Efficient Misalignment-Robust Representation for Real-Time Face Recognition

Meng Yang, Lei Zhang, and David Zhang

Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong
{csmyang,cslzhang}@comp.polyu.edu.hk

Abstract. Sparse representation techniques for robust face recognition have been widely studied in the past several years. Recently face recognition with simultaneous misalignment, occlusion and other variations has achieved interesting results via robust alignment by sparse representation (RASR). In RASR, the best alignment of a testing sample is sought subject by subject in the database. However, such an exhaustive search strategy can make the time complexity of RASR prohibitive in large-scale face databases. In this paper, we propose a novel scheme, namely misalignment robust representation (MRR), by representing the misaligned testing sample in the transformed face space spanned by all subjects. The MRR seeks the best alignment via a two-step optimization with a coarse-to-fine search strategy, which needs only two deformation-recovery operations. Extensive experiments on representative face databases show that MRR has almost the same accuracy as RASR in various face recognition and verification tasks but it runs tens to hundreds of times faster than RASR. The running time of MRR is less than 1 second in the large-scale Multi-PIE face database, demonstrating its great potential for real-time face recognition.

1 Introduction

After many years investigation of face recognition (FR) techniques [1], there are mainly two branches of FR research and development. One focuses on face verification with face images captured in uncontrolled or less controlled environment. The representative databases include LFW [2] and PubFig [3], with the representative methods such as [3], [4], [5], [6], etc. The other emphasizes on proposing new frameworks for (semi) controlled scenarios and with cooperative subjects, which have extensive applications including access control, computer systems, automobiles or automatic teller machines, etc [7]. The goal of the latter branch is for high robustness and high accuracy, and many state-of-the-art works [7], [8], [9], [10], [11], [12], [13], [14] have been proposed along this line to address various challenges, including face corruption, occlusion, misalignment and the variations of illumination, expression, etc.

The recently developed sparse representation based FR methods belong to the above mentioned second branch, and they target one important category of applications where many well-controlled training images are available. The

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part I, LNCS 7572, pp. 850–863, 2012.

pioneer work in [8], i.e., sparse representation-based classification (SRC), casts the recognition problem as finding a sparse linear representation of the test image over the training images. Furthermore, by assuming that the outlier pixels in the face image are sparse and using an identity matrix to code the outliers, SRC shows good robustness to face occlusion and corruption. The success of SRC inspires many following works, such as structured sparse representation [9], robust sparse coding [10], SRC for continuous occlusion [11], etc.

Although the well-aligned training images could be prepared, the testing images have to be automatically cropped by using some detector, e.g., the Viola and Jones' face detector [15]. Inevitably there will be certain registration error of several pixels, which will deteriorate much the performance of many FR methods [16], including SRC [8]. To solve this problem, by adding a deformation term to face representation [17], FR methods such as robust subspace learning to misalignment [14] and simultaneous image alignment and sparse representation [7], [12] have been proposed, where misalignment, occlusion and other variations (e.g., illumination) could be simultaneously handled. Though some face image registration methods, such as Active Appearance Models [17], Active Shape Models [18] and Unsupervised Joint Alignment [5], have advantages in dealing with variations in expression and pose, their goal is for face image alignment but not for recognition, and their complexity can be too high for the application of real-time FR. Therefore, in this paper we aim to propose a new robust FR method along the line of [7], [8], [12], and [14].

The simultaneous face image alignment and representation [7], [12], [14] proposes a promising framework for robust FR with occlusion, misalignment, illumination and expression changes, etc. However, there are still significant concerns on them. The approach in [14] adopts an indirect model (by using neighboring pixels' relation) to recover image transformation, which would complicate the original problem and weaken the capability of handling misalignment. Different from [14], direct recovery of image transformation and sparse representation is adopted in [7] and [12]. However, deforming training samples instead of the testing image [12] makes the size of dictionary for sparse representation very large, which dramatically increases the difficulty and time complexity of image representation. The recent work in [7] uses an integral model of robust alignment by sparse representation (RASR), which is free of the shortcomings in [12] and [14]. However, the model in [7] is hard to optimize due to the coupling of image spatial transformation and unknown identity, and the authors proposed a suboptimal algorithm via subject-by-subject exhaustive search, whose time complexity increases linearly as the number of subjects. Such a time-consuming optimization makes RASR prohibitive in large-scale and real-time FR systems.

This paper will present an efficient misalignment-robust representation (MRR) for real-time FR. We will show that the exhaustive search yet suboptimal optimization used in [7] is not necessary. By analyzing why simultaneous image alignment and representation is difficult, we design a misalignment-robust model via correspondence-based representation, which could effectively avoid falling into a bad local minimum. The proposed MRR scheme is free of the time-consuming

sparsity constraint on representation coefficients, and can be efficiently solved by a two-step optimization algorithm with a coarse-to-fine search strategy. Compared to RASR [7], the time complexity of MRR is nearly independent of subject number (denoted by c) in the database, and the speedup of MRR over RASR is more than $c/2$. Our experiments on benchmark face databases clearly show that MRR has very competitive FR results with RASR, and more importantly, it can be a truly real-time FR method; e.g., it is over 150 times faster than RASR in the large-scale Multi-PIE database.

The rest of this paper is organized as follows. Section 2 briefly reviews the RASR method in [7]. Section 3 presents the model and algorithm of the proposed MRR. Section 4 analyzes the time complexity. Section 5 conducts experiments and Section 6 concludes the paper.

2 Robust Alignment by Sparse Representation (RASR)

Different from the previous face alignment methods [5], [17], [18], which may have advantages in dealing with large variations in expression and pose, the RASR [7] focuses on deformations with fewer degrees of freedom, i.e., similarity transformations, and uses the the training images themselves as the appearance model.

Suppose that $\boldsymbol{y}$ is the observed query face image which is warped due to misalignment and denote by $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m] \in \mathbb{R}^{n \times m}$ the matrix with all vectorized training samples as its column vectors. RASR assumes that the deformation-recovered image $\boldsymbol{y}_0 = \boldsymbol{y} \odot \boldsymbol{\tau}$ has a sparse representation over $\boldsymbol{A}$: $\boldsymbol{y}_0 = \boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{e}$, where $\boldsymbol{\tau}$ represents some kind of spatial transformation (e.g., similarity, affine, etc.) but with unknown parameters, $\boldsymbol{\alpha}$ is the sparse coding vector and $\boldsymbol{e}$ is the coding residual vector. The model of RASR [7] is

$$\langle \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\tau}} \rangle = \arg\min_{\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{e}} \|\boldsymbol{\alpha}\|_1 + \|\boldsymbol{e}\|_1 \ \text{ s.t. } \ \boldsymbol{y} \odot \boldsymbol{\tau} = \boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{e} \tag{1}$$

where the sparsity of $\boldsymbol{\alpha}$ is claimed to provide a strong cue for finding the correct deformation $\boldsymbol{\tau}$. Due to the difficulty of solving Eq. (1), the authors turned to seek for the best alignment of $\boldsymbol{y}$ via a subject-to-subject optimization:

$$\langle \hat{\boldsymbol{\tau}}_i, \hat{\boldsymbol{e}}_i \rangle = \arg\min_{\boldsymbol{\alpha}_i, \boldsymbol{\tau}_i, \boldsymbol{e}_i} \|\boldsymbol{e}_i\|_1 \ \text{ s.t. } \ \boldsymbol{y} \odot \boldsymbol{\tau}_i = \boldsymbol{A}_i\boldsymbol{\alpha}_i + \boldsymbol{e}_i \tag{2}$$

where $\boldsymbol{A}_i$ is the matrix associated with subject i, and $\boldsymbol{\tau}_i$ is the transformation aligning $\boldsymbol{y}$ to subject i.

After an exhaustive alignment to every subject in the face database, the top S candidates $k_1, \cdots, k_S$ with the smallest residuals $||\hat{\boldsymbol{e}}_i||_1$ are selected to construct a new dictionary $\boldsymbol{D} = \left[\boldsymbol{A}_{k_1} \odot \hat{\boldsymbol{\tau}}_{k_1}^{-1}, \boldsymbol{A}_{k_2} \odot \hat{\boldsymbol{\tau}}_{k_2}^{-1}, \cdots, \boldsymbol{A}_{k_S} \odot \hat{\boldsymbol{\tau}}_{k_S}^{-1}\right]$, where $\boldsymbol{A}_{k_i} \odot \hat{\boldsymbol{\tau}}_{k_i}^{-1}$ means aligning each training sample of subject k_i to $\boldsymbol{y}$ and forming $\boldsymbol{A}_{k_i} \odot \hat{\boldsymbol{\tau}}_{k_i}^{-1}$ with the aligned training samples. Then the sparse vector $\boldsymbol{\alpha}$ is computed via

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}, \boldsymbol{e}} \|\boldsymbol{\alpha}\|_1 + \|\boldsymbol{e}\|_1 \ \text{ s.t. } \ \boldsymbol{y} = \boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{e} \tag{3}$$

Finally, like SRC [8] RASR classifies $\boldsymbol{y}$ by evaluating which class could yield the least reconstruction error [7].

While RASR has shown impressive results [7], it has a few drawbacks as described below.

1. For a large-scale face database with c subjects, Eq. (2) needs to be solved c times, making RASR have a high time complexity.
2. Aligning well-cropped training samples to poorly-cropped testing samples may lose some facial features and introduce disturbances (i.e., background)..
3. The accuracy of solving Eq.(2) is based on the good representation ability of the training samples $\boldsymbol{A}_i$, which could not be ensured in the lack of enough training samples.

3 Misalignment-Robust Representation (MRR)

3.1 Simultaneous Alignment and Representation

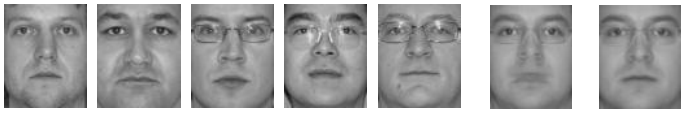
The problem of simultaneous alignment and representation could be represented by:

$$\boldsymbol{y} \odot \boldsymbol{\tau} = \boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{e} \tag{4}$$

where $\boldsymbol{\alpha}$ in $\boldsymbol{y}_0 = \boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{e}$ is unknown for the image alignment sub-problem (i.e., $\boldsymbol{y}_0 = \boldsymbol{y} \odot \boldsymbol{\tau}$) while $\boldsymbol{\tau}$ in $\boldsymbol{y}_0 = \boldsymbol{y} \odot \boldsymbol{\tau}$ is unknown for the image representation sub-problem (i.e., $\boldsymbol{y}_0 = \boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{e}$). Because the joint optimization of ($\boldsymbol{\tau}$, $\boldsymbol{\alpha}$) is neither convex nor smooth (for example when $\boldsymbol{\alpha}$ is regularized by l_1-norm [7]), the alternative optimization of ($\boldsymbol{\tau}$, $\boldsymbol{\alpha}$) may have many local minima, making $\boldsymbol{\tau}$ be estimated inaccurately and $\boldsymbol{\alpha}$ indicate the face identity incorrectly.

For the application of FR, fortunately we have two important priors for simultaneous image alignment and representation. One prior is that 2D similarity transformation $\boldsymbol{\tau}$ could well handle the misalignment problem in FR, while 2D projective transformation could handle moderate pose variation well. This prior has been adopted in [12], [7]. The other prior is the fact that face images from different subjects share big similarities, which is much ignored in previous works [12], [7]. It is not difficult to see that all people's key facial features (i.e., eyes, nose, mouth, etc.) are somewhat similar in appearance and they also have similar locations in face. This is why human can accurately manually align a face image even with pose variation according to another person's reference face image.

However, the similarities of face images could not be well exploited except that a suitable representation model is used, for example, the correspondence-based representation (please refer to Eq. (5) and the related explanations for more information). Fig. 1 gives an example. Fig. 1(a) shows five face images, whose eyes' centers are in the same position. The direct average of these five images is shown in Fig. 1(b). We can see obvious artifacts in the nose and mouth areas because the facial features (except for eyes) are not well aligned in the average. Fig. 1(c) shows the mean image of the five images with correspondence-based representation. Clearly, a much better mean face is produced.



(a) (b) (c)

Fig. 1. Face image representation with and without correspondence. (a) shows five face images with aligned eye centers; (b) shows the mean face of these five images, where clear artifacts can be seen in the nose and mouth areas; (c) presents the mean face of the five images with correspondence-based representation, which looks much better.

The correspondence-based representation could make the face space spanned by the training face images as close to the true face space as possible, and hence help to prevent the simultaneous alignment and representation from falling into a bad local minimum.

3.2 Model of MRR

Suppose there is a well cropped and centered face template $\boldsymbol{y}_t$ for all face images. If both the query image $\boldsymbol{y}$ and the training image set $\boldsymbol{A}$ can be aligned to $\boldsymbol{y}_t$, then the facial structures of $\boldsymbol{y}$ can be corresponded well to those of $\boldsymbol{A}$. With this virtual template $\boldsymbol{y}_t$ (it does not need to be obtained explicitly in our approach) as a bridge to make $\boldsymbol{y}$ correspond to $\boldsymbol{A}$, the proposed correspondence-based simultaneous alignment and representation model is

$$\boldsymbol{y} \odot \boldsymbol{\tau} = (\boldsymbol{A} \odot \boldsymbol{T})\,\boldsymbol{\alpha} + \boldsymbol{e} \tag{5}$$

where $\boldsymbol{T} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_m]$, and the operations, $\boldsymbol{y} \odot \boldsymbol{\tau}$ and $\boldsymbol{A} \odot \boldsymbol{T}$, align the query image $\boldsymbol{y}$ and each training image $\boldsymbol{a}_i$ to $\boldsymbol{y}_t$ via the transformation $\boldsymbol{\tau}$ and $\boldsymbol{t}_i, i = 1, \cdots, m$, respectively. Therefore, the proposed misalignment-robust representation (MRR) model for FR is formulated as

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{e}, \boldsymbol{T}} \|\boldsymbol{e}\|_1 \;\text{ s.t. }\; \boldsymbol{y} \odot \boldsymbol{\tau} = (\boldsymbol{A} \odot \boldsymbol{T})\,\boldsymbol{\alpha} + \boldsymbol{e} \tag{6}$$

In the above model of MRR, the operation of $\boldsymbol{A} \odot \boldsymbol{T}$ aligns the training samples, which makes the linear combination of all the training samples, $(\boldsymbol{A} \odot \boldsymbol{T})\boldsymbol{\alpha}$, more accurate to represent a query face image and benefit the accurate recovery of transformation $\boldsymbol{\tau}$. The l_1-norm minimization of representation error aims to increase the robustness of MRR to image occlusions, such as disguise, block occlusions or pixel corruption.

Usually the training data set in a face database includes more than tens of subjects. Given well-aligned training samples of each subject, the representation coefficient $\boldsymbol{\alpha}$ of a well-aligned query sample (say from class i) can be sparse since using only the training samples from class i can represent the query sample well. However, the sparsity constraint on $\boldsymbol{\alpha}$ will make the optimization of representation very time-consuming [13], especially for the alternative optimization of $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$, which may need many iterations.

Actually, it is not necessary to solve a sparse representation problem in the alternative optimization of Eq. (6) because the sparse coefficient $\boldsymbol{\alpha}$ here is only used to do representation but not classification. We rewrite the dictionary $\boldsymbol{A} \odot \boldsymbol{T}$ via singular value decomposition (SVD): $\boldsymbol{A} \odot \boldsymbol{T} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrixes and $\boldsymbol{\Sigma}$ is a diagonal matrix with descending-order diagonal values. Therefore the original MRR model is transformed into

$$\min_{\boldsymbol{\beta},\, \boldsymbol{\tau},\, \boldsymbol{e}} \|\boldsymbol{e}\|_1 \;\text{ s.t. }\; \boldsymbol{y} \odot \boldsymbol{\tau} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{e} \tag{7}$$

where $\boldsymbol{\beta} = \boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{\alpha}$ and only the first several elements of $\boldsymbol{\beta}$ will have big absolute values. Therefore, Eq.(7) could be approximated as

$$\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\beta}_\eta, \boldsymbol{\tau}, \boldsymbol{e}} \|\boldsymbol{e}\|_1 \;\text{ s.t. }\; \boldsymbol{y} \odot \boldsymbol{\tau} = \boldsymbol{U}_\eta\boldsymbol{\beta}_\eta + \boldsymbol{e} \tag{8}$$

where $\boldsymbol{U}_\eta$ is formed by the first η column vectors of $\boldsymbol{U}$. Due to the fact that $\boldsymbol{U}_\eta$ is a tall matrix and almost all the element of $\boldsymbol{\beta}_\eta$ have significant values, the representation on $\boldsymbol{U}_\eta$ will be stable enough and the regularization (e.g., $\|\cdot\|_1$) on $\boldsymbol{\beta}_\eta$ is not necessary.

After optimizing Eq. (8), the coding coefficient $\boldsymbol{\alpha}$ (regularized by l_2-norm as [13]) could be solved by

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{e}\|_{l_p} + \lambda\|\boldsymbol{\alpha}\|_2^2 \;\text{ s.t. }\; \boldsymbol{y} \odot \hat{\boldsymbol{\tau}} = (\boldsymbol{A} \odot \boldsymbol{T})\boldsymbol{\alpha} + \boldsymbol{e} \tag{9}$$

where $p = 1$ for face with occlusion and $p = 2$ for face without occlusion (in that case, Eq. (9) is equal to $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} \odot \hat{\boldsymbol{\tau}} - (\boldsymbol{A} \odot \boldsymbol{T})\,\boldsymbol{\alpha}\|_2^2 + \lambda\,\|\boldsymbol{\alpha}\|_2^2$).

3.3 Coarse-to-Fine Search of MRR

Instead of optimizing via an exhaustive search subject by subject [7], with MRR we directly align the query image $\boldsymbol{y}$ to the training data $\boldsymbol{A}$ of all subjects. Specifically, we propose a two-step coarse-to-fine search strategy. Before the online coarse-to-fine search for $\boldsymbol{y}$, we first estimate offline the transformation $\boldsymbol{T}$.

3.3.1. Estimate $\boldsymbol{T}$

In the MRR model, we assume that there is a universal template $\boldsymbol{y}_t$ to align query image $\boldsymbol{y}$ and training samples $\boldsymbol{A}$. However, such a universal template $\boldsymbol{y}_t$ is hard to get in practice. One more practical way is to estimate adaptively a template $\boldsymbol{y}_t$ for each given database. Furthermore, since our goal is to align $\boldsymbol{y}$ to $\boldsymbol{A}$, we do not need to explicitly have a template $\boldsymbol{y}_t$. In particular, we can align all the samples in $\boldsymbol{A}$ to each other first, and then align $\boldsymbol{y}$ to the already aligned dataset $\boldsymbol{A} \odot \boldsymbol{T}$. The alignment of training images $\boldsymbol{A}$ could be done offline using methods such as AAM [17] and robust alignment by sparse and low rank decomposition (RASL) [19]. In this paper, we adapt RASL to our model since it does not need to manually locate the many feature points except for the initial locations of eyes.

Let $\boldsymbol{B}$ be the aligned dataset of $\boldsymbol{A}$ after some processing. Due to the high similarity existed in face images, the well aligned face vectors in $\boldsymbol{B}$ will be highly correlated. That is, $\boldsymbol{B}$ will have a low rank. Therefore, the transformation $\boldsymbol{T}$ could be estimated by solving the following optimization:

$$\hat{\boldsymbol{T}} = \underset{\boldsymbol{B},\boldsymbol{T},\boldsymbol{E}}{\arg\min}\ \text{rank}\,(\boldsymbol{B}) + \gamma\|\boldsymbol{E}\|_0 \quad \text{s.t.}\ \ \boldsymbol{A} \odot \boldsymbol{T} = \boldsymbol{B} + \boldsymbol{E} \tag{10}$$

where $\gamma > 0$ is a parameter that trades off the rank of the aligned face images $\boldsymbol{B}$ and the sparsity of error $\boldsymbol{E}$. Different from RASL [19] where $\boldsymbol{A}$ contains the training images from the same subject, here $\boldsymbol{A}$ consists of training images from all subjects.

3.3.2. The Coarse Search of MRR

With the transformation computed in Eq. (10), the aligned training samples are $\hat{\boldsymbol{A}} = \boldsymbol{A} \odot \hat{\boldsymbol{T}}$ and the SVD of $\hat{\boldsymbol{A}}$ is $\hat{\boldsymbol{A}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$. We can then do alignment and representation of $\boldsymbol{y}$ over $\hat{\boldsymbol{A}}$ via coarse-to-fine search. The coarse search of MRR is as follows:

$$\hat{\boldsymbol{\tau}}_1 = \underset{\boldsymbol{\beta}_1,\boldsymbol{\tau}_1,\boldsymbol{e}_1}{\arg\min} \|\boldsymbol{e}_1\|_1 \ \ \text{s.t.}\ \boldsymbol{y} \odot \boldsymbol{\tau}_1 = \boldsymbol{U}_1\boldsymbol{\beta}_1 + \boldsymbol{e}_1 \tag{11}$$

where $\boldsymbol{U}_1$ is formed by the first η_1 columns of $\boldsymbol{U}$.

Then the aligned query image $\boldsymbol{y} \odot \hat{\boldsymbol{\tau}}_1$ is coded via Eq. (9), and the reconstruction error of each class to represent $\boldsymbol{y}$ is

$$r_i^1 = \left\| \boldsymbol{y} \odot \hat{\boldsymbol{\tau}}_1 - \hat{\boldsymbol{A}}_i\hat{\boldsymbol{\alpha}}_i \right\|_{l_p} \tag{12}$$

where $\hat{\boldsymbol{\alpha}}_i$ and $\hat{\boldsymbol{A}}_i$ are the coding vector and aligned training sample matrix associated with class i, respectively.

3.3.3. The fine search of MRR

The top S candidates $k_1, \cdots, k_S$ with the smallest residuals r_i^1 are selected to build a new dictionary, $\boldsymbol{D}_f = \left[\hat{\boldsymbol{A}}_{k_1}, \cdots, \hat{\boldsymbol{A}}_{k_S}\right]$. Denote the SVD of $\boldsymbol{D}_f$ as $\boldsymbol{D}_f = \boldsymbol{U}'\boldsymbol{\Sigma}'\boldsymbol{V}'^T$. The fine optimization of MRR is

$$\hat{\boldsymbol{\tau}}_2 = \underset{\boldsymbol{\beta}_2,\boldsymbol{\tau}_2,\boldsymbol{e}_2}{\arg\min} \|\boldsymbol{e}_2\|_1 \ \ \text{s.t.}\ \boldsymbol{y} \odot \boldsymbol{\tau}_2 = \boldsymbol{U}_2\boldsymbol{\beta}_2 + \boldsymbol{e}_2 \tag{13}$$

where the first η_2 column vectors of $\boldsymbol{U}'$ form $\boldsymbol{U}_2$.

Then the coding vector $\hat{\boldsymbol{\alpha}}_f$ of the aligned query image with $\hat{\boldsymbol{\tau}}_2$ is solved via Eq. (9), and the identity of the query image is classified as

$$\text{identity}\,(\boldsymbol{y}) = \arg\min_i \left\| \boldsymbol{y} \odot \hat{\boldsymbol{\tau}}_2 - \hat{\boldsymbol{A}}_i\hat{\boldsymbol{\alpha}}_f^i \right\|_{l_p} \tag{14}$$

where $\hat{\boldsymbol{\alpha}}_f^i$ is the coding vector associated with class i. It should be noted that the estimated transformation $\hat{\boldsymbol{\tau}}_1$ could be used as the initial value of $\boldsymbol{\tau}_2$ in optimizing Eq. (13) and the final representation could also be performed on $\boldsymbol{D}_f$.

The algorithm of MRR is summarized in Algorithm 1. In the coarse-to-fine search of MRR, the solving of Eq. (11) and Eq. (13) is the same as the optimization of Eq. (2) in RASR [7], which iteratively linearizes the current estimate of $\boldsymbol{\tau}$ and seek for representations like (take Eq. (11) as an example):

$$\boldsymbol{y} \odot \boldsymbol{\tau} + \boldsymbol{J}\Delta\boldsymbol{\tau} = \boldsymbol{U}_1\boldsymbol{\beta}_1 + \boldsymbol{e}_1 \tag{15}$$

where $\boldsymbol{J} = \frac{\partial}{\partial \boldsymbol{\tau}} \boldsymbol{y} \odot \boldsymbol{\tau}$ is the Jacobian of $\boldsymbol{y} \odot \boldsymbol{\tau}$ with respect to the transformation parameters $\boldsymbol{\tau}$.

Algorithm 1. Algorithm of Misalignment-Robust Representation (MRR)

1: **Input**
Training data matrix $\boldsymbol{A}$, query image $\boldsymbol{y}$, and initial transformation $\boldsymbol{\tau}_0$ of $\boldsymbol{y}$.
2: **Offline estimation of the transformation $\boldsymbol{T}$**
By Eq. (10), the transformation $\hat{\boldsymbol{T}}$ of $\boldsymbol{A}$ is estimated, and the aligned training images are $\hat{\boldsymbol{A}} = \boldsymbol{A} \odot \hat{\boldsymbol{T}}$.
3: **Coarse search**
Estimate coarse transformation $\hat{\boldsymbol{\tau}}_1$ by Eq. (11) and calculate the reconstruction error associated to each subject.
4: **Fine search**
Choose top S candidates to form a new dictionary, and estimate fine transformation $\hat{\boldsymbol{\tau}}_2$ by Eq. (13).
5: **Output**
Represent the well aligned test sample, $\boldsymbol{y} \odot \hat{\boldsymbol{\tau}}_2$, via Eq. (9), and output the identity of $\boldsymbol{y}$.

4 Complexity Analysis

In this section, we compare the time complexity of the proposed MRR with two state-of-the-art sparse-representation based robust FR methods: Huang's method [12] and RASR [7]. The main time complexity of MRR and RASR [7] contains two parts: simultaneous alignment and representation (SAR) and image representation (IR) for final classification; while the most time-consuming part of Huang's method [12] is the iterative process of SAR.

The complexity of SAR in MRR for one step (coarse or fine step) search, denoted by $O_{\boldsymbol{\tau}}$, is similar to that of SAR in RASR for one subject because they has the same optimization process [7] with similar-size dictionaries ($\boldsymbol{U}_l$ and $\boldsymbol{A}_i$ have similar number of columns for $l = 1, 2$ and $i = 1, 2, \cdots, c$). For the IR with an $n \times m$ dictionary, due to the l_2-norm regularization of $\boldsymbol{\alpha}$ of MRR, Eq. (9) with $p = 2$ has the time complexity of $O(mn)$, while for $p = 1$ Eq. (9) could be efficiently solved by Augmented Lagrange Multiplier (ALM) algorithm [20] with the time complexity of $O(kmn)$, where k is the iteration number (usually less than 50) of ALM. It can be seen that MRR is much faster than the sparse representation in RASR, whose time complexity is $O(n^2(mS/c)^{\varepsilon})$ where $\varepsilon \geq 1.2$ [21] and S is the number of candidates selected from all the subjects. The overall

time complexities of MRR and RASR are listed in Table 1, which clearly shows that MRR has much lower time complexity than RASR.

For Huang's method [12], taking affine spatial transformation as an example, its dictionary size for SAR in each iteration is $n \times 7m$ (each training sample generates 6 column vectors as the dictionary atoms to deal with transformation). The time complexity for one iteration is $O(7^{\varepsilon} n^2 m^{\varepsilon}) \approx O(10 n^2 m^{\varepsilon})$. Therefore the total complexity of Huang's method is $O(10qn^2m^{\varepsilon})$, where q is the iteration number. Since one-step SAR costs less time than sparse coding, i.e., $O_{\boldsymbol{\tau}} < O(n^2 m^{\varepsilon})$, the overall complexity of MRR $(2O_{\boldsymbol{\tau}} + 2O(kmn))$ is much lower than Huang's method [12], which is the slowest one among the three methods.

Table 1. Time complexity of MRR and RASR

Step	SAR	IR	Comparison comment
RASR [7]	$cO_{\boldsymbol{\tau}}$	$O(n^2(mS/c)^{\varepsilon})$	MRR usually has over $\frac{c}{2}$ times speedup over
MRR	$2O_{\boldsymbol{\tau}}$	$2O(nm)$ or $2O(knm)$	RASR (c is the number of subjects).

5 Experimental Result

We perform experiments on benchmark face databases to demonstrate the effectiveness of MRR. We first discuss the parameter selection of MRR in Section 5.1; in Section 5.2, we evaluate the alignment of MRR via simulating 2D deformation. In Section 5.3, we test the robustness of MRR to the number of training samples and block occlusion. In Section 5.4, we verify the effectiveness of MRR on real face recognition and verification, followed by the running time comparison in Section 5.5. All the training face images are manually cropped based on the locations of eyes, while the testing face images in all experiments (except the simulation on 2D deformation) are automatically detected by using Viola and Jone's face detector [15] without manual intervention. The supplementary material which includes more results and the Matlab source code of this paper can be downloaded at `http://www4.comp.polyu.edu.hk/~cslzhang/code.htm`.

5.1 Discussion of Parameter Selection

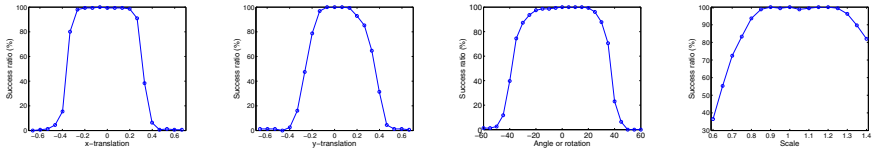
In MRR, apart from γ in estimating $\boldsymbol{T}$ (we use the default value of γ in [19]), there are four parameters (λ ,η_1, η_2 and S) need to be set beforehand. Among them, λ, η_2 and S are relatively easy to set and they can be fixed for all experiments, while η_1 depends much on the face subspace generated by $\boldsymbol{U}_1\boldsymbol{\beta}_1$ in Eq. (11). A small η_1 will reduce the representation power of $\boldsymbol{U}_1\boldsymbol{\beta}_1$ but increase its robustness to big misalignment. Thus, when the misalignment is small, a big $\boldsymbol{U}_1\boldsymbol{\beta}_1$ is preferred, and vice versa. In this following, if no specific instructions, we fix $\eta_2 = 40$, $\lambda = 0.01$, and $S = 25$ for all the experiments. For the simulation in Section 5.2 and the experiment of robustness to occlusion, η_1 is set as a small value (i.e., 4), while for all the other cases, η_1 is fixed as 25. In addition, 2D spatial similarity transformation $\boldsymbol{\tau}$ is used in the experiments.

5.2 Simulation on 2D Deformation

We first verify the capability of MRR to deal with 2D deformation (including translation, rotation and scaling) using the CMU Multi-PIE database [22]. As in [7], all the subjects in Session 1, each of which has 7 frontal images with extreme illuminations $\{0, 1, 7, 13, 14, 16, 18\}$ and neutral expression, are used for training, and the subjects from Session 2 with illumination $\{10\}$ are used for testing. The images were down-sampled to 80×64 with the distance between the two outer eye corners as 46 pixels. Artificial deformation of translations, rotation and scale are introduced to the testing images based on the coordinates of eye corners located manually. We compute the success ratio as N_1/N_2, where N_1 is the number of misaligned testing samples (i.e., with artificial deformation) correctly classified by MRR, and N_2 is the number of well-aligned testing samples (i.e., manually cropped without any deformation) correctly classified by using the classifier employed in MRR.

Fig. 2 shows the success ratio for each single artificial deformation: 2(a) and 2(b) for x and y translations, 2(c) for rotation, and 2(d) for scaling. It can be seen that MRR works well when the translation in x or y direction is less than 20% of the eye distance and when the in-plane rotation is less than 30%. This performance is similar to RASR [7], while it should be noted that the generic alignment to all subjects in MRR is much harder than the specific alignment to one subject in RASR. In addition, from Fig. 2(d) we can see that MRR performs well up to 20% change in scale, better than RASR [7] (up to 15% scale variation). It is also interesting to see that face scaling up (e.g., $1 \sim 1.4$ scaling) is easier to handle than face scaling down (e.g., $0.6 \sim 1$ scaling), which is because small cropped face regions would lose some discriminative facial features.

We then compare MRR with three state-of-the-art methods, SRC [8], Huang's method (H's) [12] and RASR [7], by performing FR experiments on the Extended Yale B (EYB) [23] and Multi-PIE [22] databases. The experimental settings on Multi-PIE remains the same as above except that an artificial translation of 5 pixels in both x and y directions is introduced into the test image. For the settings on EYB, as in [12][7] 20 subjects are selected and for each subject 32 frontal images (selected randomly) are used for training, with the remaining 32 images for testing. An artificial translation of 10 pixels in both x and y directions is introduced into the test image. The image is cropped to 192×168. Because there are only 20 subjects in EYB, S is set as 10 here. Table 2 shows the recognition rates on these two datasets by the four competing methods. It can be seen that SRC is very sensitive to misalignment, and it gives the worst performance. Both MRR and RASR have much higher recognition rates than Huang's method [12]. Compared with RASR, MRR achieves 0.6% improvement on Multi-PIE and achieves almost the same rate on EYB. However, as we will see in Section 5.5, MRR is tens to hundreds of times faster than RASR.



(a) (b) (c) (d)

Fig. 2. Success ratio of MRR versus 2D deformation: (a) translation (percent of eye distance) in x direction only; (b) translation (percent of eye distance) in y direction only; (c) in-plane rotation only and (d) scale variation only

Table 2. Recognition rates with translations (MPIE: Multi-PIE)

	SRC[8]	RASR[7]	H's[12]	MRR
MPIE	24.1%	92.2%	67.5%	**92.8%**
EYB	51.1%	**93.7%**	89.1%	**93.6%**

Table 3. Recognition rate vs. the number of training samples on the MPIE database

Sample number	3	5	7
RASR [7]	78.2%	95.8%	96.8%
MRR	**82.0%**	**97.5%**	**97.5%**

5.3 Robustness Evaluation

We first evaluate MRR's robustness to the number of training samples in comparison with RASR [7] on Multi-PIE. The first 100 subjects in Session 1 and Session 3 are used as the training and testing sets, respectively. For each person, 7 frontal images with the same illuminations as those in Section 5.2 are used for training, while 4 frontal images with illuminations $\{3, 6, 11, 19\}$ are used for testing. Three tests with the first 3, 5 and 7 training samples per person are performed. The recognition results of MRR and RSAR versus the number of training samples are shown in Table 3. We can see that MRR is better than RASR in all cases (about 3.8%, 1.7% and 0.7% improvement in the cases of 3, 5, and 7 training samples, respectively), which shows that MRR is more robust to the small sample size (SSS) problem. The reason is that when the number of training samples per subject is small, the subspace of one specific subject cannot be well built, and hence the performance of RASR will be much reduced since it works subject by subject; in contrast, MRR uses the training samples from all subjects to collaboratively represent the query sample, which can alleviate much the SSS problem [13], making it more robust to sample size than RASR.

Next we test the robustness of MRR to various levels of block occlusion on Multi-PIE. A randomly located block of the face image is replaced by the image Baboon. As [7], the training set remains the same as before, while the frontal images with illumination {10} from Session 1 are used for testing. The comparison of MRR and RASR is listed in Table 4. We see that MRR has very similar recognition rates to RSAR, and both of them still have good recognition accuracy up to 30% occlusion.

5.4 Face Recognition and Validation

In this section, a large-scale Multi-PIE face dataset [22] with 337 subjects is used for practical face recognition and verification tests, where all the 249 subjects

Table 4. Recognition rates (%) under various levels of random block occlusion

Percent	10%	20%	30%	40%	50%
RASR [7]	99.6	94.9	79.6	46.5	19.8
MRR	99.6	95.2	79.5	43.4	20.1

Table 5. Face recognition on the Multi-PIE database

Rec. Rates	Session2	Session3	Session4
RASR [7]	**93.9%**	**93.8%**	92.3%
MRR	**93.7%**	92.8%	**93.0%**

in Session 1 are used as training samples and the remaining 88 subjects are used as "imposters", or invalid images. For each of the training subjects, the 7 frontal images with the same illuminations as before are used for training. Here the images are resized to 60×48. It should be noted that all the testing face images are automatically detected by the Viola and Jones' face detector [7], and a rectangular window is used to crop the facial region.

For face recognition experiments, as [7] the frontal images with all 20 illuminations from Sessions 2-4 (recorded at different times over a period of several months) are used as testing samples. The recognition rates of MRR and RSAR are presented in Table 5. We can see that MRR has very similar recognition rate to RASR in average. Specifically, MRR is better than RASR in Session 4 with 0.7% improvement, almost the same as RASR in Session 2, and slightly worse than RASR in Session 3 with 1.0% gap.

For face validation experiments, from Session 2 we choose the subjects appearing in Session 1 as the customers. The 88 imposters are from Session 2 (37 subjects with ID between 251 and 292) and Session 3 (51 subjects with ID between 293 and 346) with 10 even-number illuminations for each subject. To be identical to [7], we also use the Sparsity Concentration Index (SCI) proposed in [8] to do validation after getting the coding coefficients. With the alignment method in [7], Nearest Neighbor (NN) and Nearest Subspace (NS) are also employed for comparison. Fig. 3 plots the receiver operating characteristic (ROC) curves by sweeping the threshold through the entire range of possible values for each algorithm. It can be seen that MRR and RASR significantly outperform NN and NS. MRR generally has very similar performance to RASR, although they locally across each other. For instance, the true positive rates of MRR and RASR are 87.5% and 86.7%, respectively, when the false positive rate is 5%; and the true positive rates of MRR and RASR are 91.8% and 92.5%, respectively, when the false positive rate is 15%.

5.5 Running Time

From Sections 5.2∼5.4, we can see that MRR and RASR achieve almost the same results in various tests. Then let's compare their running time, which is one of the most important concerns in practical FR systems.

We do face recognitions on Multi-PIE with the same experimental setting as that in Section 5.4 except that the number of subjects is set as 10, 50, 100, 150, 200, and 249, respectively. The programming environment is Matlab version 2011a. The desktop used is of 3.16 GHz CPU and with 3.25G RAM. The average running time of MRR and RASR (our reimplementation) is listed in Table 6.

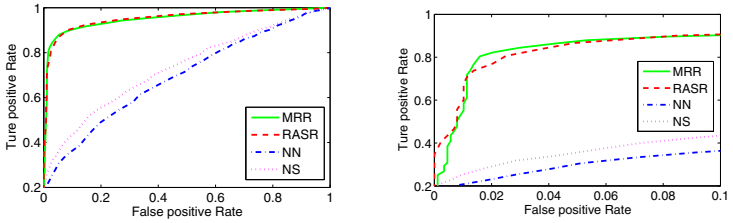


Fig. 3. ROC curves with false positive rate as $0 \sim 1$ (left) and $0 \sim 0.1$ (right) for subject verification on the Multi-PIE database

Table 6. The average running time (second) of MRR and RASR vs. subject number

Num	10	50	100	150	200	249
RASR	6.1	30.5	60.8	91.6	122.2	152.7
MRR	**0.59**	**0.81**	**0.84**	**0.87**	**0.90**	**0.91**
Speedup	10.3	37.7	72.4	105.3	135.8	167.8

It can be seen that in all cases, the running time of MRR is less than 1 second, validating that MRR is suitable for real-time FR systems. Compared to MRR, the running time of RASR is much longer, over 1 minute when the number of subjects is more than 100. Especially, we can see that RASR's running time linearly increases with the number of subjects, while the running time of MRR is nearly independent of the number of subjects. The speedup of MRR to RASR is more than half of the number of subjects, for example, 105 times speedup when the number of subjects is 150. This accords with our time complexity analysis in Section 4 very well.

6 Conclusion

We proposed a novel misalignment-robust representation (MRR) model in order for real-time face recognition. An efficient two-step optimization algorithm with a coarse-to-fine search strategy was developed to implement MRR. MRR has strong robustness to face misalignment coupled with illumination variation and occlusions, and more importantly, it can do face recognition at a real-time speed (less than 1 second under the Matlab programming environment). We evaluated the proposed MRR on various misaligned face recognition and verification tasks. The extensive experimental results clearly demonstrated that MRR could achieve similar accuracy to state-of-the-arts but with much faster speed, making it a good candidate for use in real-time face recognition systems.

References

1. Zhao, W., Chellppa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Survey 35, 399–458 (2003)
2. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report Technical Report 07-49, University of Massachusetts (2007)

3. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV (2009)
4. Nowak, E., Jurie, F.: Learning visual similarity measures for comparing never seen objects. In: CVPR (2007)
5. Huang, G.B., Jain, V., Learned-Miller, E.: Unsupervised joint alignment of complex images. In: ICCV (2007)
6. Yin, Q., Tang, X.O., Sun, J.: An associate-predict model for face recognition. In: CVPR (2011)
7. Wagner, A., Wright, J., Ganesh, A., Zhou, Z.H., Mobahi, H., Ma, Y.: Towards a practical face recognition system: robust alignment and illumination by sparse representation. IEEE PAMI 34, 372–386 (2012)
8. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE PAMI 31, 210–227 (2009)
9. Elhamifar, E., Vidal, R.: Robust classification using structured sparse representation. In: CVPR (2011)
10. Yang, M., Zhang, L., Yang, J., Zhang, D.: Robust sparse coding for face recognition. In: CVPR (2011)
11. Zhou, Z., Wagner, A., Mobahi, H., Wright, J., Ma, Y.: Face recognition with contiguous occlusion using markov random fields. In: ICCV (2009)
12. Huang, J.Z., Huang, X.L., Metaxas, D.: Simultaneous image transformation and sparse representation recovery. In: CVPR (2008)
13. Zhang, L., Yang, M., Feng, X.C.: Sparse representation or collaborative representation which helps face recognition? In: ICCV (2011)
14. Yan, S.C., Wang, H., Liu, J.Z., Tang, X.O., Huang, T.S.: Misalignment robust face recognition. IEEE IP 19, 1087–1096 (2010)
15. Viola, P., Jones, M.J.: Robust real-time face detection. Int'l J. Computer Vision 57, 137–154 (2004)
16. Shan, S., Chang, Y., Gao, W., Cao, B., Yang, P.: Curse of mis-alignment in face recognition: problem and a novel mis-alignment learning solution. In: AFGR (2004)
17. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE PAMI 23, 681–685 (2001)
18. Cootes, T., Taylor, C.: Active shape models - 'smart snakes'. In: BMVC (1992)
19. Peng, Y.G., Ganesh, A., Wright, J., Xu, W.L., Ma, Y.: RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In: CVPR (2010)
20. Yang, A.Y., Ganesh, A., Zhou, Z.H., Sastry, S.S., Ma, Y.: Fast l_1-minimizaton algorithms and application in robust face recognition. Technical Report UCB/EECS-2010-13, UC Berkeley (2010)
21. Kim, S.J., Koh, K., Lustig, M., Boyd, S., Gorinevsky, D.: A interior-point method for large-scale l_1-regularized least squares. IEEE Journal on Selected Topics in Signal Processing 1, 606–617 (2007)
22. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-PIE. Image and Vision Computing 28, 807–813 (2010)
23. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE PAMI 23, 643–660 (2001)