ is the ℓ_2-norm, defined as $||x||_2 = \sqrt{\sum_i x_i^2}$. Based on [13], when $||x_0||_0 < \rho m$ and $||z||_2 \le \varepsilon$, we can learn that there are constants ρ and ζ satisfied with

$$||\hat{x}_1 - x_0|| \le \zeta\varepsilon \tag{8}$$

So we can use the formulas (8) to examine the computed $\hat{x}_1$.

For x , $\delta_i(x)$ is a new vector whose only nonzero entries are the entries in x that are associated with class i . So we can approximate the given test sample y as $\hat{y}_i = A\delta_i(\hat{x}_1)$. We then classify y based on these approximations by assigning it to the object class that minimizes the residual between y and $\hat{y}_i$:

$$\min_i \ r_i(y)\|y - A\delta_i(\hat{x}_1)\|_2 \tag{9}$$

3 Sparse Representation with Regularization Term for Face Recognition

3.1 Why Did We Introduce the Regularization Term?

The SRC algorithm is under the assumption that image and training images are in very good agreement. The results show that when there are enough training samples which can cover all changes, y can be correctly expressed. Therefore, SRC may not obtain satisfactory results at the case where y are not aligned and dictionary contains a small amount of sample. At the same time, due to the sparsity, when the samples are highly correlated, SRC may have the problem of unstable. If the object and the query image are similar, the SRC method tends to choose a random object instead of choosing them all. This means that, SRC does not capture the relevant structure of the dictionary that plays crucial role in the face recognition [14].

Good performances of SRC comes from the collaborative representation of y is on all training samples [9]. CRC can make good use of the advantages of data correlation [9]. Therefore, in the CRC, the images is represented by an over complete dictionary which use ℓ_2 -norm rather than use ℓ_1 -norm to control coding vector. The object function of ℓ_2 -norm is as follows.

$$(\ell^2): \hat{x}_2 = \arg\min \|x\|_2 \quad s.t \quad Ax = y \tag{10}$$

Considering the noise problem, the equation can be changed into

$$(\ell^2): \hat{x}_2 = \arg\min \|x\|_2 \quad s.t \quad \|Ax\text{-}y\|_2 \le \varepsilon \tag{11}$$

ℓ^2 made CRC obtain a stable results through the use of a more dense vector, but when the training samples were not highly correlated, the CRC would not be able to get good results.

Only when the training sample is large, the SRC method can show a good recognition performance and it can’t use the correlation data to obtain useful information. While the CRC method can get a good result by the correlation, but when the correlations of training samples are limited, it may not perform well. the trace norm classification method based on sparse representation (TSRC) overcome the disadvantages of SRC and CRC. For fully considering the sparsity and correlation, we combine structure of matrix A and coding coefficient x and introduce the trace norm inspired by [15]. Giving a matrix M, diag(x) indicates converting the matrix M into a vector in which the i-th entry is M_{ii} located in the diagonal of M. The ℓ_1-norm of M is defined as $||M||_1 = \sum_{ij} |m_{ij}|$ and the trace $||M||_*$ is regarded as the sum of all the singular values of the matrix M. Thus we get following linear representation model:

$$\min ||ADiag(x)||_* \quad s.t. \quad y = Ax \tag{12}$$

where, $||ADiag(x)||_*$ is a regularization term defined as $\Omega_A(x)$. With the regularization term, we will no longer ignore sparsity or correlation.

3.2 Extreme Value of the Regularization Term

There are two extreme cases for the trace norm regularization term which can be discussed as follows.

1). When the columns of matrix A are not related and A is an orthogonal matrix, that is $A^T A = I$. And then we can get

$$\begin{aligned} &|| ADiag(x) ||_* = Tr[(ADiag(x))^T (ADiag(x))]^{1/2} \\ &= Tr[(Diag(x))^T (Diag(x))]^{1/2} \\ &= || x ||_1 \end{aligned} \tag{13}$$

Thereby, $\Omega_A(x)$ is equivalent to ℓ_1-norm. So we can change (12) into

$$\min ||x||_1 \quad s.t. \quad y = Ax \tag{14}$$

2). In the case where the images of different subjects look similar to a_1, that is $A = a_1 E^T$ and $A^T A = EE^T$ (E is a vector of size n whose elements are one), we can express $\Omega_A(x)$ as:

$$\Omega_A(x) = ||a_1 x^T||_* = ||a_1||_2 ||x||_2 = ||x||_2 \tag{15}$$

Then (12) can be changed into

$$\min ||x||_2 \quad s.t. \quad y = Ax \tag{16}$$

Generally, the images in the dictionary are neither too independent of each other nor look the same. Our model is able to obtain the correlation structure of the training samples. So we can easily know that

$$||x||_2 \le \Omega_A(x) \le ||x||_1 \tag{17}$$

It is show that x obtained by $\Omega_A(x)$ is more sparse than the one obtain by ℓ_2, but is less sparse than the one obtained by ℓ_1. This means that we can take advantage of ℓ_1 and ℓ_2 .

3.3 Our Novel Sparse Representation Model (TSRC)

If the noise obeys Gauss distribution, the objective function can be turned into

$$\min \|ADiag(x)\|_* \quad s.t. \quad \|y\text{-}Ax\|_2 \le \varepsilon \tag{18}$$

Instead if the occlusion follows the Laplace distribution, we consider the following problem:

$$\min \|ADiag(x)\|_* \quad s.t. \quad \|y\text{-}Ax\|_1 \le \varepsilon \tag{19}$$

Problem (19) is more robust to occlusion and variations than problem (18) [16]. Problem (19) can be changed into the following problem:

$$\min \|y - Ax\|_1 + \lambda \|ADiag(x)\|_* \tag{20}$$

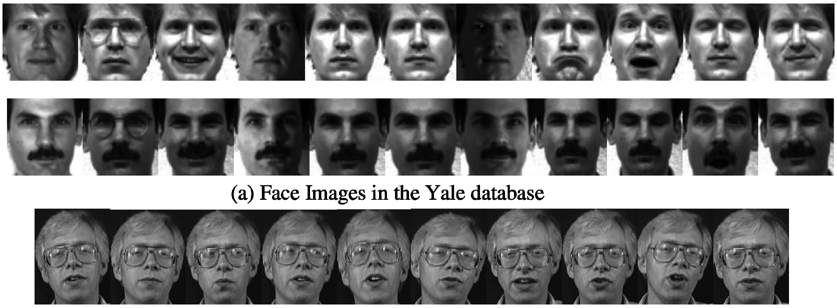
where $\lambda > 0$ is the regularization parameter. We adopt Alternating Direction Met hod (ADM) [17] to solve problem (20).
The final novel sparse representation model is expressed as follows:

$$\min \|y - Ax\|_1 + \lambda \|ADiag(x)\|_* + \eta \|x\|_2 \tag{21}$$

η is an optional parameter whose value depends on the distribution type of noise. The type of noise determines the value of the parameter η. If the noise obeys Gauss distribution, $\eta = 1$ and if the occlusion follows the Laplace distribution, we set the η to 0.

4 Experiment

This part we will demonstrate the recognition effect of TSRC in 2 face image databases in Fig. 1, respectively, to select two groups of images totally having 22 images on Yale face database (each group has 11 images) and one group ORL face database images totally having 10 images.



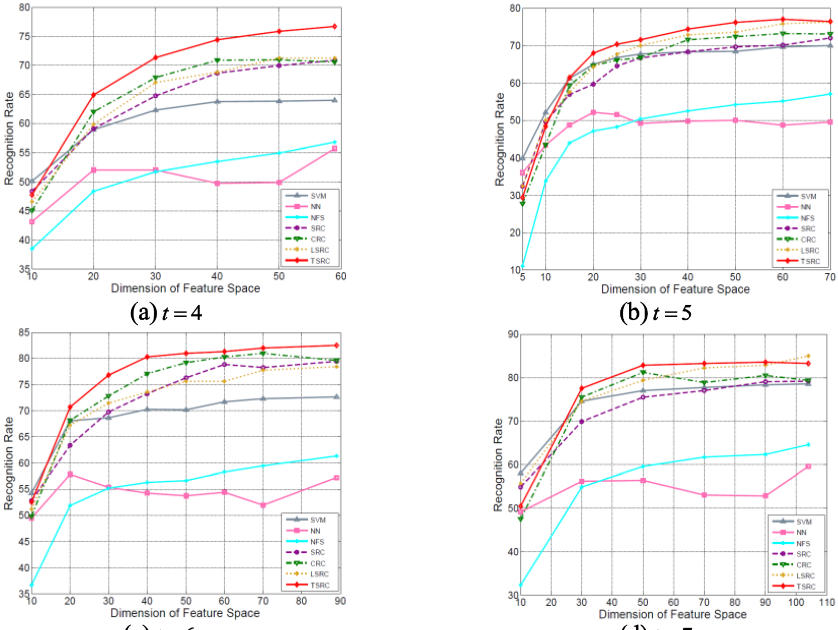
(a) Face Images in the Yale database

(b) Face Images in the ORL database

Fig. 1. Yale and ORL face databases

4.1 Experiments on Yale Face Database

The Yale face database is composed of 15 volunteers, 11 gray face images per person for a total of 165 different images of 32×32 pixel. These images with different expressions are obtained in different illumination. As we can see in Fig. 1(a) expressions of each individual are different. In this experiment, we randomly selected $t(=4,5,6,7)$ images of each individual as a training sample and the rest of the images as test images. For different t with different dimension of the feature space (in 10 increment), we record the average precision in Fig. 2 and the maximum average accuracy as well as the standard deviation corresponding to the value in the Table 1.



(a) $t=4$ (b) $t=5$

(c) $t=6$ (d) $t=7$

Fig. 2. In the Yale face database, recognition rate of various method under different t and dimension of feature space

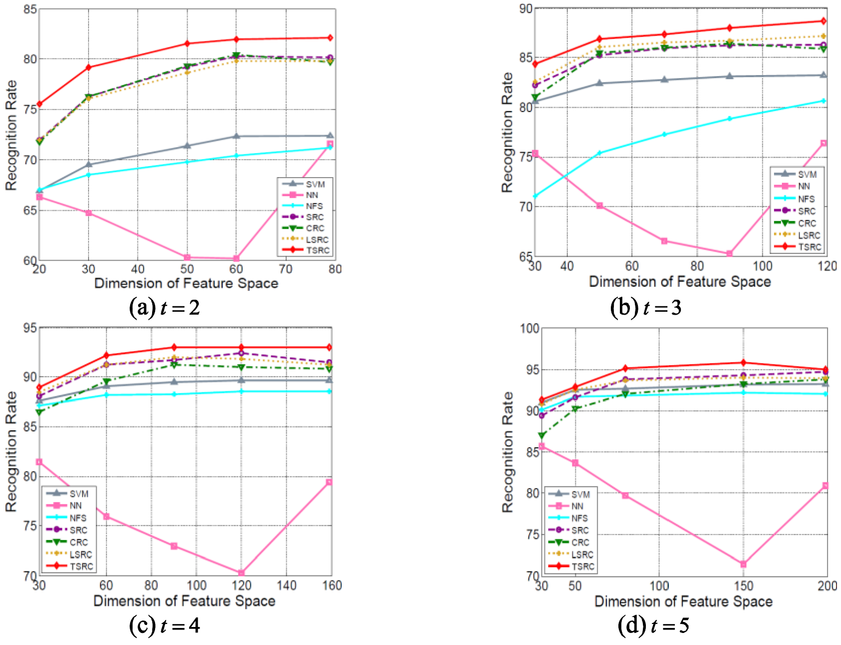
Table 1. Maximum average accuracy and standard deviation of different methods

Algorithms	$t=4$	$t=5$	$t=6$	$t=7$
SVM[10]	64.00±2.57(59)	69.89±5.48(70)	72.67±3.28(89)	78.50±6.73(104)
NN	55.71±4.65(59)	52.11±4.30(20)	57.87±4.92(20)	59.50±3.69(104)
NFS[11]	56.76±5.30(59)	57.00±4.74(70)	61.33±5.96(89)	64.50±5.21(104)
SRC[3]	70.86±4.56(59)	72.00±4.02(70)	79.47±3.68(89)	79.17±3.17(104)
CRC[9]	70.95±4.67(50)	73.11±4.79(60)	80.93±3.93(89)	81.17±3.93(50)
LSRC[6]	71.24±2.49(50)	76.22±3.93(70)	78.40±3.86(70)	85.00±5.56(104)
TSRC	74.56±4.71(59)	77.00±3.98(60)	81.31±2.67(89)	83.17±4.89(104)

We can see from the Fig. 2 and Table 1 that TSRC has better recognition rate than other methods at different t values with the change of feature space. When the t value is small (as $t = 4,5$), the maximum value of recognition is generally less than the larger value of t (as $t = 6,7$), which just corresponds to the case of SRC. Because TSRC can keep the dictionary correlation and sparsity, so when the t value is small, the good identification rate can manifest its advantages. Smaller t values mean that the number of training samples is small, TSRC can still get changes of the query images by choosing training samples with sufficient correlation so as to get better recognition rate. When t increasing, TSRC, SRC, CRC and LSRC have better recognition rate and when $t = 4$ and $t = 6$, SRC, CRC, LSRC have relatively similar curve. When $t = 5$ and $t = 7$, LSRC is superior to SRC and CRC's performance, this is due to the partial information of dictionary considered by LSRC. However, the methods are not good as TSRC proposed in this paper in addition to the extremely individual points because TSRC can accurately grasp the structural information of dictionary and enable it to better adapt to the query image.

4.2 Experiments on ORL Face Database

The ORL face database consists of 40 individuals and each individual has 10 gray images including different illumination, facial and detail changes, as Fig. 1 (b) shown. We select training samples of number t and the remaining are the query images in this experiment.The experimental results are shown in Fig. 3 and Table 2.



(a) $t = 2$

(b) $t = 3$

(c) $t = 4$

(d) $t = 5$

Fig. 3. In the ORL face database, recognition rate of various method under different t and dimension of feature space.

Table 2. Maximum average accuracy and standard deviation of different methods as well as the feature space dimension when the maximal accuracy is obtained.

Algorithms	$t = 2$	$t = 3$	$t = 4$	$t = 5$
SVM	73.38±4.00(79)	83.25±1.94(119)	89.67±1.90(120)	93.20±1.60(199)
NN	71.59±3.23(79)	79.36±2.37(119)	81.46±1.89(30)	85.70±2.37(30)
NFS	71.19±3.48(79)	80.64±1.70(119)	88.54±2.03(120)	92.20±1.72(120)
SRC	80.28±2.52(60)	86.29±1.58(119)	92.37±0.88(120)	94.70±1.44(199)
CRC	80.44±2.41(60)	86.39±2.07(90)	91.21±1.72(90)	93.75±2.12(199)
LSRC	79.81±2.46(60)	87.14±1.87(119)	92.00±1.22(90)	94.00±2.12(150)
TSRC	81.36±2.58(79)	88.52±1.96(119)	93.00±1.87(159)	95.69±2.09(150)

We can see from Fig. 3 and Table 2 that recognition rate of TSRC is higher than other methods. Comparison of several figures, the most obvious is that the recognition rate of NN method is the lowest and it is unstable. When the number of training is low ($t = 2,3$), SRC, CRC and LSRC have similar recognition rate. As can be seen from table 2, the recognition rate of TSRC shows a rise tendency as a whole, but the feature space dimension when the maximal accuracy is obtained is not rising. For example, when $t = 5$, the feature space dimension is 150 when the maximal accuracy is obtained which is small than other t, that is to say when the training samples are sufficient, the blindly increase of feature space dimension may not increase the recognition rate. However, TSRC have better recognition performance than other methods.

4.3 Summary

In general, SRC, CRC and LSRC have stable recognition rate in most cases. When the training sample size was small, CRC showed better recognition performance because it considered the correlation of data while sparsity showed lower effect. LSRC can get good recognition rate than SRC because the local information and sparsity of sample date were taken into account. But when the local information is not sufficient, TSRC can consider correlation and sparsity of the sample, so in most cases TSRC can get better recognition results than other methods. Therefore, the experiments proved that TSRC is a good method for face recognition.

5 Conclusions

We do have proved that the TSRC method have better recognition performance than other face recognition methods, such as NN, SVM, CRC and LSRC. It can benefit from sparsity and correlation. Specifically, TSRC can obtain comparable results to SRC when the dictionary is with low correlation, and performs as well as CRC when the data are with high correlation. TSRC can make good use of correlation between the query images and training samples, and then it can obtain relatively much information. LSRC only considers the limited information of few local samples in a small number of training samples. Experimental results on face database clearly show that the proposed TSRC method outperforms many state-of-the-art face recognition methods.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No. 61101248) and the Fundamental Research Funds for the Central Universities of China (No. JB140315).

References

1. d'Aspremont, A., Ghaoui, L.E., Jordan, M., Lanckriet, G.: A Direct Formulation of Sparse PCA Using Semidefinite Programming. SIAM Rev. **49**, 434–448 (2007)
2. Zhang, D.Q., Zhou, Z.H.: Two-directional two-dimensional PCA for efficient face representation and recognition. Neurocomputing **69**, 224–331 (2005)
3. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. on Pattern Analysis and Machine Intelligence **31**(2), 210–227 (2009)
4. Pillai, J.K., Patel, V.M., Chellappa, R.: Sparsity inspired selection and recognition of iris images. In: Proc. IEEE Third International Conference on Biometrics: Theory, Applications and Systems, pp. 1–6 (2009)
5. Hang, X., Wu, F.-X.: Sparse representation for classification of tumors using gene expression data. Journal of Biomedicine and Biotechnology (2009). doi:10.1155/2009/403689
6. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality constrained linear coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3360–3367 (2010)
7. Lu, J., Tan, Y., Wang, G., Gao, Y.: Image-to-set face recognition using locality repulsion projections and sparse reconstruction-based similarity measure. IEEE Trans. on Circuits and Systems for Video Technology **23**(6), 1070–1080 (2013)
8. Wang, M., Gao, Y., Lu, K., Rui, Y.: View-based discriminative probabilistic modeling for 3d object retrieval and recognition. IEEE Trans. on Image Processing **22**(4), 1395–1407 (2013)
9. Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: Which helps face recognition? In: IEEE International Conference on Computer Vision, pp. 471–478 (2011)
10. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(27), 1–27 (2011)
11. Shan, S., Gao, W., Zhao, D.: Face identification from a single example image based on face-specific subspace (fss). In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 2125–2128 (2002)
12. Amaldi, E., Kann, V.: On the Approximability of Minimizing Nonzero Variables or Unsatisfied Relations in Linear Systems. Theoretical Computer Science **209**, 237–260 (1998)

13. Donoho, D.: For Most Large Underdetermined Systems of Linear Equations the Minimal ℓ_2-norm -Norm Near Solution Approximates the Sparest Solution. Comm. Pure and Applied Math. **59**(10), 907–934 (2006)
14. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(2), 301–320 (2005)
15. Lu, C.-Y., Min, H., Zhao, Z.-Q., Zhu, L., Huang, D.-S., Yan, S.: Robust and efficient subspace segmentation via least squares regression. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 347–360. Springer, Heidelberg (2012)
16. Wright, J., Ganesh, A., Yang, A., et al.: Sparsity and Robustness in face recognition (2011). arXiv:1111.1014
17. Lin, Z., Chen, M., Wu, L., Ma, Y.: The augmented lagrange multiplier method for exact recovery of a corrupted low-rank matrices, UIUC Technical Report UILU-ENG-09-2215, Tech. Rep. (2009)