SUPERVISED TRACE LASSO FOR ROBUST FACE RECOGNITION

Jian Lai and Xudong Jiang

School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore 639798,
{jlai1, exdjiang}@ntu.edu.sg

ABSTRACT

In this paper, we address the robust face recognition problem. Recently, trace lasso was introduced as an adaptive norm based on the training data. It uses the correlation among the training samples to tackle the instability problem of sparse representation coding. Trace lasso naturally clusters the highly correlated data together. However, the face images with similar variations, such as illumination or expression, often have higher correlation than those from the same class. In this case, the result of trace lasso is contradictory to the goal of recognition, which is to cluster the samples according to their identities. Therefore, trace lasso is not a good choice for face recognition task. In this work, we propose a supervised trace lasso (STL) framework by employing the class label information. To represent the query sample, the proposed STL approach seeks the sparsity of the number of classes instead of the number of training samples. This directly coincides with the objective of the classification. Furthermore, an efficient algorithm to solve the optimization problem of proposed method is given. The extensive experimental results have demonstrated the effectiveness of the proposed framework.

Index Terms— Face recognition, sparse representation, trace lasso.

1. INTRODUCTION

Face recognition, as one of the most active application in machine learning and computer vision, has been studied by numerous researchers around the world for more than two decades. The studies mainly focus on two parts. One is the feature extraction and the other is the classification. As the face images are often lying on a very high dimensional space, directly handling them requires very expensive computational resource and leads to the so-called "curse of dimension" [1, 2] problem. Many representative methods for feature extraction are proposed to alleviate this problem. Eigenface [3, 4] is introduced to minimize the reconstruction error in the low dimension; Fisherface [5, 6] searches an embedding to maximize the discriminative information; Laplacianface [7] is proposed to maintain the local structure in subspace. For the classification, linear classifiers are most commonly used in face recognition. Nearest neighborhood (NN) is popular for its simplicity. Nearest feature line (NFL) [8] searches the nearest distance from the query sample to the line connecting two training samples. Similar to NFL, nearest subspace plane (N-SP) [9] finds the closest point in the feature plane consisting of three training samples to the test sample. Linear regression classification (LRC) [10] considers a probe image as a linear combination of class-specific galleries. Although these methods perform well in the controlled environment, they are not robust to the query image with occlusion or disguise. In [11], a robust linear regression method, named trimmed linear regression (TLR), is proposed to alleviate this problem.

In [12], Wright *et al.* proposed a sparse representation coding (SRC) method by applying the sparse representation [13, 14, 15] to face recognition. Without using the class label information, SRC reconstructs the query image as a sparse linear combination of the training samples from all subjects. To handle the corruption in the query samples, in [12], a modified sparse representation is proposed to measure the reconstructed error by l_1-norm, which is more robust to real world contamination. The experimental results are impressive. The advantages of SRC attract many further studies in recent years [16, 17, 18, 19].

To achieve a sparse representation, SRC tends to choose one from the highly correlated samples randomly [20]. Due to the randomness, if the selected sample is not from the correct subject, the classification result will be wrong. To alleviate the shortcoming of SRC, Grave *et al.* [20] proposed a trace lasso norm. It is proved that trace lasso interpolates between the l_2-norm and l_1-norm. Its behavior is adaptively related to the correlation of the training data. Compared with SRC, trace lasso is more stable as it is in favor of using more highly correlated samples instead of one [20].

Face image includes much information, such as identity and variations (e.g., illumination and expression). In the uncontrolled environment, the variation information can be more significant than the identity. In this case, face images from different subjects with similar variations could have higher correlation than those from same subject but with different variations. As a result, trace lasso naturally clusters the samples with similar variations together. The outcome of trace lasso is contradictory to the goal of identification, which is to cluster the samples according to their identities. For a recognition task, the label information, which is often available, is much more crucial than the sample correlation. By cooperating the label information with trace lasso, in this paper, we propose a method named supervised trace lasso (STL). To reconstruct the query sample, the proposed STL approach seeks the sparsity of the number of classes instead of the number of training samples. This directly coincides with the objective of the classification. Furthermore, an efficient algorithm to solve the optimization problem of proposed method is given.

2. RELATED WORK

2.1. Sparse Representation Coding (SRC)

For NN, NFL [8], NSP [9], LRC [10] and TLR [11], they all try to utilizing samples from one subject to represent an unlabel face image. However, in SRC [12], Wright *et al.* considered the query sample as a linear combination of all training samples. It is assumed that the query sample are lying on the subspace spanned by the training samples from the same subject [12]. As a result, only a few training samples significantly contribute to the representation of the query image. Therefore, the face recognition problem can be formulated as

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_0, \tag{1}$$

where $\mathbf{y} \in \Re^m$ is an unlabel query sample, each column of $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n] \in \Re^{m\times n}$ represent a training sample, and $\|\cdot\|_0$ is the l_0-norm. Here, m is the feature dimension and n is the number of training samples. However, (1) is a NP-hard problem, which can only be solved by exhausting searching all subsets of entries of $\mathbf{x}$.

It has been proved that, if $\mathbf{x}$ is sparse enough, the solution of l_0 minimization problem (1) is equivalent to the solution of the following l_1 minimization problem [13, 14, 15]:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{2}$$

here, $\|\cdot\|_1$ is the l_1-norm. Many efficient methods have been proposed to solve problem (2) [21, 22, 23].

If training data are highly correlated, to achieve the sparse goal, SRC tends to randomly select one from the highly correlated samples [20]. Due to the randomness, SRC may employ the samples from the wrong subject, which leads to misclassify the query sample.

2.2. Trace Lasso

To alleviate the randomness problem of SRC, Grave *et al.* proposed a new norm named trace lasso $\|\mathbf{A}D(\mathbf{x})\|_*$ [20]. Here, $\|\cdot\|_*$ is the nuclear norm and $D(\mathbf{x}) \in \Re^{n\times n}$ is a diagonal matrix, whose diagonal elements are $\mathbf{x}$. It is shown that trace lasso interpolates between l_2-norm and l_1-norm [20]. Its behavior depends on the correlation of training data. When all columns of $\mathbf{A}$ are orthogonal to each other, the result of trace lasso is the same as l_1-norm of $\mathbf{x}$. When all columns of $\mathbf{A}$ are the same, the result of trace lasso is the same as l_2-norm of $\mathbf{x}$. Different from SRC, which tends to select one representative sample from a group of highly correlated samples [20], trace lasso shares the gain with more highly correlated samples. To replace the l_1-norm in (2) with the trace lasso norm, it [20] derives the following problem

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{A}D(\mathbf{x})\|_*. \tag{3}$$

As trace lasso naturally clusters the highly correlated data, it has been applied to the subspace segmentation [24].

3. PROPOSED APPROACH

Except for the identity information, face image often contains many other variations (e.g., illumination and expression) as well. In the uncontrolled environment, the variation information can be dominant. In this case, face images from different subjects with similar variations could have higher correlation than those from the same subject with different variations. As a result, trace lasso naturally clusters the samples with similar variations together. Therefore, trace lasso is contradictory to the goal of face identification, which is to cluster the samples by their identities. For a recognition task, the label information, which is available in most of the cases, is much more crucial than the sample correlation. To cooperate the label information with trace lasso, in this paper, we propose a method named supervised trace lasso (STL) as

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_1 + \lambda\|\mathbf{G}D(\mathbf{x})\|_*. \tag{4}$$

The first term of (4) is used to measure the reconstruction error with l_1-norm, which is much more robust than l_2-norm to handle real world contamination. Different from the trace lasso, which depends on the correlation of the training data $\mathbf{A}$, we introduce a proposed class dependent matrix $\mathbf{G} \in \Re^{m\times n}$. $\mathbf{G} = [\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_c]$ and $\mathbf{G}_i \in \Re^{m\times n_i}$, here n_i is the number of training samples in the ith class and c is the number of training classes. In $\mathbf{G}_i$, all elements in the ith row are one and those in the other rows are zero. With the proposed $\mathbf{G}$, the correlation of column vectors within the class is one and that between the classes is zero.

The supervised trace lasso term $\|\mathbf{G}D(\mathbf{x})\|_*$ can be considered as an approximation of rank of $\mathbf{G}D(\mathbf{x})$ [25]. Minimizing the second term of problem (4) is equivalent to minimize the rank of $\mathbf{G}D(\mathbf{x})$. $\mathbf{G}D(\mathbf{x})$ can be rewrite as

$$\mathbf{G}D(\mathbf{x}) = [x_1\mathbf{g}_1, x_2\mathbf{g}_2, \ldots, x_n\mathbf{g}_n], \tag{5}$$

here, x_j is the jth element of $\mathbf{x}$ and $\mathbf{g}_j$ is the jth column of $\mathbf{G}$. Minimizing the rank reduces the number of the linear independent columns in $\mathbf{G}D(\mathbf{x})$. As the columns associated with different classes of $\mathbf{G}$ are independent (orhogonal), to achieve a low rank, only coefficients from a few classes should be selected to represent the query image. Therefore, the proposed method automatically seeks a sparsity of the number of classes to represent the query image. The desired output of a classification system is one for correct class and zero for all others, which is the most class-wise sparse case. The proposed method directly coincides with this objective of the classification.

Let $\mathbf{x}_i$ denote the sub-vector of $\mathbf{x}$ associated with the ith class. $\mathbf{x}_i^k$ is the kth element of $\mathbf{x}_i$. The result of $\mathbf{G}_i D(\mathbf{x}_i)$ is

$$\begin{pmatrix} 0 & \dots & 0 \\ \vdots & \dots & \vdots \\ x_i^1 & \dots & x_i^{n_i} \\ \vdots & \dots & \vdots \\ 0 & \dots & 0 \end{pmatrix}, \tag{6}$$

where the ith row is $(\mathbf{x}_i)^T$ and zero for all other rows.

From (6), we can see that he rank of $\mathbf{G}_i D(\mathbf{x}_i)$ with only one nonzero element in $\mathbf{x}_i$ is the same as that with all nonzero elements in $\mathbf{x}_i$, which are rank one matrices. To minimize the representation error defined by the first term of (4), more training samples should be selected. Therefore, if one sample is selected, the proposed method is in favor of using more samples from the same class. In this case, the rank of $\mathbf{G}D(\mathbf{x})$ will not increase, while the reconstruction error will reduce.

Fig. 1 visualizes some results of the proposed method comparing to SRC and trace lasso. The training set contains 500 face images from AR database, where each of 100 persons has 5 images randomly selected from Session 1. The details of data setting is given in Section 4.2. The query sample shown in Fig. 1(a) is a face image with both side lights on and the correct subject has no training sample with similar lighting condition to the query sample. As a result, the largest coefficients of the both SRC and Trace Lasso are associated with a same training sample, which has the similar lighting condition to the query sample but is originated from a wrong subject. On contrast, the largest coefficient of the proposed method is associated with a training sample from the correct subject though its lighting condition is different from that of the query image. Moreover, compared with SRC and trace lasso, our method achieves the most class-wise sparse representation and selects the most samples of the correct subject. This illustrates that the proposed class-wise sparsity enhances the identity information of images and deemphasizes the image variation information within the subject.

3.1. Optimization

To solve the optimization (4), the original problem is converted to the following equivalent problem

$$\min_{\mathbf{J},\mathbf{x},\mathbf{e}} \|\mathbf{e}\|_1 + \lambda \|\mathbf{J}\|_* \text{ s.t. } \mathbf{e} = \mathbf{y} - \mathbf{A}\mathbf{x}, \mathbf{J} = \mathbf{G}D(\mathbf{x}). \tag{7}$$

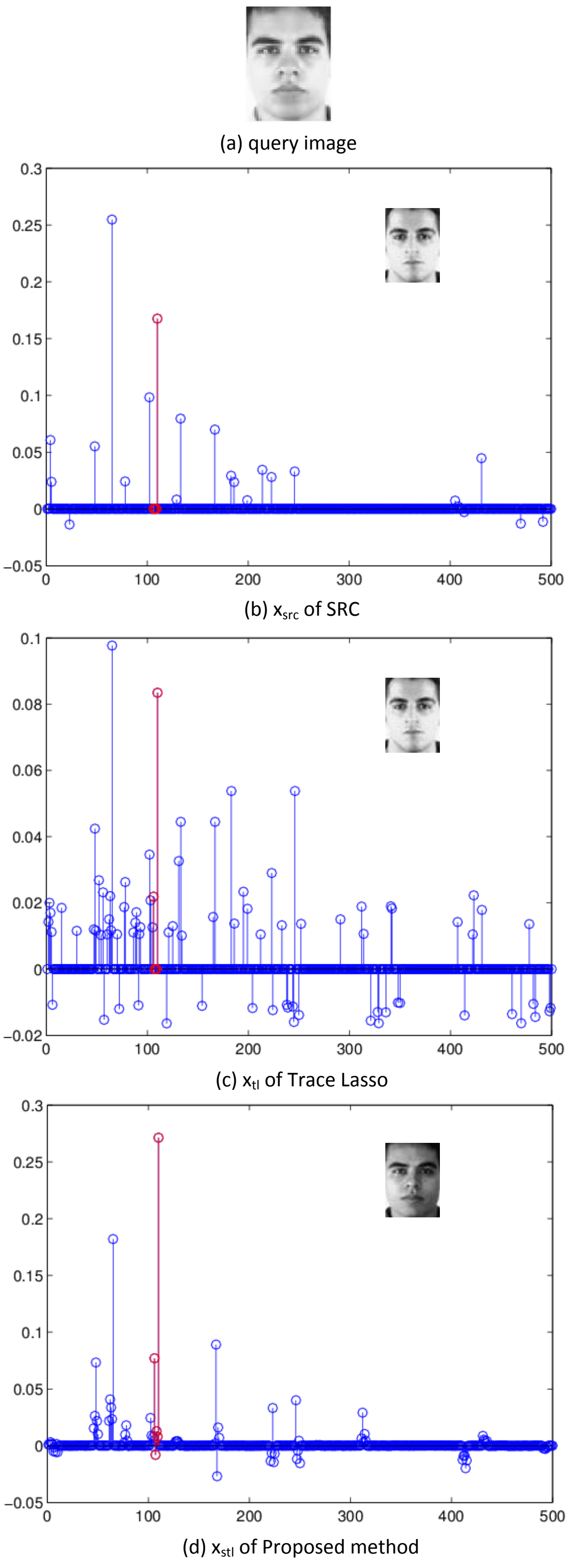

Fig. 1: (a) The query sample. Comparison the coefficient vector $\mathbf{x}$ of SRC (b), trace lasso (c) and STL (d). In (b), (c) and (d), the red circles denotes the coefficient associated with the correct subject and the face images are the samples associated with the largest coefficients.

The solution of problem (7) is difficult to be achieved directly. We adopt the Augmented Lagrange Multiplier scheme [26] to derive the following unconstrained optimization problem:

$$\begin{aligned}\min_{\mathbf{J},\mathbf{x},\mathbf{e}} \quad & \|\mathbf{e}\|_1 + \lambda\|\mathbf{J}\|_* \\ + \quad & \boldsymbol{\theta}^T(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}) + tr(\mathbf{Y}^T(\mathbf{G}D(\mathbf{x}) - \mathbf{J})) \\ + \quad & \frac{\mu}{2}(\|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}\|_2^2 + \|\mathbf{G}D(\mathbf{x}) - \mathbf{J}\|_F^2), \quad (8)\end{aligned}$$

where $\mathbf{Y} \in \Re^{m\times n}$ and $\boldsymbol{\theta} \in \Re^m$ are the lagrangian multipliers, $\mu > 0$ is the penalty parameter, $tr(\cdot)$ is the trace function, and $\|\cdot\|_F$ is the Frobenius norm.

Instead of optimizing all arguments simultaneously, as $\mathbf{J}, \mathbf{x}, \mathbf{e}$ are separable, we solve them individually and iteratively.

By fixing $\mathbf{x}$ and $\mathbf{e}$, we optimize $\mathbf{J}$ by the following subprolbem

$$\begin{aligned}&\min_{\mathbf{J}} \lambda\|\mathbf{J}\|_* + tr(\mathbf{Y}^T(\mathbf{G}D(\mathbf{x}) - \mathbf{J})) + \frac{\mu}{2}\|\mathbf{G}D(\mathbf{x}) - \mathbf{J}\|_F^2 \\ &= \frac{\lambda}{\mu}\|\mathbf{J}\|_* + \frac{1}{2}\|\mathbf{J} - (\mathbf{G}D(\mathbf{x}) + \mathbf{Y}/\mu)\|_F^2. \quad (9)\end{aligned}$$

Problem (9) can be solved by Singular Value Thresholding [27].

To update $\mathbf{x}$, the following subproblem is solved

$$\begin{aligned}&\min_{\mathbf{x}} \boldsymbol{\theta}^T(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}) + tr(\mathbf{Y}^T(\mathbf{G}D(\mathbf{x}) - \mathbf{J})) \\ &+ \frac{\mu}{2}(\|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}\|_2^2 + \|\mathbf{G}D(\mathbf{x}) - \mathbf{J}\|_F^2) \\ &= tr(\mathbf{Y}^T G(\mathbf{x}) - \boldsymbol{\theta}^T\mathbf{A}\mathbf{x} + \frac{\mu}{2}D(\mathbf{x})\mathbf{G}^T\mathbf{G}D(\mathbf{x}) \\ &-\mu\mathbf{J}^T\mathbf{G}D(\mathbf{x}) + \frac{\mu}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - \mu(\mathbf{y} - \mathbf{e})^T\mathbf{A}\mathbf{x}) \quad (10)\end{aligned}$$

The solution of (10) can be achieved via solving a linear system.

The optimized $\mathbf{e}$ can be obtained as

$$\begin{aligned}&\min_{\mathbf{e}} \|\mathbf{e}\|_1 + \boldsymbol{\theta}^T(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}\|_2^2 \\ &= \frac{1}{\mu}\|\mathbf{e}\|_1 + \frac{1}{2}\|\mathbf{e} - (\mathbf{y} - \mathbf{A}\mathbf{x} + \boldsymbol{\theta}/\mu)\|_2^2. \quad (11)\end{aligned}$$

Problem (11) can be solved by the soft-thresholding operator [28].

The lagrangian multipliers are updated as

$$\begin{aligned}\mathbf{Y} &= \mathbf{Y} + \mu(\mathbf{G}D(\mathbf{x}) - \mathbf{J}), \\ \boldsymbol{\theta} &= \boldsymbol{\theta} + \mu(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}). \quad (12)\end{aligned}$$

The steps (9)-(12) are repeated until the convergence conditions are attained. Algorithm 1 summarizes the procedures to solve the optimization problem.

Algorithm 1: Supervise Trace Lasso (STL)

Input: Matrixes of training sample set $\mathbf{A}$, query image $\mathbf{y}$, parameter λ.
Initialization:
$\mu = 10^{-3}, \mu_{max} = 10^6, \rho = 1.1, \epsilon = 10^{-3}$.
while *not converged* **do**
1. update the $\mathbf{J}$ as Problem (9)
2. update the $\mathbf{x}$ as Problem (10)
3. update the $\mathbf{e}$ as Problem (11)
4. update the multipliers
$$\mathbf{Y} = \mathbf{Y} + \mu(\mathbf{G}D(\mathbf{x}) - \mathbf{J})$$
$$\boldsymbol{\theta} = \boldsymbol{\theta} + \mu(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e})$$
5. update the parameter μ by $\mu = min(\rho\mu, \mu_{max})$
6. check the convergence conditions:
$\|\mathbf{J} - \mathbf{G}D(\mathbf{x})\|_\infty < \epsilon$ and $\|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}\|_\infty < \epsilon$
end
Output: Coefficient vector $\mathbf{x}$.

3.2. Classification

Once $\mathbf{x}$ is obtained, the next step is to classify the query sample. To be consistent with the measurement of the reconstruction error in problem (4), we use the l_1-norm. Therefore, the residual of each subject is calculated as

$$r(i) = \|\mathbf{y} - \mathbf{e} - \mathbf{A}\delta_i(\mathbf{x})\|_1, \quad (13)$$

where, $\delta_i(\mathbf{x})$ only retains the coefficients associated with the ith subject.

Finally, query sample is labeled to the subject with the minimum residual as following

$$i^* = \arg\min_i r(i). \quad (14)$$

Algorithm 2 below summarizes the complete recognition procedure.

Algorithm 2: Supervise Trace Lasso for Face Recognition

Input: Matrixes of training sample set $\mathbf{A}$, query image $\mathbf{y}$.
1. Solve the optimization problem as Algorithm 1.
2. Calculate the residual $r(i)$, for $i = 1, 2, \ldots, c$.
Output: The identity of $\mathbf{y} = \arg\min_i r(i)$.

4. EXPERIMENTS

In this section, the proposed framework is compared with the state-of-the-art approaches including SRC [12], SRC using l_1-norm to model the reconstruction error [12] (SRC-I), group lasso [29] and trace lasso [20]. We use two popular

Fig. 2: Face samples from the Extended Yale B Database.

face databases: Extended Yale B database [30] and AR Face Database [31]. For the trace lasso, once the coefficient vector is achieved, it uses the same classification procedure as SRC. The parameters for each method are finely tuned to achieve its best results. We directly utilize the gray level as feature in all experimental scenarios for all approches.

4.1. Extended Yale B Database

The cropped Extended Yale B Database includes 38 subject-s. There are about 64 face images per subject captured under different illuminations. Fig. 2 shows some samples with various lighting condition. The image size is downsampled from 198×168 to 48×42. We randomly select t images from each subject for training and rest images are used as testing set. For a fixed $t = 20, 15$, and 8, the averages recognition rates over 10 runs across various methods are reported in Table 1.

In Table 1, it can been seen that all extension of SRC (SRC-I, group lasso, trace lasso and STL) outperform SRC in different degrees. The proposed method achieves the best results with all number of training samples, While there are no method always achieving the second best consistently. In particular, the recognition rates of STL are 98.80%, 97.35%, and 88.07% for $t = 20, 15$, and 8, respectively. For $t = 20$, as each subject has enough training samples, all methods performance very well (e.g., at least 96.5%) and our method is slightly (1.09%) better than the second best one. As the number of training sample decreases to 15, the performances of all but the proposed method are lower than 95%. STL gets at least 2.44% gains over all other approaches. For $t = 8$, as the training samples cannot sufficiently cover the whole illumination variations, the recognition rates further degrades. However, our method still achieves around 2% improvement.

4.2. AR Face Database

AR database is consist of 126 subject. For each subject, 26 face images are taken in two separate sessions. Each session is with the expression, illumination, disguise variation. In this paper, a subset of 100 subject with only expression or illumination changes are used. Hence, there are total 1400 images. Some samples are shown in Fig. 3. All images are resized to 50×40. The same as the previous experiment, for each subject, t face images from Session 1 are used for training

Table 1: Face Recognition Rate on Extended Yale B Database.

number t	20	15	8
SRC[12]	96.51%	93.88%	84.94%
SRC-I[12]	97.71%	94.91%	85.37%
group lasso[29]	96.93%	94.37%	85.37%
trace lasso[20]	97.17%	94.15%	86.13%
Ours	**98.80%**	**97.35%**	**88.07%**

Fig. 3: Face samples from the AR database.

and all samples from Session 2 are used for testing. With each $t = 6, 5, 4$, we randomly select 10 times. Table 2 details the performances of different methods with different training samples.

Table 2: Face Recognition Rate on AR Face Database.

number t	7	6	5	4
SRC[12]	92.82%	91.71%	88.76%	87.5%
SRC-I[12]	93.28%	92.13%	89.13%	87.55%
group lasso[29]	92.99%	92.13%	91.56%	89.13%
trace lasso[20]	92.56%	91.70%	91.13%	89.56%
Ours	**96.42%**	**95.14%**	**94.71%**	**91.56%**

In this dataset, the results are similar to previous experiment. The performances of all approaches decrease as the number of training sample decreases. Our method achieves the best recognition rates in all scenarios. Especially, STL obtains 3.14%, 3.01%, 3.15%, and 2.00% gains over the second best method for $t = 7, 6, 5$, and 4, respectively.

5. CONCLUSION

Trace lasso tackles the instability problem of SRC based on the correlation of training data. It naturally clusters the highly correlated data together. However, the face images with similar variations, such as illumination or expression, often have higher correlation than those from the same class. In this case, the result of trace lasso is contradictory to the goal of identification, which is to cluster the samples according to their identities. In this paper, we propose a supervised trace lasso (STL) framework to address the face recognition problem. By cooperating the class label information with trace lasso, the proposed STL approach seeks the sparsity of the number of classes instead of the number of training samples. Compared with SRC and trace lasso, which are unsupervised methods, our method can achieve a more discriminative representation for classification task. Furthermore, an efficient algorithm to solve the optimization problem of proposed method is given. The extensive experimental results have demonstrated the effectiveness of the propose framework.

6. REFERENCES

[1] A.K. Jain, R.P.W. Duin, and Jianchang Mao, "Statistical pattern recognition: a review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 1, pp. 4 –37, Jan. 2000.

[2] X. D. Jiang, "Linear subspace learning-based dimensionality reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16–26, Mar. 2011.

[3] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[4] X. D. Jiang, "Asymmetric principal component and discriminant analyses for pattern classification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 5, pp. 931–937, May 2009.

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 711 –720, July 1997.

[6] X. D. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 3, pp. 383–394, Mar. 2008.

[7] X. F. He, S. C. Yan, Y. X. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 3, pp. 328 –340, Mar. 2005.

[8] S. Z. Li and J. W. Lu, "Face recognition using the nearest feature line method," *IEEE Trans. Neural Networks*, vol. 10, no. 2, pp. 439 –443, Mar. 1999.

[9] J. T. Chien and C. C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 12, pp. 1644 – 1649, Dec. 2002.

[10] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 11, pp. 2106–2112, Nov. 2010.

[11] J. Lai and X.D. Jiang, "Robust face recognition using trimmed linear regression," in *ICASSP*, 2013, pp. 2979–2983.

[12] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[13] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ?1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[14] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[15] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, Dec. 2006.

[16] L.S. Qiao, S.C. Chen, and X.Y. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognition*, vol. 43, no. 1, pp. 331 – 341, 2010.

[17] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized robust coding for face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1753–1766, May 2013.

[18] J. Lai and X.D. Jiang, "Modular weighted global sparse representation for robust face recognition," *IEEE Signal Processing Letters*, vol. 19, no. 9, pp. 571–574, Sept 2012.

[19] J. Lai and X.D. Jiang, "Discriminative sparsity preserving embedding for face recognition," in *ICIP*, 2013, pp. 3695–3699.

[20] E. Grave, G. Obozinski, and F. Bach, "trace lasso a trace norm regularization for correlated designs," in *NIPS*, 2012.

[21] K. Koh, S.J. Kim, and S. Boyd, "An interior-point method for large-scale l1-regularized logistic regression," *J. Mach. Learn. Res.*, vol. 8, pp. 1519–1555, Dec. 2007.

[22] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[23] A.Y. Yang, Zihan Zhou, A.G. Balasubramanian, S.S. Sastry, and Yi Ma, "Fast l_1-minimization algorithms for robust face recognition," *Image Processing, IEEE Transactions on*, vol. 22, no. 8, pp. 3234–3246, 2013.

[24] C.Y. Lu, J.S. Feng, Z.C. Lin, and S.C. Yan, "Correlation adaptive subspace segmentation by trace lasso," in *ICCV*, 2013, vol. 2.

[25] E. J. Candès, X. D. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, June 2011.

[26] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, 1996.

[27] J. Cai, E. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[28] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[29] A. Majumdar and R.K. Ward, "Classification via group sparsity promoting regularization," in *ICASSP*, 2009, pp. 861–864.

[30] K. C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[31] A. Martinez and R. Benavente, "The ar face database," *CVC Technical Report*, , no. 24, June 1998.