General Regression and Representation Model for Face Recognition

Jianjun Qian, Jian Yang
School of Computer Science and Engineering
Nanjing University of Science and Technology, Nanjing, China
qjjtx@126.com, csjyang@njust.edu.cn

Abstract

Recently, the regularized coding-based classification method (e.g. SRC and CRC) shows a great potential for face recognition. However, most existing coding methods ignore the statistical information from the training data, which actually plays an important role in classification. To address this problem, we develop a general regression and representation model (GRR) for classification. GRR not only has advantages of CRC, but also introduces the prior information and the specific information to enhance the classification performance. In GRR, we combine the leave-one-out strategy with K Nearest Neighbors to learn the prior information from the training data. Meanwhile, the specific information is obtained by using the iterative algorithm to update the feature weights of the test sample. Finally, we classify the test sample based on the reconstruction error of each class. The proposed model is evaluated on public face image databases. And the experimental results demonstrate the advantages of GRR over state-of-the-art methods.

1. Introduction

Linear regression has been widely applied to pattern classification. To prevent overfitting, the L_2-regularizer is generally used in the linear regression model. In the past years, the L_1-regularizer, which is closely linked to sparse representation, becomes a hot theme in information theory, signal/image processing and related areas. Meanwhile, numerous findings of neuroscience and biology form a physiological base for sparse representation [1-3].

Recently, many efforts have been made to apply sparse representation methods to pattern classification tasks, including signal/image classification and face recognition etc. Labusch et al. presented a simple sparse-coding strategy for digit recognition and achieved state-of-the-art results on the MNIST benchmark [4]. J.C. Yang et al. addressed the problem of generating a super-resolution (SR) image from a single low-resolution input image via sparse representation [5]. J. Mairal et al. elaborated a framework for learning multi-scale sparse representations of images with applications to image denoising and inpainting [6]. J.C. Yang et al. employed sparse coding instead of vector quantization to capture the significant properties of local image descriptors for image classification [7]. Particularly, Wright et al. introduced a sparse representation based classification (SRC) and successfully applied it to identify human faces with varying illumination changes, occlusion and real disguise [8]. A test sample image is coded as a sparse linear combination of the training images, and then the classification is achieved by identifying which class yields the least residual. Subsequently, M. Yang and L. Zhang constructed a Gabor occlusion dictionary for SRC to reduce the computation cost by using Gabor feature [9]. Although the newly-emerging SRC shows great potential for pattern classification, it lacks theoretical justification. J. Yang et al. provided an insight into SRC and analyzed the role of L_1-optimizer [10]. They thought that L_1-optimizer contains two properties sparsity and closeness. However, L_0-optimizer can only achieve the sparsity. Sparstiy determines a small number of nonzero representation coefficients and closeness makes the nonzero representation coefficients concentrate on the training samples with the same class label as the given test sample. Wright et al. give a recent review of sparse representation for computer vision and pattern recognition [11]. In addition, many related tasks have been reported [12-16].

With the widely use of sparse representation for classification, some scholars question the role of sparseness for image classification [17, 18]. L. Zhang et al. analyzed the working principle of SRC and believed that it is the collaborative representation that improves the image classification accuracy rather than the L_1-norm sparstiy. Consequently, L. Zhang et al. presented a collaborative representation based classification with regularized least square (CRC) [19]. Compared with SRC, CRC delivers very competitive classification results with little computation time. Subsequently, M. Yang et al. proposed a relaxed collaborative representation model (RCR) which effectively captures the similarity and distinctiveness of different features for pattern classification [20].

Most of the previous works focus on investigating the importance of the L_1-regularizer/L_2-regularizer. However,

these works ignore some statistical information hidden in the training data. This paper aims to explore the prior information (learned from the training set offline) and the specific information (learning from the testing sample online) so as to enhance the classification performance under different conditions. To this end, we propose a model named General Regression and Representation (GRR) for pattern classification. The overview of GRR is shown in Fig. 1. Compared to other classification method, the novelty of the proposed model is twofold:

- First, GRR captures the prior information from the training set via the generalized Tikhonov regularization in conjunction with the leave-one-out strategy and K Nearest Neighbor method;
- Second, we propose two models of GRR: Basic GRR (B-GRR) and Robust GRR (R-GRR) by combining the prior information and the specific information with different strategies;

To evaluate the proposed method, we perform experiments on the AR and the Extended Yale B databases, and the experimental results demonstrate the effectiveness and robustness of the proposed GRR.

2. General Regression and Representation Model for Classification (GRR)

In this study, GRR contains two models: Basic GRR and Robust GRR. We will introduce these two models in detail as follows.

2.1. Basic GRR

Most of current works don't make full use of the statistical information of training set. To address this problem, we introduce the concept of prior information and propose a basic general regression and representation model for classification. Specifically, let **A** be the matrix formed by the K nearest neighbors of the test sample from training set and y be the test sample. **P** is the weight matrix of reconstruction errors, and **Q** is the weight matrix of representation coefficients. **P** and **Q** are matrices containing the prior information. Our model is

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_{\mathbf{P}}^2 + \|\mathbf{x}\|_{\mathbf{Q}}^2 \tag{1}$$

We call the above model as the basic general regression and representation (B-GRR). Actually, this model can be reformulated as follows:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} (\mathbf{y} - \mathbf{A}\mathbf{x})^T \mathbf{P}(\mathbf{y} - \mathbf{A}\mathbf{x}) + \mathbf{x}^T\mathbf{Q}\mathbf{x} \tag{2}$$

If **P** and **Q** are known, from [26], we know there is a close-form solution:

$$\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{P}\mathbf{A} + \mathbf{Q})^{-1}\mathbf{A}^T\mathbf{P}\mathbf{y} \tag{3}$$

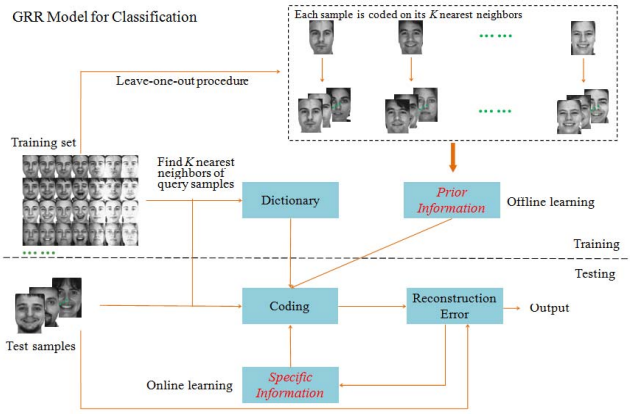

Figure 1: An overview of General Regression and Representation model for classification

However, **P** and **Q** are unknown beforehand. The remaining problem is how to learn **P** and **Q**. Here, we employ a generative method to evaluate these two matrices. Specifically, we use the leave-one-out strategy in conjunction with K Nearest Neighbor to learn the prior information matrices **P** and **Q**. Denote by $\mathbf{y}_i^{(t)}$ the *i*-th sample of training set. $\mathbf{A}_i^{(t)}$ is the *K* nearest neighbors of $\mathbf{y}_i^{(t)}$ from the training set. We can initialize $\mathbf{P}_0$, $\mathbf{Q}_0$ as $\mathbf{P}_0$=**I** and $\mathbf{Q}_0$=**I**. The $\mathbf{y}_i^{(t)}$ is coded on $\mathbf{A}_i^{(t)}$ as follows:

$$\hat{\mathbf{x}}_i^{(t)} = \arg\min_{\mathbf{x}_i^{(t)}} \|\mathbf{y}_i^{(t)} - \mathbf{A}_i^{(t)}\mathbf{x}_i^{(t)}\|_{\mathbf{P}_0}^2 + \|\mathbf{x}_i^{(t)}\|_{\mathbf{Q}_0}^2 \tag{4}$$

Then, **P** and **Q** can be directly derived by using Eq. (5) and Eq. (6) respectively:

$$\mathbf{P} = \{\mathrm{cov}(\mathbf{E}) + \lambda_1\mathbf{I}\}^{-1} \tag{5}$$

where $\mathbf{E}=[\boldsymbol{e}_1,\cdots,\boldsymbol{e}_i\cdots,\boldsymbol{e}_K]$ and $\boldsymbol{e}_i = \|\mathbf{y}_i^{(t)} - \mathbf{A}_i^{(t)}\hat{\mathbf{x}}_i^{(t)}\|^2$. Here, λ_1 is the regular parameter, and cov(·) is an operator to compute the covariance matrix.

$$\mathbf{Q} = \{\mathrm{cov}(\mathbf{X}) + \lambda_2\mathbf{I}\}^{-1} \tag{6}$$

where $\mathbf{X}=[\hat{\mathrm{x}}_1^{(t)},\cdots,\hat{\mathrm{x}}_i^{(t)}\cdots,\hat{\mathrm{x}}_K^{(t)}]$. λ_2 is the regular parameter.

In the testing stage, the solution of $\hat{\mathbf{x}}$ in Eq. (1) is easily derived by using Eq. (3). We can reconstruct the test sample **y** as $\hat{\mathbf{y}}_i = \mathbf{A}\delta_i(\hat{\mathbf{x}})$ by employing the coefficients associated with *i*-th class. The corresponding reconstruction error of *i*-th class is defined:

$$\mathrm{r}_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}\delta_i(\hat{\mathbf{x}})\|_2 / \|\delta_i(\hat{\mathbf{x}})\|_2 \tag{7}$$

The decision rule is: if $\mathrm{r}_l(\mathbf{y}) = \min_i \mathrm{r}_i(\mathbf{y})$, y is assigned to Class *l*.

Algorithm 1 B-GRR

Input: Dictionary **A**, test sample **y**. Initial values $\mathbf{P}_0$and $\mathbf{Q}_0$

1. Normalize the columns of **A** to have unit L_2-norm.
2. The prior information matrices **P** and **Q** are learned from training set by using the generalized Tikhonov regularization, leave-one-out strategy and KNN.
3. The test sample **y** is coded on its *K* nearest neighbors **A** via Eq. (1).
4. Compute the residuals of each class.

Output: **y** is assigned to the class which yields the minimum residual.

B-GRR makes full use of the prior information of the training set. It works well when the testing samples share the same probability distribution with the training samples. The algorithm of B-GRR for classification is summarized in Algorithm 1.

2.2. Robust GRR

In face recognition problems, illumination, expression or pose changes may cause differences between test samples and training samples. Therefore, it is necessary to introduce the specific information of the test sample to alleviate the effect caused by the differences between test samples and training samples. This specific information is to give a weight to each image pixel, which can be learned online via the iteratively reweighted method, as adopted in RSC [12].

Based on this idea, we present a robust general regression and representation model (R-GRR) for classification. Compared with B-GRR, R-GRR not only includes the prior information matrices **P** and **Q**, but also owns the specific information matrix **W**. The model is given below:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left\| \mathbf{W}^{1/2}(\mathbf{y} - \mathbf{A}\mathbf{x}) \right\|_{\mathbf{P}}^2 + \left\| \mathbf{x} \right\|_{\mathbf{Q}}^2 \tag{8}$$

If **P**, **Q** and **W** are known, the above model has a close-form solution:

$$\hat{\mathbf{x}} = [\mathbf{A}^{\mathrm{T}}(\mathbf{W}^{1/2})^{T}\mathbf{P}(\mathbf{W}^{1/2})\mathbf{A} + \mathbf{Q}]^{-1}\mathbf{A}^{\mathrm{T}}(\mathbf{W}^{1/2})^{T}\mathbf{P}(\mathbf{W}^{1/2})\mathbf{y} \tag{9}$$

Since **P** and **Q** can be learned offline using the same method as in Basic GRR, the remaining problem is to learn the specific information (matrix **W**) online. Specifically, given a test sample **y**, we firstly compute the coding residuals $\boldsymbol{e}$ of **y** so as to initialize the weight. The residual $\boldsymbol{e}$ is initialized as $\boldsymbol{e}$=**y**-$\mathbf{y}_{ini}$, and $\mathbf{y}_{ini}$ is the initial estimation of the true images from the observe samples. In this study, we simply set $\mathbf{y}_{ini}$ as the mean image of all samples in the coding dictionary since we don't know which class the test image **y** belongs to. With the initialized $\mathbf{y}_{ini}$, our method can estimate the **W** iteratively. **W** actually is a diagonal matrix, $\mathbf{W}_{i,i}$ (i.e. $\omega_\theta(e_i)$) is the weight assigned to each pixel of the test image. The weight function [12] is:

$$\omega_\theta(e_i) = \frac{\exp(\alpha\beta - \alpha(e_i)^2)}{1 + \exp(\alpha\beta - \alpha(e_i)^2)} \tag{10}$$

where, α and β are positive scalars.

In addition, Eq. (9) is the explicit solution of Eq. (8). The convergence is achieved when the difference of the weights between adjacent iterations satisfies the following condition:

$$\left\| \mathbf{W}^{(t)} - \mathbf{W}^{(t-1)} \right\|_2 / \left\| \mathbf{W}^{(t-1)} \right\| < \varepsilon \tag{11}$$

where, ε is a small positive value.

The R-GRR algorithm for classification is summarized in Algorithm 2.

Algorithm 2 R-GRR

Input: Dictionary **A**, test sample **y**. Initial values $\mathbf{P}_0$, $\mathbf{Q}_0$ and $\mathbf{y}_{ini}$.

1. Normalize the columns of **A** to have unit L_2-norm, test sample y with L_2-norm and y^t initialized as y_{ini}.
2. The prior information matrices **P** and **Q** are learned from the training set by using the generalized Tikhonov regularization, leave-one-out strategy and KNN.
3. The test sample **y** is coded on its *K* nearest neighbors **A**.
   a) Compute residual $\boldsymbol{e}^{(t)} = \mathbf{y} - \mathbf{y}^{(t)}$
   b) Estimate weights
   $$\omega_\theta(e_i^{(t)}) = \frac{\exp(\alpha^{(t)}\beta^{(t)} - \alpha^{(t)}(e_i^{(t)})^2)}{1 + \exp(\alpha^{(t)}\beta^{(t)} - \alpha^{(t)}(e_i^{(t)})^2)}$$
   c) Code $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left\| (\mathbf{W}^{(t)})^{1/2}(\mathbf{y} - \mathbf{A}\mathbf{x}) \right\|_{\mathbf{P}}^2 + \left\| \mathbf{x} \right\|_{\mathbf{Q}}^2$
   d) Compute the reconstructed test sample $\mathbf{y}^{(t)} = \mathbf{A}\mathbf{x}^{(t)}$, and let t = t + 1
   e) Go back to step a) until the maximal number of iterations is reached, or convergence is met as shown in Eq. (11)
4. Compute the residuals of each class.

Output: **y** is assigned to the class which yields the minimum residual.

3. Further Analysis of GRR

In this section, we will further analysis the role of **P**, **Q** and **W** in GRR. **P** (a symmetric matrix) is the weight matrix of reconstruction errors and learned from the training set. The diagonal elements of matrix **P** stand for the importance of image pixels or features. The non-diagonal elements represent the cross relationship between image pixels or

Figure 2: Two classes of samples from the AR database

Test image R-GRR RSC B-GRR CRC

(a) Recovered clean image and occluded part via four methods for the image with sunglasses

Test image R-GRR RSC B-GRR CRC

(b) Recovered clean image and occluded part via four methods for the image with scarf

Figure 3: The advantages of R-GRR

features. The matrix **Q** can be considered as a regularization term. In general, the regularization term is determined manually. Here, **Q** is learned from the training set and represents the relationship between samples in the coding dictionary.

We also give an example to show the advantages of Robust GRR (R-GRR) as shown in Fig. 3. In our example, two classes of face images from the AR database, as shown in Fig. 2, are used for training. We test two cases of real-world disguise images: the images with sunglasses and the images with scarves. In Fig. 3 (a) and Fig. 3 (b), the left column contains the disguise images. In our test, we use R-GRR, RSC, B-GRR and CRC to deal with occlusion. For each occluded image, the reconstructed images (recovered clean image) and the residual images (recovered occlusion) are shown in Fig. 3. From Fig. 3, we can see that R-GRR achieves comparable result with RSC and significantly outperforms other methods.

4. Experimental Results

In this section, we perform experiments on public face image databases and compare the proposed model GRR with state-of-the-art methods. Note that here in SRC and RSC, the matlab function "l1-ls" [21] is used to calculate the sparse representation coefficients. In the following experiment, the parameters α and β of Eq. (10) are determined by the rule in paper [12]. λ_1 and λ_2 are set to 0.1, 10^{-7}, respectively. The parameter K is determined by experiments. Fig. 7 plots the recognition rates versus the variation of the parameter K on the different experiments.

Dim	54	120	300
NN	68.0	70.1	71.3
LRC	71.0	75.4	76.0
SRC	83.3	89.5	93.3
CRC[19]	80.5	90.0	93.7
RSC[12]	***86.8***	94.0	96.0
B-GRR	81.3	90.4	93.6
R-GRR	85.6	***95.3***	***97.3***

Table 1. The recognition rates (%) of each classifier for face recognition on the AR database

Dim	50	100	200
NN	78.5	85.8	90.2
LRC	93.3	94.8	95.2
SRC	93.7	94.7	95.6
CRC	91.9	94.7	96.5
RSC	94.2	97.0	98.2
B-GRR	92.1	94.9	97.1
R-GRR	***94.3***	***97.5***	***98.4***

Table 2. The recognition rates (%) of each classifier for face recognition on the Extended Yale B database

4.1. Face Recognition without Occlusion

We evaluate the performance of B-GRR and R-GRR on the AR and the Extended Yale B database with illumination and expression changes but without occlusion. In these experiments, PCA is first used to reduce the dimensionality of face image.

AR Database

The AR face database [22] contains over 4000 color face images of 126 persons, including frontal views of faces with different facial expression, lighting conditions and occlusions. The pictures of 120 individuals were taken in two sessions (separated by two weeks) and each session contains 13 color images. Fourteen face images (each session contains 7) of 100 individuals are selected and used in our experiment. The face portion of each image is manually cropped and then normalized to 60×43 pixels.

In this experiment, images from the first session are used for training, and images from the second session are used for testing. Then LRC (linear regression classification [14]), SRC, CRC, RSC, B-GRR and R-GRR are employed for classification. The NN classifier is also used to provide a

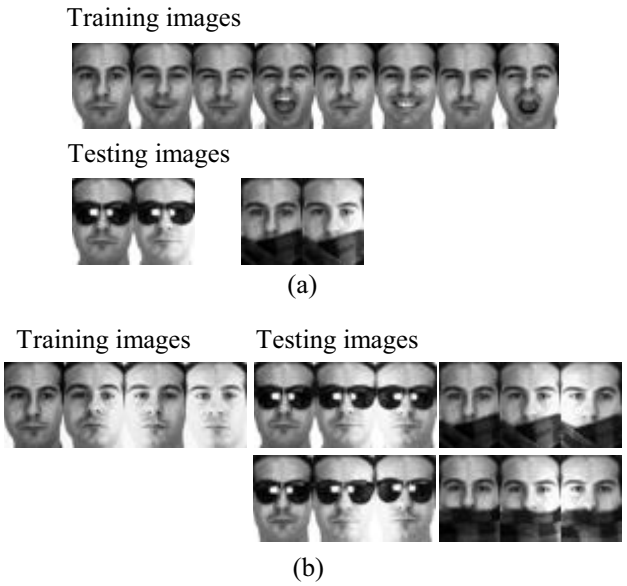

Figure 4: Sample images for one person of AR database. (a) Sample images of the first experiment. (b) Sample images of the Second experiment.

baseline. The parameter K of R-GRR means we choose the K nearest neighbors of test image from training set to form the coding dictionary. K is set to 650 here. The recognition rates of each classifier versus the variation of dimensions are listed in Table 1. From Table 1, we can see that R-GRR gives better performance than state-of-the-art methods in all dimensions except that R-GRR is slightly worse than RSC when dimension is 54. However, it's difficult to achieve better performance when dimension is low for all the methods. The maximal recognition rates of NN, LRC, SRC, CRC, RSC, B-GRR and R-GRR are achieved when the dimension is 300.

Extended Yale B Database

The extended Yale B face image database [23] contains 38 human subjects under 9 poses and 64 illumination conditions. The 64 images of a subject in a particular pose are acquired at camera frame rate of 30 frames / second, so there are only small changes in head pose and facial expression for those 64 images. All frontal-face images marked with P00 are used in our experiment, and each is resized to 48×42 pixels.

In our experiment, we use the first 32 images of each individual for training and the remaining images are used for testing. Based on the PCA-transformed features, NN, LRC, SRC, CRC, RSC, B-GRR and R-GRR are employed for classification. The parameter K is 800. The recognition rates of each classifier corresponding to the variation of feature dimensions are listed in Table 2. Table 2 shows that the proposed R-GRR achieves the best recognition rates in all dimensions for face recognition. When the feature dimension is 100, R-GRR gives about 3% improvement of recognition rate over LRC, SRC and CRC, respectively.

Methods	Sunglasses	Scarves
CRC	65.5	88.5
SRC[8]	87.0	59.5
GSRC[9]	93.0	79.0
CESR[15]	99.0	42.0
RSC[12]	99.0	97.0
R-GRR	***99.5***	***99.0***

Table 3. The recognition rates (%) of each classifier for face recognition on AR database with disguise occlusion

4.2. Face Recognition with Occlusion

In this section, we examine the robustness of R-GRR when face images suffer different occlusions, such as real disguise or block occlusion. Here, R-GRR combines advantages of the prior information **Q** and the specific information **W** to enhance the classification performance and set **P** as unit matrix. As we know, **P** reflects the probability distribution of reconstruction errors. If there are great differences between the test sample and the training samples, the resulting reconstruction error does not follow the original distribution. In this case, using **P** may cause negative effect on the classification performance. In the following experiments, we mainly compared our methods with CRC, SRC, RSC, correntropy-based sparse representation (CESR) [15] and Gabor-SRC [9].

Face Recognition with Real Disguise

A subset of AR face image database is used in our experiment. The subset contains 100 individuals, 50 males and 50 females. All the individuals have two session images and each session contains 13 images. The face portion of each image is manually cropped and then normalized to 42×30 pixels.

In the first experiment, we choose the first four images (with various facial expressions) from the session 1 and session 2 of each individual to form the training set. The total training images is 800. There are two image sets (with sunglasses and scarves) for testing. Each set contains 200 images (one image per session of each individual with neutral expression). The sample images of one person as shown in Fig. 4 (a). The parameter K is 300 for the test set with sunglasses and 760 for the test set with scarves. The face recognition results of each method on the two testing set are listed in Table 3. From Table 3, we can see that R-GRR achieves the best recognition results among all the methods. Moreover, the performances of RSC and CESR

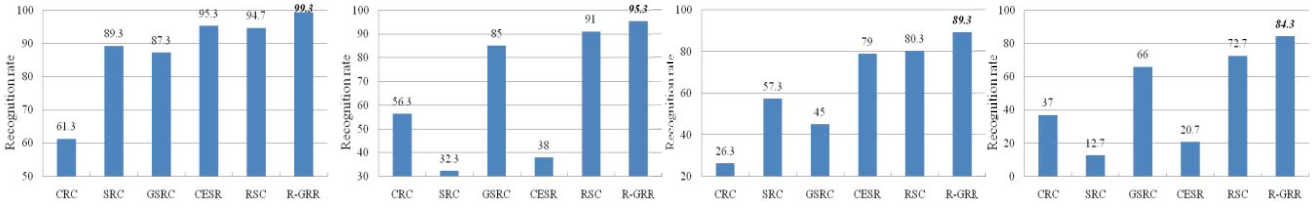

Figure 5: The recognition rates (%) of each classifier for face recognition on AR database with disguise occlusion. (a) The testing images with sunglasses from session 1. (b) The testing images with scarves from session 1. (c) The testing images with sunglasses from session 2. (d) The testing images with scarves from session 2.

are both higher when facial image with sunglasses. However, CESR only achieves 42% when facial images with scarves.

In the second experiment, four neutral images with different illumination from the first session of each individual are used for training. The disguise images with various illumination and glasses or scarves per individual in session 1 and session 2 for testing. The sample images of one person as shown in Fig. 4 (b). We set the parameter K as 220, 300, 240 and 320 for the four test set, respectively. The recognition results of each method are shown in Fig. 5. From Fig. 5, we can see clearly that R-GRR gives better performance than CRC, SRC, GSRC, CESR and RSC on different testing subsets. Both SRC and CESR do well on the subsets with sunglasses but poor in the cases with scarves. However, GSRC achieves better result on the subsets with scarves and worse result on the subsets with sunglasses. Compared to RSC, at least 4.3% improvement is achieved by R-GRR for each testing set. Meanwhile, it is worth noticing that the recognition rate of R-GRR is 71.6%, 63.6% higher than SRC and CESR on the testing images with scarves from session 2, and 44.3% higher than GSRC on the testing images with sunglasses from session 2.

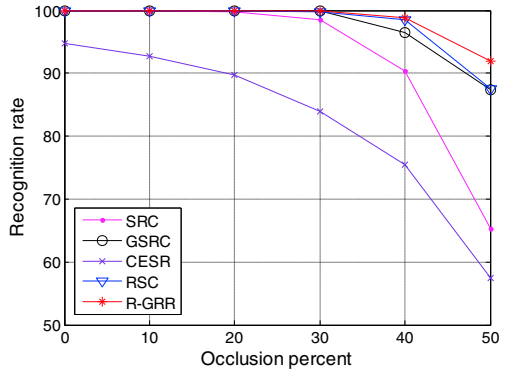

Figure 6: The recognition rates (%) of SRC, GSRC, CESR, RSC and R-GRR under the occlusion percentage from 0 to 50

Face Recognition with Block Occlusion

In this experiment, we use the same experiment setting as in [8, 12] to test the robustness of R-GRR. Subsets 1 and 2 of Extended Yale B are used for training and subset 3 is used for testing. The face images are resized to 96×84. The parameter K is 200. Fig. 6 shows recognition rates curve of SRC, GSRC, CESR, RSC and R-GRR versus the various levels of occlusion (from 0 percent to 50 percent). From Fig. 6, we can see that the proposed R-GRR overall outperforms SRC, GSRC, CESR and RSC. When the occlusion percentage is 50%, GRR achieves the best recognition rate 91.9, compared to 65.3 for SRC, 87.4 for GSRC, 57.4 for CESR, and 87.6 for RSC. It's surprising that the performance of CESR is very poor. Probably, it is not suit to deal with this block occlusion case.

5. Conclusions

This paper presents a general regression and representation (GRR) model for face recognition. In GRR, we learn the prior information from the training set by combining the leave-one-out strategy and KNN in the framework of generalized Tikhonov regularization. Also, we learn the specific information from the test sample by using the iteratively reweighted algorithm. Actually, we provide two models: B-GRR and R-GRR, which combine the prior information and the specific information using different strategies. Experiments on face datasets demonstrated that the validity of our model and its performance advantages over state-of-the-art classification methods.

References

[1] W. E. Vinje, J. L. Gallant, Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision, Science, 2000, Vol. 287. no. 5456, pp. 1273 – 1276

[2] B. A. Olshausen, D. J. Field, Sparse coding of sensory inputs, Current Opinion in Neurobiology, 2004, Vol. 14, No. 4., pp. 481-487.

[3] T. Serre, "Learning a Dictionary of Shape-Components in Visual Cortex: Comparison with Neurons, Humans and Machines," PhD dissertation, MIT, 2006.

[4] Zhou, H.; Hastie, T.; and Tibshirani, R. 2004. Sparse principle component analysis. Technical Report, Statistics Department, Stanford University.

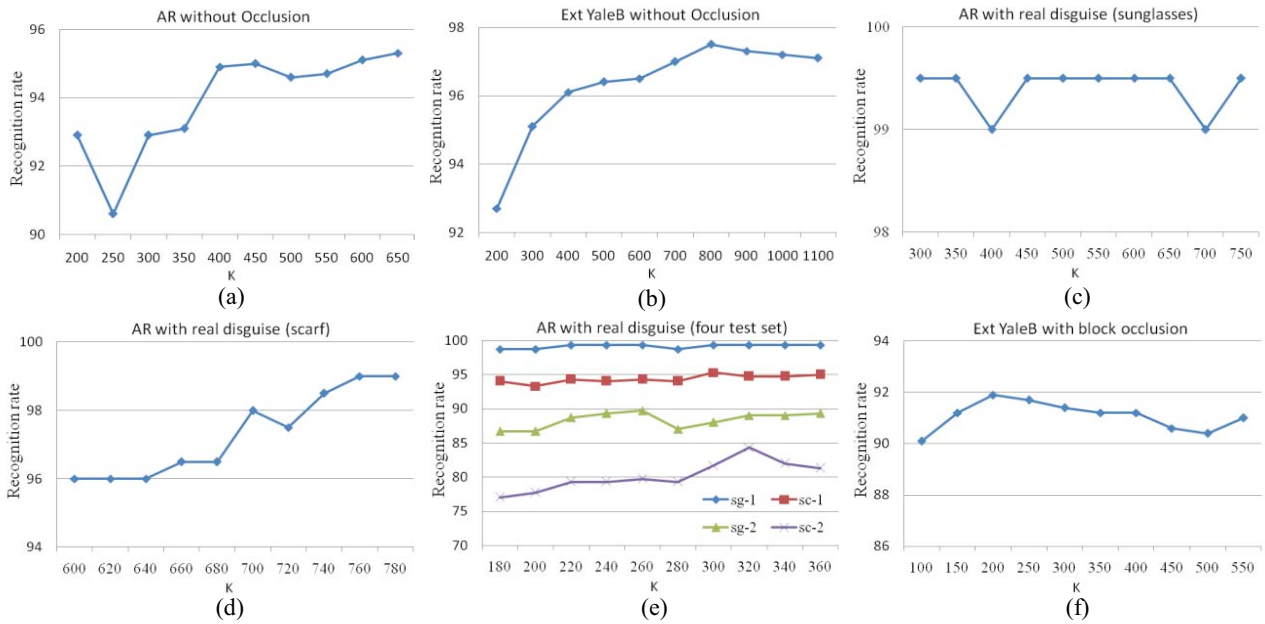

Figure 7: The recognition rate curves of R-GRR versus the variation of parameter K on the different experiments. (a) the images without occlusion for test; (b) the images without occlusion for test; (c) the images with sunglasses for test; (d) the images with scarf for test; (e)the images with sunglasses (sg-X) or scarf (sc-X) in session X for test; (f) the images with block occlusion (50%) for test.

[5] J. C. Yang, John Wright, Thomas Huang, and Yi Ma, Image Super-resolution as Sparse Representation of Raw Image Patches, In *CVPR*, 2008

[6] J. Mairal, G. Sapiro, and M. Elad. "Learning multiscale sparse representations for image and video restoration", *SIAM MMS, 7(1):214–241,April 2008*.

[7] J. Yang, K. Yu, Y. Gong and T. Huang. Linear spatial pyramid matching using sparse coding for image classification, In *CVPR* 2009.

[8] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE PAMI*, 31(2):210–227, 2009.

[9] M. Yang and L. Zhang. Gabor Feature based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary. In *ECCV*, 2010.

[10] J. Yang, L. Zhang, Y. Xu and J.Y. Yang, Beyond Sparsity: the Role of L1-optimizer in Pattern Classification, *Pattern Recognition*, 45 (2012) 1104–1118

[11] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. Proceedings of IEEE, Special Issue on Applications of Compressive Sensing & Sparse Representation, 98(6):1031-1044, 2010.

[12] M. Yang, L. Zhang, J. Yang and D. Zhang, "Robust sparse coding for face recognition," In *CVPR*, 2011.

[13] A. Wagner, J. Wright, A. Ganesh, Z.H. Zhou, and Y. Ma, Towards a practical face recognition system: robust registration and illumination by sparse representation. In *CVPR* 2009.

[14] I. Naseem, R. Togneri, and M. Bennamoun. Linear regression for face recognition. *IEEE PAMI*, 32(11):2106-2112, 2010.

[15] R. He, W.S. Zheng, B.G. Hu, and X.W. Kong, A regularized correntropy framework for robust pattern recognition, *Neural Computation*, vol. 23, pp. 2074-2100, 2011.

[16] R. He, W.S. Zheng, and B.G. Hu, Maximum correntropy criterion for robust face recognition, *IEEE PAMI*, vol. 33, no. 8, pp. 1561-1576, 2011.

[17] R. Rigamonti, M. Brown and V. Lepetit. Are Sparse Representations Really Relevant for Image Classification? In *CVPR* 2011.

[18] Q. Shi, A. Eriksson, A. Hengel, C. Shen. Is face recognition really a compressive sensing problem? In *CVPR* 2011.

[19] L. Zhang, M. Yang, and X. C. Feng. Sparse representation or collaborative representation which helps face recognition? In *ICCV*, 2011.

[20] M. Yang, L. Zhang, D. Zhang and S.L Wang, Relaxed Collaborative Representation for Pattern Classification, In *CVPR*, 2012

[21] S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. A interior-point method for large-scale l1 -regularized least squares. *IEEE Journal on Selected Topics in Signal Processing*, 1(4):606–617, 2007.

[22] A. Martinez and R. benavente. The AR face database. Tech-nical Report 24, CVC, 1998.

[23] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE PAMI*, 27(5):684–698, 2005.

[24] S.X. Liao, M. Pawlak, On image analysis by moments, *IEEE PAMI.,* 1996, 18(3), 254-266.

[25] Y.H. Tseng, C.C. Kuo and H.J. Lee, Speeding Up Chinese Character Recognition in An Automatic Document Reading System, *Pattern Recognition*, 31(11) (1998) 1601-1612.

[26] http://en.wikipedia.org/wiki/Ridge_regress