Multi-illumination Face Recognition from a Single Training Image per Person with Sparse Representation

Die Hu, Li Song, and Cheng Zhi

Institute of Image Communication and Information Processing,
Department of Electronic Engineering, Shanghai Jiao Tong University,
Shanghai, 200240, China
{butterflybubble,song_li,zhicheng}@sjtu.edu.cn

Abstract. In real-world face recognition systems, traditional face recognition algorithms often fail in the case of insufficient training samples. Recently, the face recognition algorithms of sparse representation have achieved promising results even in the presence of corruption or occlusion. However a large over-complete and elaborately designed discriminant training set is still required to form sparse representation, which seems impractical in the single training image per person problems. In this paper, we extend Sparse Representation Classification (SRC) to the one sample per person problem. We address this problem under variant lighting conditions by introducing relighting methods to generate virtual faces. Our diverse and complete training set can be well composed, which makes SRC more general. Moreover, we verify the recognition under different lighting environments by a cross-database comparison.

1 Introduction

In the past decades, the one sample per person problem has aroused significant attention in Face Recognition Technology (FRT) due to the wide applications such as law scenarios and passport identifications. It is defined as: given a stored database of faces with only one image per person, the goal is to identify a person from the database later in time in any different and unpredictable poses, lighting, etc from an individual image [1]. However, when there is only one sample per person available, many difficulties, i.e. reliability of parameter estimation, may arise when directly putting traditional recognition methods into use.

Recently, many research efforts have been focusing on applying the sparse representation to computer vision tasks since sparse signal representation has proven to be an extremely powerful tool for acquiring, representing, and compressing high-dimensional signals [2]. In FRT, Wright et al. [3] has cast face recognition as the problem of finding the sparsest representation of the test image over the training set, which have demonstrated promising results even in the presence of corruption and occlusion. However they have to provide a large over-complete and elaborately designed discriminant training set to form sparse

R. Kimmel, R. Klette, and A. Sugimoto (Eds.): ACCV 2010, Part II, LNCS 6493, pp. 672–682, 2011.

representation. It's beyond doubt that this constraint provokes a great challenge in one sample per person problem. Details will be discussed in Section 3.

In this paper, we propose an improved method of SRC in the one sample per person recognition problem under different lighting environments. Different from the traditional face recognition methods focused on illumination compensation, we make full use of the diversity representative illumination information to generate novel realistic faces. Given a single image input, we could synthesize the image space of it, so that the test image could form an efficient sparse representation when projected onto the constructed space. By doing that, we could not only satisfy the over-complete constraint in one sample per person problem, but also guarantee the discrimination capability of the training set in terms of selecting varying illumination conditions. We conduct extensive experiments on public databases, and especially, verify the recognition under different lighting environments by a cross-database comparison, of which the classical illumination record doesn't come from the same database as the images to be recognized.'

2 Preliminaries

2.1 SRC

As is mentioned above, in FRT, [3] has cast the face recognition problem as finding the sparsest representation of the test image over the whole training set. The well-aligned n_i training images of individual i taken under varying illumination are stacked as columns of a matrix $A_i = [a_{i,1}, a_{i,2}, ..., a_{i,n}] \in \mathbb{R}^{m \times n_i}$, each normalized to be the l^2 norm. Then we could define a dictionary as a concentration of all the N object classes $A = [A_1, A_2, ..., A_N] \in \mathbb{R}^{m \times n}$. Here, n is the number of total training images and $n = \sum_{i=1}^{N} n_i$. Since images of the same face under varying illumination lie near a special low dimensional subspace [4], given a test image $y \in \mathbb{R}^m$, it can be represented by a sparse linear combination of the training set $y = Ax$, where x is a coefficient vector whose items are all zero except for those associated with the i^{th} object. However in consideration of the gross error brought by partial corruption or occlusion, the formulation is $y = Ax + e$. Since the desired solution (x, e) is the not only sparse but also the sparsest solution of the system, it can be recovered by the following equation

$$(x, e) = \arg\min \|x\|_0 + \|e\|_0 \quad s.t. \quad y = Ax + e \tag{1}$$

Here, the l^0 norm counts the nonzero number in a vector. However, to solve the l^0 norm problem above is NP-hard. Based on the theoretical result of l^0 and l^1 minimization equivalence, if the solution (x, e) sought is sparse enough, the authors would rather seek the convex relaxation

$$\min \|x\|_1 + \|e\|_1 \quad s.t. \quad y = Ax + e \tag{2}$$

The above problem can be solved by linear programming methods in polynomial time efficiently. In succession, to analyze the coefficient to what extent it concentrates on one subject, we can judge which class it belongs to and whether it belongs to any subjects in the training database.

2.2 Quotient Image

Quotient Image (QI) [5] is firstly introduced by Riklin-Raviv and Shashua, which uses the color ratio to define an illumination invariant signature image and enables a rendering of the image space with varying illumination. The main algorithm of Quotient Image is formulated on the Lambertian model

$$I(x,y) = \rho(x,y)n(x,y)^T s \tag{3}$$

Where ρ is the albedo (surface texture) of face, $n(x,y)^T$ is the surface normal of the object, and s is the point light source direction which can be arbitrary. Then the quotient image Q_y of an object y against the object a is defined by

$$Q_y(u,v) = \frac{\rho_y(u,v)}{\rho_a(u,v)} = \frac{\rho_y(u,v)n(u,v)^T s_y}{\rho_a(u,v)n(u,v)^T s_y} \tag{4}$$

Let s_1, s_2, s_3 be the three linearly independent vectors of a basis, thus any point light source direction can be reconstructed by $s_y = \sum_j x_j s_j$. Hence

$$Q_y(u,v) = \frac{\rho_y(u,v)n(u,v)^T s_y}{\rho_a(u,v)n(u,v)^T \sum_j x_j s_j} = \frac{I_y}{\sum_{j=1}^{3} I_j x_j} \tag{5}$$

Here, I_y is the image of illumination source direction s_y. If the illumination set $\{I_j\}_{j=1}^{3}$ in (5) is devised in advance, given a reference image of object y, the quotient image Q_y could be computed. Then by product of Q_y and the different choice of I_j, the image space of a reference object y could be synthesized by the set of three linearly independent illumination images of other objects. Through the above method, given a single input image of an object, and a database of images with varying illumination of other objects of the same general class, we can re-render the input image to simulate new illumination conditions using QI.

3 SRC for One Training Image per Person

3.1 Motivation for Synthesizing Virtual Samples

To extend SRC method to the one sample image per person problem, the first difficulty to handle is that we must provide over-complete training atoms. One of the ways is synthesizing virtual samples from a single image. Furthermore, to give a better performance of SRC, the dictionary must be discriminant. We test the algorithm of [3] under the following conditions. The Extended Yale B, one of the renowned face databases, consists of 2,414 frontal faces of 38 individuals [4]. We choose the 30^{th} to the 59^{th} images in Extended Yale B of each individual for training, and the 1^{st} to the 29^{th} to test, then down-sample all the images to 12×10 at a rate of 1/16, which is the same feature dimension as in [3]. The recognition rate is about 82.3%, and it is much lower than the results of randomly selecting half of the images for training, which was reported in [3] at a rate between 92.1% and 95.6%. This is because of the uniform distribution of

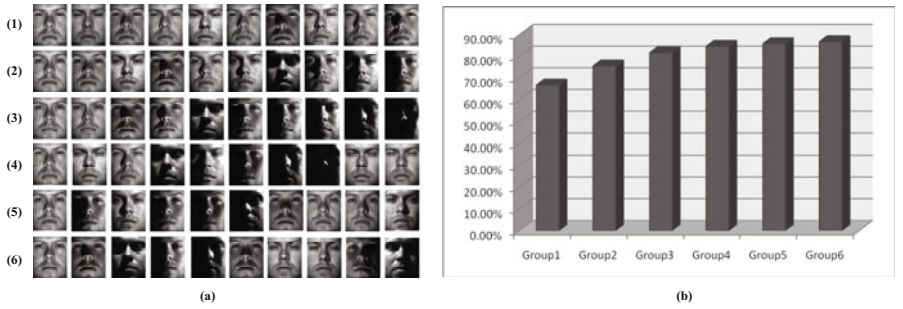

Fig. 1. Left: Six groups of training set in our experiment in rows. In the top Row (1), the first 10 consecutive images are selected. In Row (2) we choose every other image, which is the 1st, 3rd, 5th images and so on. Then in Row (6), we select one in every six consecutive images. Right: The corresponding recognition results to the Left.

the training samples. Since all the images in this training set are collected under 64 different illuminations, the lighting condition varies to the consecutive one gradually.

To discuss how to design the training set for robustness, we conduct experiments as follows. For each individual, we select 10 images for training, with the others left for testing. We divide our training set into six groups. For the first group, the top 10 consecutive images are selected as the training dictionary. For the second set, we select every other images. And for the third, we select one in every three consecutive pictures and so on. As shown in Figure 1(a), there are six groups of images in rows. And the recognition results are in Figure 1(b) respectively. It can be deduced that, the more diversely illumination varies from each other, the better recognition results it gives. In our method, we make full use of illumination environment information of an illumination record instead of traditional compensation ways. We can generate complete and diversity training atoms from a single input image of an individual by devising a small representative illumination record. Therefore, synthesizing virtual samples could not only satisfy the sparse constraint but also form a discriminant dictionary for the one sample per person problem.

3.2 Our Approach of Recognition with One Training Image per Person

Suppose that there is only one image y_{s_i} (the reference image) available for each individual i, firstly we should generate the sufficient and robust training set for each object. Based on the technology of QI, given a database of varying illumination, we could generate new images of another subject of the same condition. Let B_j be a matrix whose columns are the pictures of object j with albedo function ρ_j. The illumination set $\{B_j\}_{j=1}^{K}$, which contains only K (a small number) other objects, can be elaborately designed manually to ensure the discrimination

capability. From (5), if we know the correct coefficient x_j, we can get the quotient image Q_{y_i}. In [5], it proves that the energy function

$$f(\hat{x}) = \frac{1}{2}\sum_{j=1}^{K} |B_j\hat{x} - \alpha_j y_{s_i}|^2 \tag{6}$$

has a global minimum $\hat{x} = x$, if the albedo ρ_y of an object y is rationally spanned by the illumination set. To recover x in (6), it is only a least-squares problem. The global minima x_0 of the energy function $f(\hat{x})$ is:

$$x_0 = \sum_{j=1}^{K} \alpha_j v_j \tag{7}$$

where

$$v_j = (\sum_{r=1}^{K} B_r B_r^T)^{-1} B_j y_{s_i} \tag{8}$$

and the coefficients α_j are determined up to a uniform scale as the solution of the following equation:

$$\alpha_j y_{s_i}^T y_{s_i} - (\sum_{r=1}^{K} \alpha_r v_r)^T B_j y_{s_i} = 0 \quad s.t. \quad \sum_j \alpha_j = K \tag{9}$$

for $j = 1, \ldots, K$. Then the quotient image Q_{y_i} (defined in (4)) is computed by

$$Q_{y_i} = \frac{y_{s_i}}{\overline{B}x} \tag{10}$$

where

$$\overline{B} = \frac{1}{K}\sum_{i=1}^{K} B_i \tag{11}$$

is the average of illumination set.

Then we can approximate which class the given test sample y_t belongs to through SRC. For $z \in \mathbb{R}^n$, $\delta_i(z) \in \mathbb{R}^n$ is a new vector of z whose entries are zero except for those associated with the i^{th} class. One can approximate the given test sample by $\hat{y_i} = A\delta_i(\hat{z})$. We then classify y_t based on these approximations by assigning it to the object class that minimizes the residual between y_t and $\hat{y_i}$:

$$r_i(y_t) = \|y_t - A\delta_i(\hat{z})\|_2 \tag{12}$$

The whole procedures of our method are depicted in the flowchart of Figure 2. And Algorithm 1 summarizes the complete recognition details in one sample per person problem.

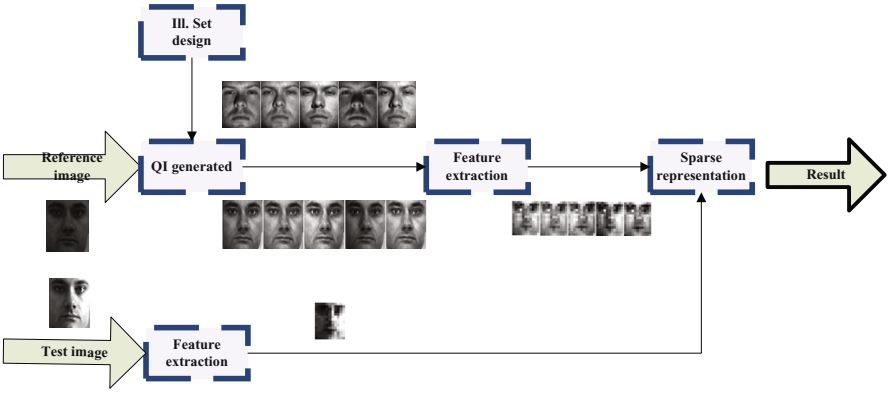

Fig. 2. The whole recognition procedures

Algorithm 1. Single Face Recognition via Sparse Representation

1: **Input**: $\{B_j\}_{j=1}^{K}$, the illumination set, where each matrix contains n_i images (as its columns). y_{s_i}, the reference image of each individual i (a vector of size m). y_t, the test image. Align all the images in the $\mathbb{R}^m$ space.

2: **Step**: Solve the equation (9). Compute x in (7) and the quotient image Q_{y_i} is generated by (10) and (11). Synthesize the training space: For $l = 1, \ldots, n_i$ and $i = 1, \ldots, N$, $A_{i,l} = Q_{y_i} \otimes \overline{B_l}$, where $A_{i,l}$ stands for the l^{th} image of the i^{th} person, $\overline{B_l}$ is the l^{th} column of matrix $\overline{B}$, and $\otimes$ is the Cartesian product (pixel by pixel multiplication).

3: **Step**: Stack all the generated matrices in $A = [A_1, A_2, \ldots, A_N] \in \mathbb{R}^{m \times n}$, and normalize the columns of A to have the unit l^2 norm. Solve the l^1 minimization problem

$$\hat{z} = \arg\min_z \|z\|_1 \quad s.t. \quad \|Az - y_t\|_2 \leqslant \varepsilon \tag{13}$$

where ε is an optional error tolerance. Compute the residuals (12) for $i = 1, \ldots, N$.

4: **Output**: identity$(y_t) = \arg\min_i r_i(y_t)$.

4 Experiments

In this section, we conduct a wide range of experiments on publicly available databases for face recognition. Firstly, we verify our algorithm on CMU-PIE database [6] and discuss its robustness in real-world applications. We then extend our method on color images. Finally, we put forward a cross-database synthesizing comparison.

4.1 Gray Scale Images Verification

Firstly we test our algorithm in the CMU-PIE database. It consists of 2,924 frontal-face images of 68 individuals. The cropped and normalized 112×92 face images were captured under various laboratory-controlled lighting conditions. There are two folders of different illuminations, of which the folder with the

ambient lights off we call it PIE-illumination package while the folder with the ambient lights on is named the PIE-light package. Each person is under 43 different illumination conditions, and the PIE-illumination contains 21 lighting environments.

We randomly choose images of 10 individuals as the illumination set in PIE-light. Since there are 22 distinct illumination, we synthesize images of the same lighting condition given a reference image of the left persons. As is shown in Figure 3 there are 22 different images generated. Here we select the first image of each face as the unique reference image to get quotient image and re-render others while the images left are all used for testing. In the period of recognition, the feature space dimension was set to 154. Since the precise choice of feature space is no longer critical [3], we just down-sample the images at a ratio of 1/8. Our method achieves recognition rate of 95.07% which gives a good result in the single training image problem.

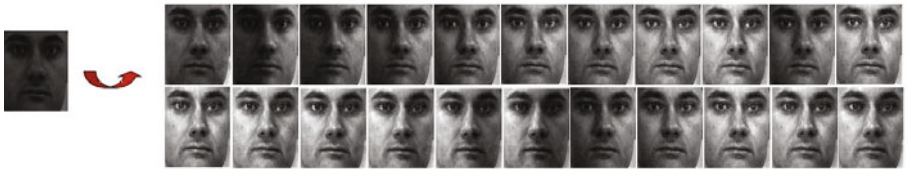

Fig. 3. Virtual samples

One may argue that the reference image we chose was well-chosen, since the first image of each object was under a favorable illumination condition. However, in real-world tasks, especially in law enforcement scenarios for illustration, we often have the record of each suspect with a piece of legible photograph. Therefore this choice of deliberately designed input image looks reasonable. Then we verify the algorithm of different input images as depicted in Figure 4. Here the leave-one-out scheme [7] is used. To name a few, each image acts as the template by turns and the others present as the test part. As is shown in the picture, the recognition result is influenced by the template we choose, and the recognition rate fluctuates between 82.35% and 95.07%. Here templates which give a most or least satisfying performance is presented as below. In Figure 5, we can see that images in column (b) and column (c), which perform worst at a rate around 82.35%, are images with shadows. This is due to the limitation of the formulation of Lambertian model, without taking into account shadow. It can be expected that other algorithms such as 3D morphable model will improve the performance.

4.2 Recognition with Color Images

The next experiment is about color images. Given a color image represented by RGB channels, we could directly convert it to a gray scale image. The gray scale

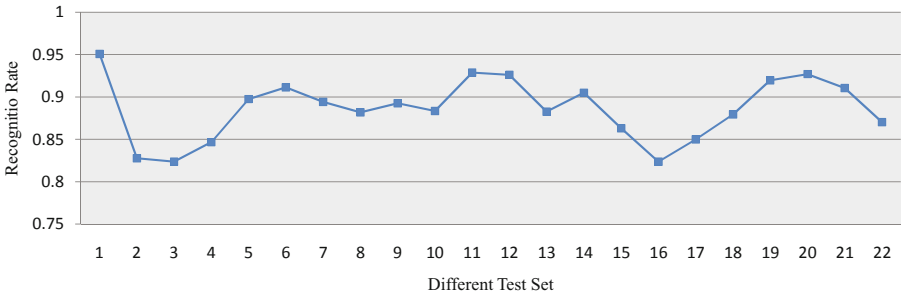

Fig. 4. Recognition Rate on CMU-PIE

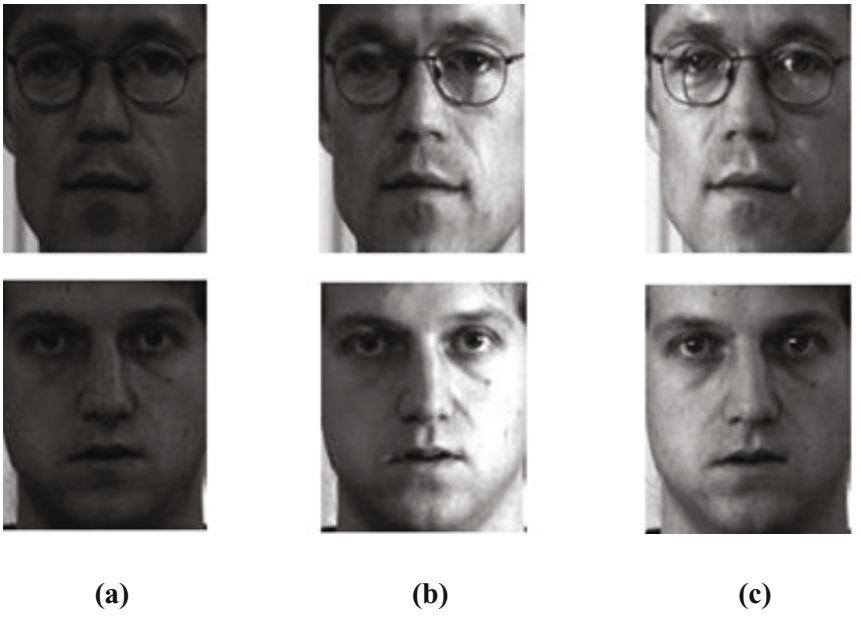

Fig. 5. Templates. The Column (a) has the most willing illumination, which achieved a recognition result at 95.07%. And the (b) and (c) columns perform worst at a rate of 82.34%.

images conserve all the lightness information because lightness of the gray is directly proportional to the number representing the brightness levels of the primary. We can assume that the varying illumination only works on the gray-value distribution without affecting the hue and saturation part. At the same time, the value channel of HSV representation, which describes the luminance (the black-and-white) information, can be transformed from the RGB color space. As is shown in Figure 6, Row (a) contains four color images with different illumination directions. And Row (b) are images from the RGB color space transforms into HSV space. While in the Row (c), we present the value in the V channel. It reveals that the Value channel in HSV space preserves the luminance information.

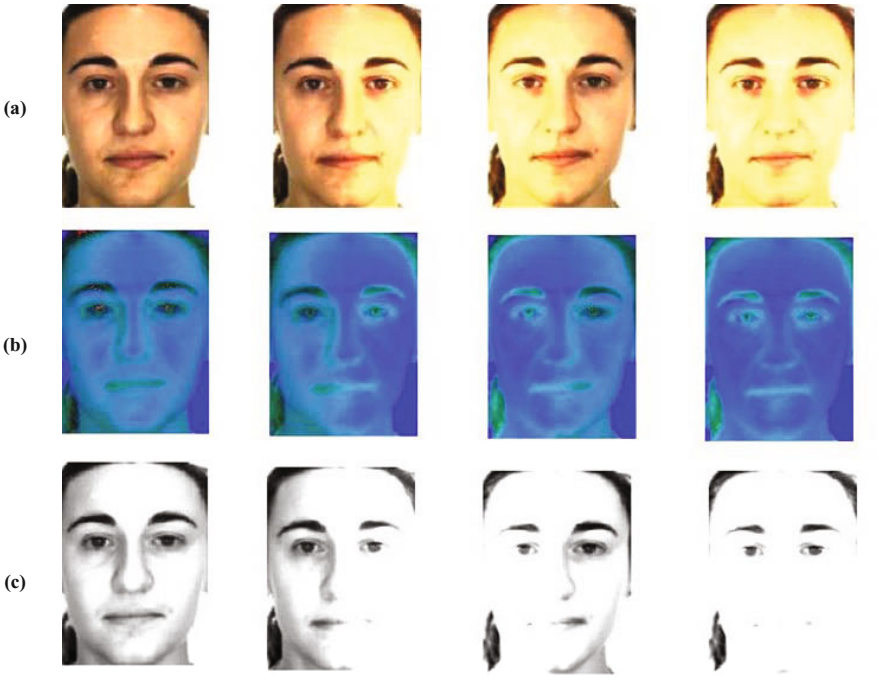

Fig. 6. The V-channel in HSV color space. There are four illumination conditions in columns. Row (a) contains four color images in RGB space. And Row (b) are images from the RGB color space transforms into HSV space. While in the Row (c), we present the value in the V channel.

To illustrate how it works in recognition, we use the AR databases. It is made up of over 4,000 frontal images of 126 individuals. For each person, 26 pictures are taken in two separate sessions named Session 1 and Session 2, of which 8 images are faces with only illumination change [8]. In our trail, we use the 4 illumination conditions of arbitrary 10 individuals in Session 1 as the illumination set. Then we randomly select one of the 8 images of the other individuals for training, and process them as described in Figure 7. Subsequently, we recognize the left faces in the 8 images of each individual after down-sampling all the matrix to 154 features. It achieves a recognition rate between 78.75% and 81.48% though we have only 4 training lighting conditions of each person.

4.3 A Cross-Database Comparison

So far we have experimented with objects and their images from the same databases. Even though the object we choose for training and testing is outside the illumination set, we still have the advantage that they are taken by the same camera, in the same laboratory-controlled environment. Here we test our algorithm under cross-dictionary conditions especially. To name a few, in our

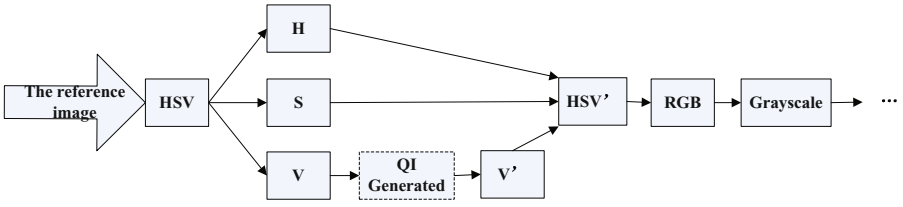

Fig. 7. Color images synthesized procedures

Table 1. Cross database performance

Ill. set	No. of Ills.	Reference image for training	Rec. rate
Extended Yale B	30	20^{th} face of PIE-illumination	76.68%
Extended Yale B	20	20^{th} face of PIE-illumination	72.41%
Extended Yale B	30	1^{st} face of PIE-light	98.67%
Extended Yale B	20	1^{st} face of PIE-light	98.18%

experiment, a cross-database test means that the classical illumination record in the set doesn't come from the same database as images to be recognized. However, in real-world scenarios, we often have possessions of illumination set in advance, but the images for training and identification are usually from a totally different collection environment.

As is shown in Table 1, in attempt to generate an incoherent dictionary [9] in the language of signal representation, we select every other image of random 10 persons in Extended Yale B as the illumination set. Then all the faces in PIE-illumination are taken for testing. The 20^{th} face of each man is the reference image to generate training atoms and the others are test cases. In the beginning, we pre-align all the images in both databases to 112×92 pixels and the down-sampling rate is also 1/8. That is, the feature space we use is 154. As a result, the recognition rate is 76.68%. Whereafter, we decrease the illumination conditions to 20 instead, and the rate is around 72.41% respectively. Finally, images in PIE-light are also used for testing instead of PIE-illumination in the same procedures, and the corresponding performance is recorded in the table.

We can get several views from the table as follows. First of all, we can use our method for one sample per person problem in real-world scenarios when the images in the illumination set are not collected in the same environment as the images to be classified. In addition, illumination diversity of the training set improves the performance. Last but not the least, it implies some tricks of choosing the reference image, since it's obvious that images in PIE-light perform much better than those in PIE-illumination. There may be some reasons related to the ambient light so that illumination distribution is more uniform. However, we can get the environment condition of the reference image in control and make some illumination compensation of it in practice.

5 Conclusions

In this paper, we present a new method for one sample per person recognition based on sparse representation, which makes use of the illumination diversity information instead of compensation pre-processing. This method could not only satisfy the over-complete constraint in the one training image per person problem, but also make the training set discriminant. Our experiments on public databases achieved good performances and a cross-database comparison was put forward. In the future, we will try to improve the virtual sample generation method with other techniques by taking pose variation into consideration. At the same time, details of dictionary design will be explored further as well.

Acknowledgement. This work was supported in part by NSFC (No.60702044, No.60632040, No.60625103), MIIT of China (No.2010ZX03004-003) and 973 Program of China (No.2010CB731401, No.2010CB731406).

References

1. Tan, X.: Face recognition from a single image per person: A survey. Pattern Recognition 39, 1725–1745 (2006)
2. Wright, J.: Sparse representations for computer vision and pattern recognition. Proceedings of IEEE (2009)
3. Wright, J.: Robust face recognition via sparse representation. IEEE Trans. (2008)
4. Geoghiades, A.: From few to many: Illumination Cone models for face recognition under variable lighting and pose. IEEE Trans. on Pattern Analysis and Machine Intelligence 23, 643–660 (2001)
5. Shashua, A.: The quotient image: Class-based re-rendering and recognition with varying illuminations. Transactions on Pattern Analysis and Machine Intelligence 23, 129–139 (2001)
6. Sim, T.: The CMU Pose Illumination and Expression (PIE) Database. In: The 5th International Conference on Automatic Face and Gesture Recognition (2002)
7. Wang, H.: Face Recognition under Varying Lighting Condition Using Self Quotient Image. In: Proceedings of International Conference on Automatic Face and Gesture Recognition, vol. 54, pp. 819–824 (2004)
8. Martinez, A.: The AR face database. CVC Tech. Report (1998)
9. Donoho, D.: For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. In: IEEE International Symposium on Comm. Pure and Applied Math., Information Theory, ISIT 2007, vol. 59, pp. 797–829 (2006)