1 Face Recognition by Sparse Representation

Arvind Ganesh, Andrew Wagner, Zihan Zhou
Coordinated Science Lab, University of Illinois, Urbana, USA
Allen Y. Yang
Department of EECS, University of California, Berkeley, USA
Yi Ma and John Wright
Visual Computing Group, Microsoft Research Asia, Beijing, China

In this chapter, we present a comprehensive framework for tackling the classical problem of face recognition, based on theory and algorithms from sparse representation. Despite intense interest in the past several decades, traditional pattern recognition theory still stops short of providing a satisfactory solution capable of recognizing human faces in the presence of real-world nuisances such as occlusion and variabilities in pose and illumination. Our new approach, called sparse representation-based classification (SRC), is motivated by a very natural notion of sparsity, namely, one should always try to explain a query image using a small number of training images from a single subject category. This sparse representation is sought via ℓ_1-minimization. We show how this core idea can be generalized and extended to account for various physical variabilities encountered in face recognition. The end result of our investigation is a full-fledged practical system aimed at security and access control applications. The system is capable of accurately recognizing subjects out of a database of several hundred subjects with state-of-the-art accuracy.

1.1 Introduction

Automatic face recognition is a classical problem in the computer vision community. The community's sustained interest in this problem is mainly due to two reasons. First, in face recognition, we encounter many of the common variabilities that plague vision systems in general: illumination, occlusion, pose, and misalignment. Inspired by the good performance of humans in recognizing familiar faces [38], we have reason to believe that effective automatic face recognition is possible, and that the quest to achieve

Figure 1.1 Examples of image nuisances in face recognition. **Left**: Illumination change. **Middle Left**: Pixel corruption. **Middle Right**: Facial disguise. **Right:** Occlusion and misalignment.

this will tell us something about visual recognition in general. Second, face recognition has a wide spectrum of practical applications. Indeed, if we could construct an extremely reliable automatic face recognition system, it would have broad implications for identity verification, access control, security, and public safety. In addition to these classical applications, the recent proliferation of online images and videos has provided a host of new applications such as image search and photo tagging (e.g. Google's Piccasa and Apple FaceTime).

Despite several decades of work in this area, high-quality automatic face recognition remains a challenging problem that defies satisfactory solutions. While there has been steady progress on scalable algorithms for face recognition in low-stake applications such as photo album organization[1], there have been a sequence of well-documented failed trials of face recognition technology in mass surveillance/watch-list applications, where the performance requirements are very demanding.[2] These failures are mainly due to the challenging structure of face data: any real-world face recognition system must simultaneously deal with variables and nuisances such as illumination variation, corruption and occlusion, and reasonable amount of pose and image misalignment. Some examples of these image nuisances for face recognition are illustrated in Figure 1.1.

Traditional pattern recognition theory stops short of providing a satisfactory solution capable of simultaneously addressing all of these problems. In the past decades, numerous methods for handling a single mode of variability, such as pose or illumination, have been proposed and examined. But much less work has been devoted to simultaneously handling multiple modes of variation, according to a recent survey [52].[3] In other words, although a method might successfully deal with one type of variation, it quickly breaks down when moderate amounts of other variations are introduced to face images.

[1] As documented, e.g., in the ongoing Labeled Faces in the Wild [26] challenge. We invite the interested reader to consult this work and the references therein.

[2] A typical performance metric that is considered acceptable for automatic mass surveillance may require both a recognition rate in high 90's and a false positive rate lower than 0.01% over a database with thousands of subjects.

[3] The literature on face recognition is vast, and doing justice to all ideas and proposed algorithms would require a separate survey of comparable length to this chapter. In the course of this chapter, we will review a few works necessary to put ours in context. We refer the reader to [52] for a more comprehensive treatment of the history of the field.

Recently, the theory of sparse representation and compressed sensing has shed some new light on this challenging problem. Indeed, there is a very natural notion of sparsity in the face recognition problem: one always tries to find only a single subject out of a large database of subjects that best explains a given query image. In this chapter, we will discuss how tools from compressed sensing, especially ℓ_1-minimization and random projections, have inspired new algorithms for face recognition. In particular, the new computational framework can simultaneously address the most important types of variation in face recognition.

Nevertheless, face recognition diverges quite significantly from the common compressed sensing setup. On the mathematical side, the data matrices arising in face recognition often violate theoretical assumptions such as the restricted isometry property or even incoherence. Moreover, the physical structure of the problem (especially misalignment) will occasionally force us to solve the sparse representation problem subject to certain *nonlinear* constraints.

On the practical side, face recognition poses new non-trivial challenges in algorithm design and system implementation. First, face images are very high-dimensional data (e.g., a 1000×1000 gray-scale image has 10^6 pixels). Largely due to lack of memory and computational resource, dimensionality reduction techniques have largely been considered as a necessary step in the conventional face recognition methods. Notable holistic feature spaces include Eigenfaces [42], Fisherfaces [3], Laplacianfaces [25] and their variants [29, 10, 47, 36]. Nevertheless, it remains an open question: what is the optimal low-dimensional facial feature space that is capable of pairing with any well-designed classifier and leads to superior recognition performance?

Second, past face recognition algorithms often work well under laboratory conditions, but their performance would degrade drastically when tested in less-controlled environments – partially explaining some of the highly publicized failures of these systems. A common reason is that those face recognition systems were only tested on images taken under the same laboratory conditions (even with the same cameras) as the training images. Hence, their training sets do not represent well variations in illumination for face images taken under different indoor and outdoor environments, and under different lighting conditions. In some extreme cases, certain algorithms have attempted to reduce the illumination effect from only a single training image per subject [12, 53]. Despite these efforts, truly illumination-invariant features are in fact impossible to obtain from a few training images, let alone a single image [21, 4, 1]. Therefore, a natural question arises: How can we improve the image acquisition procedure to guarantee sufficient illuminations in the training images that can represent a large variety of real-world lighting conditions?

In this chapter, under the overarching theme of the book, we provide a systematic exposition of our investigation over the past few years into a new mathematical approach to face recognition, which we call *sparse representation-based classification* (SRC). We will start from a very simple, almost simplistic, problem formulation that is directly inspired by results in compressed sensing. We will see generalization of this approach naturally accounts for the various physical variabilities in the face recognition problem. In turn, we will see some of the new observations that face recognition can contribute to

the mathematics of compressed sensing. The end result will be a full-fledged practical system aimed at applications in access control. The system is capable of accurately recognizing subjects out of a database of several hundred with high accuracy, despite large variations in illumination, moderate occlusion, and misalignment.

1.1.1 Organization of this chapter

In Section 1.2, starting with the simplest possible problem setting, we show how face recognition can be posed as a sparse representation problem. Section 1.3 discusses the possibility of solving this problem more efficiently by projecting the data into a randomly selected lower-dimensional feature space. In Sections 1.4 and 1.5, we then show how the SRC framework can be naturally extended to handle physical variabilities such as occlusion and misalignment, respectively. Section 1.6 and 1.7 discuss practical aspects of building a face recognition system using the tools introduced here. Section 1.6 shows how to efficiently solve sparse representation problems arising in face recognition, while Section 1.7 discusses a practical system for acquiring training images of subjects under different illuminations. In Section 1.8, we combine these developments to give an end-to-end system for face recognition, aimed at access control tasks.

1.2 Problem Formulation: Sparse Representation-based Classification

In this section, we first illustrate the core idea of SRC in a slightly artificial scenario in which the training and test images are very well-aligned, but do contain significant variations in illumination. We will see how in this setting, face recognition can be naturally cast as a sparse representation problem, and solved via ℓ_1-minimization. In subsequent sections, we will see how this core formulation extends naturally to handle other variabilities in real-world face images, culminating in a complete system for recognition described in Section 1.7.

In face recognition, the system is given access to a set of labeled training images $\{\boldsymbol{\phi}_i, l_i\}$ from the C subjects of interest. Here, $\boldsymbol{\phi}_i \in \mathbb{R}^m$ is the vector representation of a digital image (say, by stacking the $W \times H$ columns of the single-channel image as a $m = W \times H$ dimensional vector), and $l_i \in \{1 \dots C\}$ indicates which of the C subjects is pictured in the i-th image. In the testing stage, a new query image $\boldsymbol{y} \in \mathbb{R}^m$ is provided. The system's job is to determine which of the C subjects is pictured in this query image, or, if none of them is present, to reject the query sample as invalid.

Influential results due to [1, 21] suggest that if sufficiently many images of the same subject are available, these images will lie close to a low-dimensional linear subspace of the high-dimensional image space $\mathbb{R}^m$. The required dimension could be as low as nine for a convex, Lambertian object [1]. Hence, given sufficient diversity in training illuminations, the new test image $\boldsymbol{y}$ of subject i can be well represented as a linear

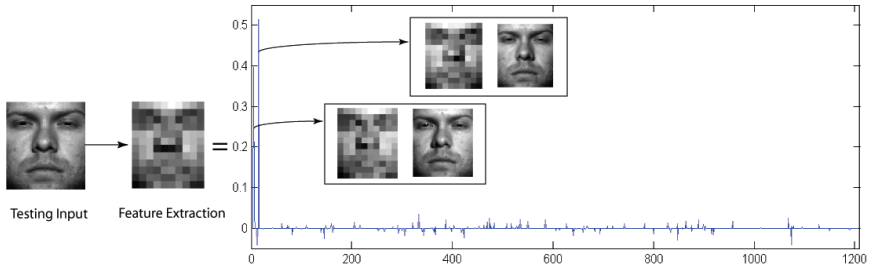

Figure 1.2 Sparse representation of a 12×10 down-sampled query image based on about 1,200 training images of the same resolution. The query image belongs to Class 1 [46].

combination of the training images of the same subject:

$$\boldsymbol{y} \approx \sum_{\{j|l_j=i\}} \boldsymbol{\phi}_j c_j \doteq \Phi_i \boldsymbol{c}_i, \tag{1.1}$$

where $\Phi_i \in \mathbb{R}^{m \times n_i}$ concatenates all of the images of subject i, and $\boldsymbol{c}_i \in \mathbb{R}^{n_i}$ is the corresponding vector of coefficients. In Section 1.7, we will further describe how to select the training samples $\boldsymbol{\phi}$ to ensure the approximation in (1.1) is accurate in practice.

In the testing stage, we are confronted with the problem that the class label i is unknown. Nevertheless, one can still form a linear representation similar to (1.1), now in terms of *all* of the training samples:

$$\boldsymbol{y} = [\Phi_1, \Phi_2, \cdots, \Phi_C]\boldsymbol{c}_0 = \Phi \boldsymbol{c}_0 \in \mathbb{R}^m, \tag{1.2}$$

where

$$\boldsymbol{c}_0 = [\cdots, \mathbf{0}^T, \boldsymbol{c}_i^T, \mathbf{0}^T, \cdots]^T \in \mathbb{R}^n. \tag{1.3}$$

Obviously, if we can recover a vector $\boldsymbol{c}$ of coefficients concentrated on a single class, it will be very indicative of the identity of the subject.

The key idea of SRC is to cast face recognition as the quest for such a coefficient vector $\boldsymbol{c}_0$. We notice that because the nonzero elements in $\boldsymbol{c}$ are concentrated on images of a single subject, $\boldsymbol{c}_0$ is a highly *sparse* vector: on average only a fraction of $\frac{1}{C}$ of its entries are nonzero. Indeed, it is not difficult to argue that in general this vector is the sparsest solution to the system of equations $\boldsymbol{y} = \Phi \boldsymbol{c}_0$. While the search for sparse solutions to linear systems is a difficult problem in general, foundational results in the theory of sparse representation indicate that in many situations the sparsest solution can be exactly recovered by solving a tractable optimization problem, minimizing the ℓ_1-norm $\|\boldsymbol{c}\|_1 \doteq \sum_i |c_i|$ of the coefficient vector (see [15, 8, 13, 6] for a sampling of the theory underlying this relaxation). This suggests seeking $\boldsymbol{c}_0$ as the unique solution to the optimization problem

$$\min \ \|\boldsymbol{c}\|_1 \quad \text{s.t.} \quad \|\boldsymbol{y} - \Phi \boldsymbol{c}\|_2 \leq \varepsilon. \tag{1.4}$$

Here, $\varepsilon \in \mathbb{R}$ reflects the noise level in the observation.

Figure 1.2 shows an example of the coefficient vector $\boldsymbol{c}$ recovered by solving the problem (1.4). Notice that the nonzero entries indeed concentrate on the (correct) first subject class, indicating the identity of the test image. In this case, the identification is correct even though the input images are so low-resolution (12×10!) that system of equations $\boldsymbol{y} \approx \Phi \boldsymbol{c}$ is underdetermined.

Once the sparse coefficients $\boldsymbol{c}$ have been recovered, tasks such as recognition and validation can be performed in a very natural manner. For example, one can simply define the concentration of a vector $\boldsymbol{c} = [\boldsymbol{c}_1^T, \boldsymbol{c}_2^T, \ldots, \boldsymbol{c}_C^T]^T \in \mathbb{R}^n$ on a subject i as

$$\alpha_i \doteq \|\boldsymbol{c}_i\|_1 / \|\boldsymbol{c}\|_1. \tag{1.5}$$

One can then assign to test image $\boldsymbol{y}$ the label i that maximizes α_i, or reject $\boldsymbol{y}$ as not belonging to any subject in the database if the maximum value of α_i is smaller than a predetermined threshold. For more details, as well as slightly more sophisticated classification schemes based on the sparse coefficients $\boldsymbol{c}$, please see [46, 44].

In the remainder of this chapter, we will see how this idealized scheme can be made practical by showing how to recover the sparse coefficients $\boldsymbol{c}_0$ even if the test image is subject to additional variations such as occlusion and misalignment. We further discuss how to reduce the complexity of the optimization problem (1.4) for large-scale problems.

1.3 Dimensionality Reduction

One major obstacle to large-scale face recognition is the sheer scale of the training data: the dimension of the raw images could be in the millions, while the number of images increases in proportion to number of subjects. The size of the data is directly reflected in the computational complexity of the algorithm - the complexity of convex optimization could be cubic or worse. In pattern recognition, a classical technique for addressing the problem of high dimensionality is to project the data into a much lower-dimensional feature space $\mathbb{R}^d$ ($d \ll m$), such that the projected data still retain the useful properties of the original images.

Such projections can be generated via principal component analysis [43], linear discriminant analysis [3], locality preserving projections [25], as well as less-conventional transformations such as downsampling or selecting local features (e.g., the eye or mouth regions). These well-studied projections can all be represented as linear maps $A \in \mathbb{R}^{d \times m}$. Applying such a linear projection gives a new observation

$$\tilde{\boldsymbol{y}} \doteq A\boldsymbol{y} \approx A\Phi\boldsymbol{c} = \Psi\boldsymbol{c} \in \mathbb{R}^d. \tag{1.6}$$

Notice that if d is small, the solution to the system $A\Phi\boldsymbol{c} = \tilde{\boldsymbol{y}}$ may not be unique. Nevertheless, under generic circumstances, the desired *sparsest* solution $\boldsymbol{c}_0$ to this system *is* unique, and can be sought via a lower complexity convex optimization

$$\min \|\boldsymbol{c}\|_1 \quad \text{s.t.} \quad \left\|\Psi\boldsymbol{c} - \tilde{\boldsymbol{y}}\right\|_2 \leq \varepsilon. \tag{1.7}$$

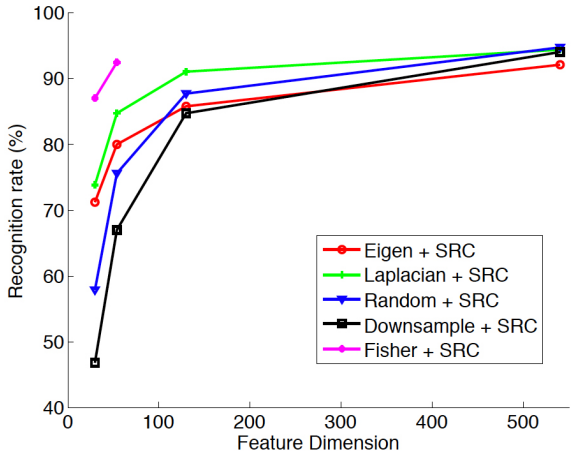

Figure 1.3 Recognition rates of SRC for various feature transformations and dimensions [46]. The training and query images are selected from the public AR database.

Then the key question is to what extent the choice of transformation A affects our ability to recover $\boldsymbol{c}_0$ and subsequently recognize the subject.

The many different projections referenced above reflect a long-term effort within the face recognition community to find the best possible set of data-adaptive projections. On the other hand, one of the key observations of compressed sensing is that *random projections* serve as a universal non-adaptive set of projections [9, 16, 17]. If a vector $\boldsymbol{c}$ is sparse in a known orthobasis, then ℓ_1-minimization will recover $\boldsymbol{c}$ from relatively small sets of random observations, with high probability. Although our matrix Φ is not an orthonormal basis (far from it, as we will see in the next section), it is still interesting to investigate random projections for dimensionality reduction, and to see to what extent the choice of features affects the performance of ℓ_1-minimization in this application.

Figure 1.3 shows a typical comparison of recognition rates across a variety of feature transformations A and feature space dimensions d. The data in Figure 1.3 are taken from the AR face database, which contains images of 100 subjects under a variety of conditions [33].[4] The horizontal axis plots the feature space dimension, which varies from 30 to 540. The results, which are consistent with other experiments on a wide range of databases, show that the choice of an optimal feature space is no longer critical. When ℓ_1-minimization is capable of recovering sparse signals in several hundred dimensional feature spaces, the performance of all tested transformations converges to a reasonably high percentage. More importantly, even random features contain enough information to recover the sparse representation and hence correctly classify most query images. Therefore, what is critical is that the dimension of the feature space is sufficiently large, and that the sparse representation is correctly computed.

It is important to note that reducing the dimension typically leads to decrease in the recognition rate; although that decrease is not large when d is sufficiently large. In large-

[4] For more detailed information on the experimental setting, please see [46].

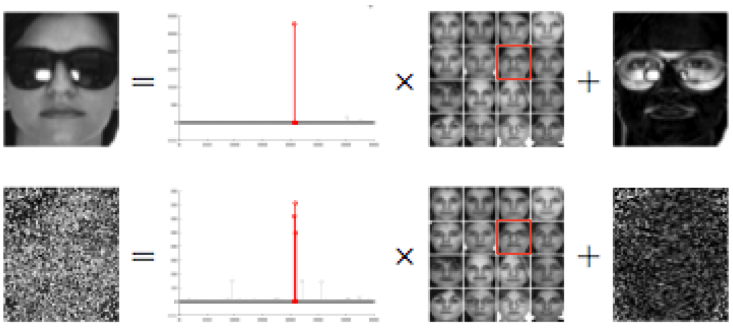

Figure 1.4 Robust face recognition via sparse representation. The method represents a test image (left), which is partially occluded (top) or corrupted (bottom), as a sparse linear combination of all the normal training images (middle) plus sparse errors (right) due to occlusion or corruption. Coefficients in red correspond to training images of the correct individual. Our algorithm determines the true identity (indicated with a red box at second row and third column) from 700 training images of 100 individuals in the standard AR face database. [46].

scale applications where this tradeoff is inevitable, the implication of our investigation is that a variety of features can confidently used in conjunction with ℓ_1-minimization. On the other hand, if highly accurate face recognition is desired, the original images themselves can be used as features, in a way that is robust to additional physical nuisances such as occlusion and geometric transformations of the image.

1.4 Recognition with Corruption and Occlusion

In many practical scenarios, the face of interest may be partially occluded, as shown in Figure 1.4. The image may also contain large errors due to self-shadowing, specularities, or corruption. Any of these image nuisances may cause the representation to deviate from the linear model $\boldsymbol{y} \approx \Phi \boldsymbol{c}$. In realistic scenarios, we are more likely confronted with an observation that can be modeled as

$$\boldsymbol{y} = \Phi \boldsymbol{c} + \boldsymbol{e}, \tag{1.8}$$

where $\boldsymbol{e}$ is an unknown vector whose nonzero entries correspond to the corrupted pixels in the observation $\boldsymbol{y}$, as shown in Figure 1.4.

The errors $\boldsymbol{e}$ can be large in magnitude, and hence cannot be ignored or treated with techniques designed for small noise, such as least squares. However, like the vector $\boldsymbol{c}$, they often are *sparse*: occlusion and corruption generally affect only a fraction $\rho < 1$ of the image pixels. Hence, the problem of recognizing occluded faces can be cast as the search for a sparse representation $\boldsymbol{c}$, up to a sparse error $\boldsymbol{e}$. A natural robust extension to the SRC framework is to instead solve a combined ℓ_1-minimization problem

$$\min \ \|\boldsymbol{c}\|_1 + \|\boldsymbol{e}\|_1 \ \text{ s.t. } \ \boldsymbol{y} = \Phi \boldsymbol{c} + \boldsymbol{e}. \tag{1.9}$$

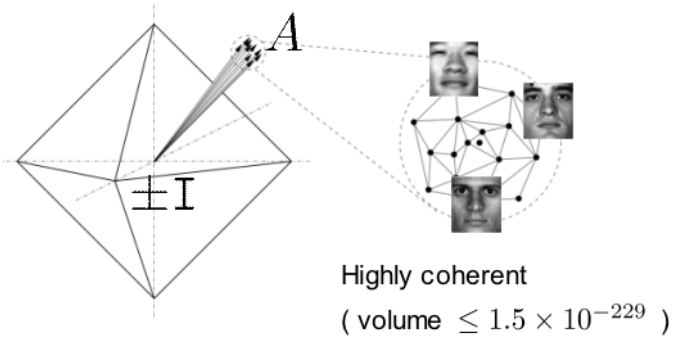

Figure 1.5 The cross-and-bouquet model for face recognition. The raw images of human faces expressed as columns of A are clustered with small variance [45].

In [46], it was observed that this optimization performs quite well in correcting occlusion and corruption, for instance, for block occlusions covering up to 20% of the face and random corruptions affecting more than 70% of the image pixels.

Nevertheless, on closer inspection, the success of the combined ℓ_1-minimization in (1.9) is surprising. One can interpret (1.9) as an ℓ_1-minimization problem against a single combined dictionary $B \doteq [\Phi\ I] \in \mathbb{R}^{m\times(n+m)}$:

$$\min\ \|\boldsymbol{w}\|_1\ \text{ s.t. }\ \boldsymbol{y} = B\boldsymbol{w}, \tag{1.10}$$

where $\boldsymbol{w} = [\boldsymbol{c}^T, \boldsymbol{e}^T]^T$. Because the columns of Φ are all face images, and hence somewhat similar in the high-dimensional image space $\mathbb{R}^m$, the matrix B fairly dramatically violates the classical conditions for uniform sparse recovery, such as the incoherence criteria [15] or the restricted isometry property [8]. In contrast to the classical compressed sensing setup, the matrix B has quite inhomogeneous properties: the columns of Φ are coherent in the high-dimensional space, while the columns of I are as incoherent as possible. Figure 1.5 illustrates the geometry of this rather curious object, which was dubbed a "cross-and-bouquet" (CAB) in [45], due to the fact that the columns of the identity matrix span a cross polytope, whereas the columns of A are tightly clustered like a bouquet of flowers.

In sparse representation, the CAB model belongs to a special class of sparse representation problems where the dictionary Φ is a concatenation of sub-dictionaries. Examples include the merger of wavelet and heaviside dictionaries in [11] and the combination of texture and cartoon dictionaries in morphological component analysis [18]. However, in contrast to most existing examples, not only is our new dictionary B inhomogeneous, in fact the solution $(\boldsymbol{c}, \boldsymbol{e})$ to be recovered is also very inhomogeneous: the sparsity of $\boldsymbol{c}$ is limited by the number of images per subject, whereas we would like to handle as dense $\boldsymbol{e}$ as possible, to guarantee good error correction performance. Simulations (similar to the bottom row of Figure 1.4) have suggested that in fact the error $\boldsymbol{e}$ can be quite dense, provided its signs and support are random [46, 45]. In [45], it is shown that

As long as the bouquet is sufficiently tight in the high-dimensional image space $\mathbb{R}^m$, ℓ_1-minimization successfully recovers the sparse coefficients $\boldsymbol{x}$ from very dense ($\rho \nearrow 1$) randomly signed errors $\boldsymbol{e}$.

For a more precise statement and proof of this result, we refer the reader to [45].

For our purposes here, it suffices to say that this result suggests that excellent error correction is possible in circumstances quite similar to the ones encountered in real-world face recognition. This is surprising for two reasons. First, as mentioned above, the "dictionary" in this problem dramatically violates the restricted isometry property. Second, the errors corrected can be quite dense, in contrast to typical results from compressed sensing in which the number of nonzero coefficients recovered (or errors corrected) is typically bounded by a small fraction of the dimension m [8, 17]. Interestingly, while the mathematical tools needed to analyze this problem are quite standard in this area, the results obtained are qualitatively different from classical results in compressed sensing. Thus, while classical results such as [8, 14] are inspiring for face recognition, the structure of the matrices encountered in this application gives it a mathematical flavor all its own.

1.5 Face Alignment

The problem formulation in the previous sections allows us to simultaneously cope with illumination variation and moderate occlusion. However, a practical face recognition system needs to deal with one more important mode of variability: misalignment of the test image and training images. This may occur if the face is not perfectly localized in the image, or if the pose of the face is not perfectly frontal. Figure 1.6 shows how even small misalignment can cause appearance-based face recognition algorithms (such as the one described above) to break down. In this section, we will see how the framework of the previous sections naturally extends to cope with this difficulty.

To pose the problem, we assume the observation $\boldsymbol{y}$ is a warped image $\boldsymbol{y} = \boldsymbol{y}_0 \circ \tau^{-1}$ of the ground-truth signal $\boldsymbol{y}_0$ under some 2-D transformation of domain τ.[5] As illustrated in Figure 1.6, when τ perturbs the detected face region away from the optimal position, directly solving a sparse representation of $\boldsymbol{y}$ against properly aligned training images often results in erroneous representation.

Nevertheless, if the true deformation τ can be efficiently found, then we can recover $\boldsymbol{y}_0$ and it becomes possible to recover a relevant sparse representation $\boldsymbol{c}$ for $\boldsymbol{y}_0$ with respect to the well-aligned training set. Based on the previous error correction model (1.8), the sparse representation model under face alignment is defined as

$$\boldsymbol{y} \circ \tau = \Phi \boldsymbol{c} + \boldsymbol{e}. \tag{1.11}$$

[5] In our system, we typically use 2-D similarity transformations, $T = \mathbb{SE}(2) \times \mathbb{R}_+$, for misalignment incurred by face cropping, or 2-D projective transformations, $T = \mathbb{GL}(3)$, for pose variation.

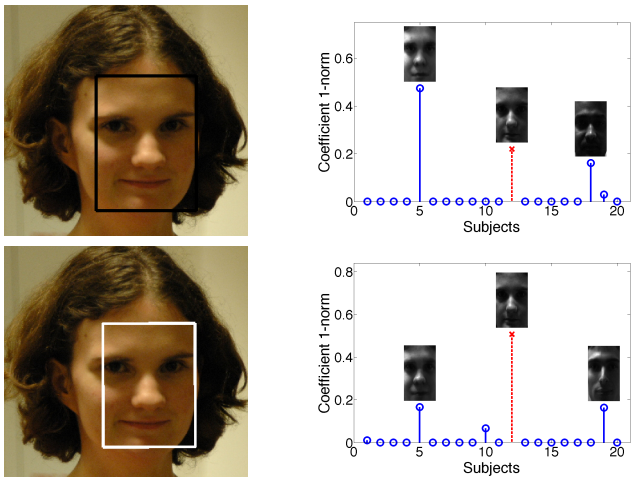

Figure 1.6 Effect of face alignment [44]. The task is to identify the girl among 20 subjects, by computing the sparse representation of her input face with respect to the entire training set. The absolute sum of the coefficients associated with each subject is plotted on the right. We also show the faces reconstructed with each subject's training images weighted by the associated sparse coefficients. The red line corresponds to her true identity, Subject 12. **Top:** The input face is from Viola and Jones' face detector (the black box). The estimated representation failed to reveal the true identity as the coefficients from Subject 5 are more significant. **Bottom:** The input face is well-aligned (the white box) with the training images by our alignment algorithm, and a better representation is obtained.

Naturally, one would like to use the sparsity as a strong cue for finding the correct deformation τ, solving the following optimization problem:

$$\min_{\boldsymbol{c},\boldsymbol{e},\tau} \|\boldsymbol{c}\|_1 + \|\boldsymbol{e}\|_1 \;\; \text{s.t.} \;\; \boldsymbol{y} \circ \tau = \Phi \boldsymbol{c} + \boldsymbol{e}. \tag{1.12}$$

Unfortunately, simultaneously estimating τ and $(\boldsymbol{c}, \boldsymbol{e})$ in (1.12) is a difficult nonlinear optimization problem. In particular, in the presence of multiple classes in the matrix Φ, many local minima arise, which correspond to aligning $\boldsymbol{y}$ to different subjects in the database.

To mitigate the above two issues, it is more practical to first consider aligning $\boldsymbol{y}$ individually to each subject k:

$$\tau_k^* = \arg \min_{\boldsymbol{c},\boldsymbol{e},\tau_k} \|\boldsymbol{e}\|_1 \;\; \text{s.t.} \;\; \boldsymbol{y} \circ \tau_k = \Phi_k \boldsymbol{c} + \boldsymbol{e}. \tag{1.13}$$

Note that in (1.13), the sparsity of $\boldsymbol{c}$ is no longer penalized, since Φ_k only contains images of the same subject.

Second, if we have access to a good initial guess of the transformation (e.g., from the output of a face detector), the true transformation τ_k in (1.13) can be iteratively sought by solving a sequence of linearized approximations as follows:

$$\min_{\boldsymbol{c},\boldsymbol{e},\Delta\tau_k} \|\boldsymbol{e}\|_1 \;\; \text{s.t.} \;\; \boldsymbol{y} \circ \tau_k^i + J_k^i \cdot \Delta\tau_k = \Phi_k \boldsymbol{c} + \boldsymbol{e}, \tag{1.14}$$

Algorithm 1.1 (Deformable SRC for Face Recognition [44]).

1: **Input:** Frontal training images $\Phi_1, \Phi_2, \ldots, \Phi_C \in \mathbb{R}^{m \times n_i}$ for C subjects, a test image $\boldsymbol{y} \in \mathbb{R}^m$ and a deformation group T considered.
2: **for** each subject k,
3: $\quad \tau_k^0 \leftarrow I$.
4: $\quad$ **do**
5: $\quad\quad \tilde{\boldsymbol{y}}(\tau_k^i) \leftarrow \frac{\boldsymbol{y} \circ \tau_k^i}{\|\boldsymbol{y} \circ \tau_k^i\|_2}$; $\quad J_k^i \leftarrow \frac{\partial}{\partial \tau_k} \tilde{\boldsymbol{y}}(\tau_k)\big|_{\tau_k^i}$;
6: $\quad\quad \Delta\tau_k = \arg\min \|\boldsymbol{e}\|_1$ s.t. $\tilde{\boldsymbol{y}}(\tau_k^i) + J_k^i \Delta\tau_k = \Phi_k \boldsymbol{c} + \boldsymbol{e}$.
7: $\quad\quad \tau_k^{i+1} \leftarrow \tau_k^i + \Delta\tau_k$;
8: $\quad$ **while** $\|\tau_k^{i+1} - \tau_k^i\| \geq \varepsilon$.
9: **end**
10: Set $\Phi \leftarrow \left[\Phi_1 \circ \tau_1^{-1} \mid \Phi_2 \circ \tau_2^{-1} \mid \cdots \mid \Phi_C \circ \tau_C^{-1}\right]$.
11: Solve the ℓ_1-minimization problem:

$$\hat{\boldsymbol{c}} = \arg\min_{\boldsymbol{c},\boldsymbol{e}} \|\boldsymbol{c}\|_1 + \|\boldsymbol{e}\|_1 \text{ s.t. } \boldsymbol{y} = \Phi \boldsymbol{c} + \boldsymbol{e}.$$

12: Compute residuals $r_k(\boldsymbol{b}) = \|\boldsymbol{c} - \Phi_k \delta_k(\hat{\boldsymbol{c}})\|_2$ for $k = 1, \ldots, C$.
13: **Output:** identity$(\boldsymbol{y}) = \arg\min_k r_k(\boldsymbol{c})$.

where τ_k^i is the current estimate of the transformation τ_k, $J_k^i = \nabla_{\tau_k}(\boldsymbol{y} \circ \tau_k^i)$ is the Jacobian of $\boldsymbol{y} \circ \tau_k^i$ with respect to τ_k, and $\Delta\tau_k$ is a step update to τ_k. [6]

From the optimization point of view, our scheme (1.14) can been seen as a generalized Gauss-Newton method for minimizing the composition of a nonsmooth objective function (the ℓ_1-norm) with a differentiable mapping from transformation parameters to transformed images. It has been extensively studied in the literature and is known to converge quadratically in a neighborhood of any local optimum of the ℓ_1-norm [35, 27]. For our face alignment problem, we simply note that typically (1.14) takes 10 to 15 iterations to converge.

To further improve the performance of the algorithm, we can adopt a slightly modified version of (1.14), in which we replace the warped test image $\boldsymbol{y} \circ \tau_k$ with the normalized one $\tilde{\boldsymbol{y}}(\tau_k) = \frac{\boldsymbol{b} \circ \tau_k}{\|\boldsymbol{y} \circ \tau_k\|_2}$. This help to prevent the algorithm from falling into a degenerate global minimum corresponding to zooming in on a dark region of the test image. In practice, our alignment algorithm can run in a multi-resolution fashion in order to reduce the computational cost and gain a larger region of convergence.

Once the best transformation τ_k is obtained for each subject k, we can apply its inverse to the training set Φ_k so that the entire training set is aligned to $\boldsymbol{y}$. Then, a

[6] In computer vision literature, the basic iterative scheme for registration between two *identical* images related by an image transformation of a few parameters has been long known as the Lucas-Kanade algorithm [31]. Extension of the Lucas-Kanade algorithm to address the illumination issue in the same spirit as ours has also been exploited. However, most traditional solutions formulated the objective function using the ℓ_2-norm as a least squares problem. One exception prior to the theory of CS, to the best of our knowledge, was proposed in a robust face tracking algorithm by Hager and Belhumeur [24], where the authors used an iterative reweighted least squares (IRLS) method to iteratively remove occluded image pixels while the transform parameters of the face region were sought.

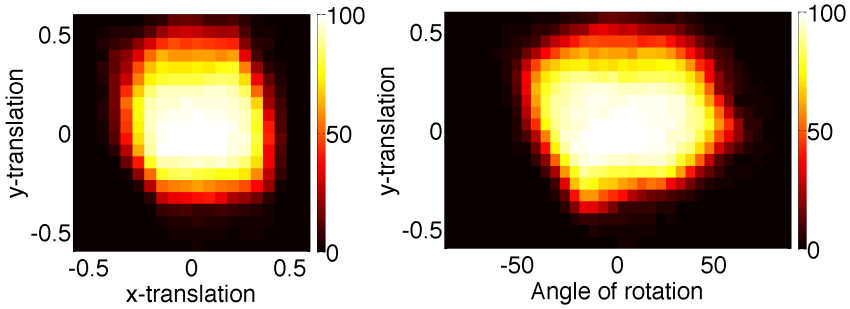

Figure 1.7 Region of attraction [44]. Fraction of subjects for which the algorithm successfully aligns a manually perturbed test image. The amount of translation is expressed as a fraction of the distance between the outer eye corners, and the amount of in-plane rotation in degrees. **Left:** Simultaneous translation in x and y directions. More than 90% of the subjects were correctly aligned for any combination of x and y translations, each up to 0.2. **Right:** Simultaneous translation in y direction and in-plane rotation θ. More than 90% of the subjects were correctly aligned for any combination of y translation up to 0.2 and θ up to 25°.

global sparse representation $\hat{\boldsymbol{c}}$ of $\boldsymbol{y}$ with respect to the transformed training set can be sought by solving an optimization problem of the form (1.9). The final classification is done by computing the ℓ_2 distance between $\boldsymbol{y}$ and its approximation $\hat{\boldsymbol{y}} = \Phi_k \delta_k(\hat{\boldsymbol{c}})$ [7] using only the training images from the k-th class, and assigning $\boldsymbol{y}$ to the class that minimize the distance. The complete algorithm is summarized in Algorithm 1.1.

Finally, we present some experimental results that characterizing the region of attraction of the proposed alignment procedure for both 2-D deformation and 3-D pose variation. We will leave the evaluation of the overall face recognition system to Section 1.8.

For 2-D deformation, we use a subset of images of 120 subjects from the CMU Multi-PIE database [23], since the ground-truth alignment is available. In this experiment, the training set consists of images under properly chosen lighting conditions, and the testing set contains one new illumination. We introduce artificial perturbation to each test image with a combination of translation and rotation, and use the proposed algorithm to align it to the training set of the same subject. For more details about the experiment setting, please refer to [44]. Figure 1.7 shows the percentage of successful registrations for all test images for each artificial perturbation. We can see that our algorithm performs very well with translation up to 20% of the eye distance (or 10 pixels) in both x and y directions, and up to 30° in-plane rotation. We have also tested our alignment algorithm with scale variation, and it can handle up to 15% change in scale.

For 3-D pose variation, we collect our own dataset using the acquisition system which will be introduced in Section 1.7. The training set includes frontal face images of each subject under 38 illuminations and the testing set contains images taken under densely sampled poses. Viola and Jones' face detector is then used for face cropping in this

[7] $\delta_k(\hat{\boldsymbol{c}})$ returns a vector of the same dimension as $\hat{\mathbf{c}}$ that only retains the nonzero coefficients corresponding to Subject k.

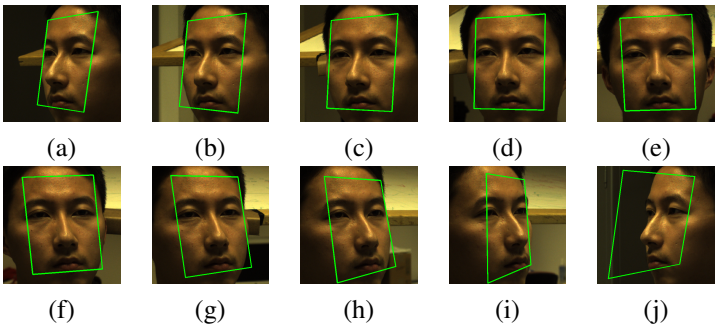

Figure 1.8 Aligning different poses to frontal training images [44]. (a) to (i): good alignment for poses from -45° to $+45^\circ$. **(j):** a case when the algorithm fails for an extreme pose ($> 45^\circ$).

experiment. Figure 1.8 shows some typical alignment results. The alignment algorithm works reasonably well with poses up to $\pm 45^\circ$, which easily exceeds the pose requirement for real-world access-control applications.

1.6 Fast ℓ_1-Minimization Algorithms

In the previous sections, we have seen how the problem of recognizing faces despite physical variabilities such as illumination, misalignment, and occlusion fall naturally into the framework of sparse representation. Indeed, all of these factors can be addressed simultaneously by solving appropriate ℓ_1-minimization problems. However, for these observations to be useful in practice, we need scalable and efficient algorithms for ℓ_1-minimization.

Although ℓ_1-minimization can be recast as a linear program and solved to high accuracy using interior-point algorithms [28], these algorithms do not scale well with the problem size: each iteration typically requires cubic time. Fortunately, interest in compressed sensing has inspired a recent wave of more scalable, more efficient first-order methods, which can solve very large ℓ_1-minimization problems to medium accuracy (see, e.g., [40] for a general survey). As we have seen in Section 1.4, ℓ_1-minimization problems arising in face recognition may have dramatically different structures from problems arising in other applications of compressed sensing, and hence require customized solvers. In this section, we describe our algorithm of choice for solving these problems, which is essentially an augmented Lagrange multiplier method [5], but also uses an accelerated gradient algorithm [2] to solve a key subproblem. We draw extensively on the survey [48], which compares the performance of various solvers in the context of face recognition.

The key property of the ℓ_1 norm that enables fast first-order solvers is the existence of an efficient solution to the "proximal minimization":

$$\mathcal{S}_\lambda[\boldsymbol{z}] = \arg\min_{\boldsymbol{x}} \lambda\|\boldsymbol{x}\|_1 + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2, \tag{1.15}$$

where $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$, and $\lambda > 0$. It is easy to show that the above minimization is solved by *soft-thresholding*, which is defined for scalars as follows:

$$\mathcal{S}_\lambda[x] = \begin{cases} x - \lambda, & \text{if } x > \lambda \\ x + \lambda, & \text{if } x < -\lambda \\ 0, & \text{if } |x| \leq \lambda \end{cases} \tag{1.16}$$

and extended to vectors and matrices by applying it element-wise. It is extremely simple to compute, and forms the backbone of most of the first-order methods proposed for ℓ_1-minimization. We will examine one such technique, namely, the method of *Augmented Lagrange Multipliers* (ALM), in this section. To keep the discussion simple, we focus our discussion on the SRC problem, although the ideas are directly applicable to the image alignment problem as well. The interested reader may refer to the Appendix of [48] for more details.

Lagrange multiplier methods are a popular tool in convex optimization. The basic idea is to eliminate equality constraints by adding an appropriate penalty term to the cost function that assigns a very high cost to infeasible points. The goal is then to efficiently solve the unconstrained problem. For our problem, we define the augmented Lagrangian function as follows:

$$L_\mu(\boldsymbol{c}, \boldsymbol{e}, \boldsymbol{\nu}) \doteq \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 + \langle \boldsymbol{\nu}, \boldsymbol{y} - \Phi\boldsymbol{c} - \boldsymbol{e} \rangle + \frac{\mu}{2} \|\boldsymbol{y} - \Phi\boldsymbol{c} - \boldsymbol{e}\|_2^2, \tag{1.17}$$

where $\mu > 0$, and $\boldsymbol{\nu}$ is a vector of Lagrange multipliers. Note that the augmented Lagrangian function is convex in $\boldsymbol{c}$ and $\boldsymbol{e}$. Suppose that $(\boldsymbol{c}^\star, \boldsymbol{e}^\star)$ is the optimal solution to the original problem. Then, it can be shown that for sufficiently large μ, there exists a $\boldsymbol{\nu}^\star$ such that

$$(\boldsymbol{c}^\star, \boldsymbol{e}^\star) = \arg\min_{\boldsymbol{c},\boldsymbol{e}} L_\mu(\boldsymbol{c}, \boldsymbol{e}, \boldsymbol{\nu}^\star). \tag{1.18}$$

The above property indicates that minimizing the augmented Lagrangian function amounts to solving the original constrained optimization problem. However, this approach does not seem a viable one since $\boldsymbol{\nu}^\star$ is not known a priori and the choice of μ is not evident from the problem. ALM methods overcome these issues by simultaneously solving for $\boldsymbol{\nu}^\star$ in an iterative fashion and monotonically increasing the value of μ every iteration so as to avoid converging to an infeasible point. The basic ALM iteration is given by [5]:

$$\begin{aligned} (\boldsymbol{c}_{k+1}, \boldsymbol{e}_{k+1}) &= \arg\min_{\boldsymbol{c},\boldsymbol{e}} L_{\mu_k}(\boldsymbol{c}, \boldsymbol{e}, \boldsymbol{\nu}_k), \\ \boldsymbol{\nu}_{k+1} &= \boldsymbol{\nu}_k + \mu_k(\boldsymbol{y} - \Phi\boldsymbol{c}_{k+1} - \boldsymbol{e}_{k+1}), \end{aligned} \tag{1.19}$$

where $\{\mu_k\}$ is a monotonically increasing positive sequence. This iteration by itself does not give us an efficient algorithm since the first step of the iteration is an unconstrained convex program. However, for the ℓ_1-minimization problem, we will see that it can be solved very efficiently.

The first step to simplifying the above iteration is to adopt an alternating minimization strategy, *i.e.,* to first minimize with respect to $\boldsymbol{e}$ and then minimize with respect to $\boldsymbol{c}$. This approach, dubbed *alternating direction method of multipliers* in [22], was first used

Algorithm 1.2 (Augmented Lagrange Multiplier Method for ℓ_1-minimization)

Input: $\boldsymbol{y} \in \mathbb{R}^m$, $\Phi \in \mathbb{R}^{m \times n}$, $\boldsymbol{c}_1 = \boldsymbol{0}$, $\boldsymbol{e}_1 = \boldsymbol{y}$, $\boldsymbol{\nu}_1 = \boldsymbol{0}$.
while not converged ($k = 1, 2, \ldots$) **do**
  $\boldsymbol{e}_{k+1} = \text{shrink}\left(\boldsymbol{y} - \Phi \boldsymbol{c}_k + \frac{1}{\mu_k}\boldsymbol{\nu}_k, \frac{1}{\mu_k}\right)$;
  $t_1 \leftarrow 1$, $\boldsymbol{z}_1 \leftarrow \boldsymbol{c}_k$, $\boldsymbol{w}_1 \leftarrow \boldsymbol{c}_k$;
  while not converged ($l = 1, 2, \ldots$) **do**
    $\boldsymbol{w}_{l+1} \leftarrow \text{shrink}\left(\boldsymbol{z}_l + \frac{1}{\gamma}\Phi^T\left(\boldsymbol{y} - \Phi \boldsymbol{v}_l - \boldsymbol{e}_{k+1} + \frac{1}{\mu_k}\boldsymbol{\nu}_k\right), \frac{1}{\mu_k \gamma}\right)$;
    $t_{l+1} \leftarrow \frac{1}{2}\left(1 + \sqrt{1 + 4t_l^2}\right)$;
    $\boldsymbol{z}_{l+1} \leftarrow \boldsymbol{w}_{l+1} + \frac{t_l - 1}{t_{l+1}}(\boldsymbol{w}_{l+1} - \boldsymbol{w}_l)$;
  end while
  $\boldsymbol{c}_{k+1} \leftarrow \boldsymbol{w}_l$;
  $\boldsymbol{\nu}_{k+1} \leftarrow \boldsymbol{\nu}_k + \mu_k(\boldsymbol{y} - \Phi \boldsymbol{c}_{k+1} - \boldsymbol{e}_{k+1})$;
end while
Output: $\boldsymbol{c}^* \leftarrow \boldsymbol{c}_k, \boldsymbol{e}^* \leftarrow \boldsymbol{e}_k$.

in [51] in the context of ℓ_1-minimization. Thus, the above iteration can be rewritten as:

$$\begin{aligned} \boldsymbol{e}_{k+1} &= \arg\min_{\boldsymbol{e}} L_{\mu_k}(\boldsymbol{c}_k, \boldsymbol{e}, \boldsymbol{\nu}_k), \\ \boldsymbol{c}_{k+1} &= \arg\min_{\boldsymbol{c}} L_{\mu_k}(\boldsymbol{c}, \boldsymbol{e}_{k+1}, \boldsymbol{\nu}_k), \\ \boldsymbol{\nu}_{k+1} &= \boldsymbol{\nu}_k + \mu_k(\boldsymbol{y} - \Phi \boldsymbol{c}_{k+1} - \boldsymbol{e}_{k+1}), \end{aligned} \tag{1.20}$$

Using the property described in (1.15), it is not difficult to show that

$$\boldsymbol{e}_{k+1} = \mathcal{S}_{\frac{1}{\mu_k}}\left[\frac{1}{\mu_k}\boldsymbol{\nu}_k + \boldsymbol{y} - \Phi \boldsymbol{c}_k\right]. \tag{1.21}$$

Obtaining a similar closed-form expression for $\boldsymbol{c}_{k+1}$ is not possible, in general. So, we solve for it in an iterative procedure. We note that $L_{\mu_k}(\boldsymbol{c}, \boldsymbol{e}_{k+1}, \boldsymbol{\nu}_k)$ can be split into two functions: $\|\boldsymbol{c}\|_1 + \|\boldsymbol{e}_{k+1}\|_1 + \langle \boldsymbol{\nu}_k, \boldsymbol{y} - \Phi \boldsymbol{c} - \boldsymbol{e}_{k+1} \rangle$ that is convex and continuous in $\boldsymbol{x}$; and $\frac{\mu_k}{2}\|\boldsymbol{y} - \Phi \boldsymbol{c} - \boldsymbol{e}_{k+1}\|_2^2$ that is convex, smooth and has Lipschitz continuous gradient. This form of the $L_{\mu_k}(\boldsymbol{c}, \boldsymbol{e}_{k+1}, \boldsymbol{\nu}_k)$ allows us to use a fast iterative thresholding algorithm, called FISTA [2], to solve for $\boldsymbol{c}_{k+1}$ in (1.20) efficiently. The basic idea in FISTA is to iteratively form quadratic approximations to the smooth part of the cost function and minimize the approximated cost function instead.

Using the above mentioned techniques, the iteration described in (1.20) is summarized as Algorithm 1.2, where γ denotes the largest eigenvalue of $\Phi^T\Phi$. Although the algorithm is composed of two loops, in practice, we find that the innermost loop converges in a few iterations.

As mentioned earlier, several first-order methods have been proposed for ℓ_1-minimization recently. Theoretically, there is no clear winner among these algorithms in terms of the convergence rate. However, it has been observed empirically that ALM offers the best trade-off in terms of speed and accuracy. An extensive survey of some of the other methods along with experimental comparison is presented in [48]. Compared to the classical interior-point methods, Algorithm 1.2 generally takes more iterations to

converge to the optimal solution. However, the biggest advantage of ALM is that each iteration is composed of very elementary matrix-vector operations, as against matrix inversions or Gaussian eliminations used in the interior-point methods.

1.7 Building a Complete Face Recognition System

In the previous sections, we have presented a framework for reformulating face recognition in terms of sparse representation, and have discussed fast ℓ_1-minimization algorithms to efficiently estimate sparse signals in high-dimensional spaces. In this section, we discuss some of the practical issues that arise in using these ideas to design prototype face recognition systems for access-control applications.

In particular, note that so far we have made the critical assumption that the test image, although taken under some unknown illumination, can be represented as a linear combination of a finite number of training illuminations. These assumptions naturally raise the following questions: *Under what conditions is the linear subspace model a reasonable assumption, and how should a face recognition system acquire sufficient training illumination samples to achieve high accuracy on a wide variety of practical, real-world illumination conditions?*

First, let us consider an approximation of the human face as a convex, Lambertian object under distinct illuminations with a fixed pose. Under those assumptions, the incident and reflected light are distributions on a sphere, and thus can be represented in a spherical harmonic basis [1]. The Lambertian reflectance kernel acts as a low-pass filter between the incident and reflected light, and as a result, the set of images of the object end up lying very close to a subspace corresponding to the low-frequency spherical harmonics. In fact, one can show that only nine (properly chosen) basis illuminations are sufficient to generate basis images that span all possible images of the object.

While modeling the harmonic basis is important for understanding the image formation process, various empirical studies have shown that even in the case when convex, Lambertian assumptions are violated, the algorithm can still get away with using a small number of frontal illuminations to linearly represent a wide range of new frontal illuminations, especially when they are all taken under the same laboratory conditions. This is the case for many public face databases, such as AR, ORL, PIE, and Multi-PIE. Unfortunately, in practice, we have observed that a training database consisting purely of frontal illuminations is not sufficient to linearly represent images of a face taken under typical indoor and outdoor conditions. To ensure our algorithm works in practice, we need to more carefully acquire a set of training illuminations that are sufficient to linearly represent a wide variety of practical indoor and outdoor illuminations.

To this end, we have designed a system that can acquire frontal images of a subject while simultaneously illuminating the subject from all directions. A sketch of the system is shown in Figure 1.9. A more detailed explanation of this system is discussed in [44].

Based on the results of our experiments, the illumination patterns projected either directly on the subject's frontal face or indirectly on the wall correspond to a total of 38 training illumination images, as an example shown in Figure 1.10. We have observed

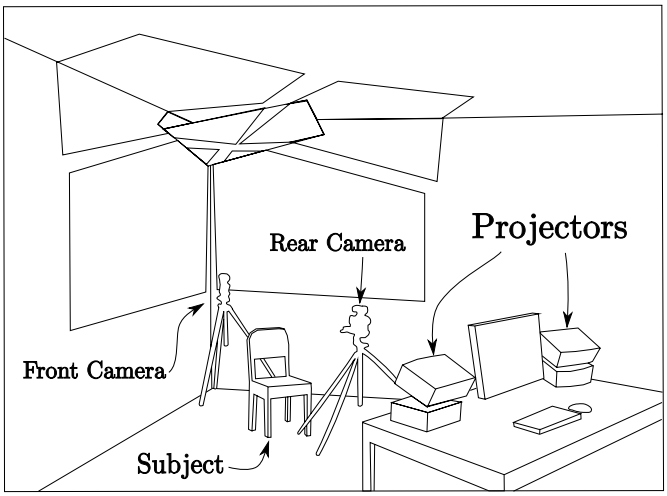

Figure 1.9 Illustration of the training acquisition system, which consists of four projectors and two cameras controlled by a computer.

Figure 1.10 38 training images of a subject collected by the system. The first 24 images are sampled using the foreground lighting patterns, and the rest 14 images using the background lighting patterns.

that further acquiring finer illumination patterns does not significantly improve the image registration and recognition accuracy [44]. Therefore, we have used those illumination models for all our large-scale experiments.

1.8 Overall System Evaluation

In this section, we present representative recognition results of our complete system on large-scale face databases. All the experiments are carried out using input directly obtained from the Viola and Jones' face detector, without any manual intervention throughout the process.

We use two different face databases to test our system. We first report the performance of our system on the largest public face database available that is suitable for testing our algorithm, the CMU Multi-PIE database [23]. This database contains images of 337 subjects across simultaneously variation in pose, expression, illumination and facial appearance over time, thus provides the most extensive test among all public databases. However, one shortcoming of the CMU Multi-PIE database for our purpose is that all the

Table 1.1. Recognition rates on CMU Multi-PIE database.

Rec. Rates	Session 2	Session 3	Session 4
LDA_d (LDA_m)	5.1 (49.4)%	5.9 (44.3)%	4.3 (47.9)%
NN_d (NN_m)	26.4 (67.3)%	24.7 (66.2)%	21.9 (62.8)%
NS_d (NS_m)	30.8 (77.6)%	29.4 (74.3)%	24.6 (73.4)%
Algorithm 1.1	**91.4** %	**90.3** %	**90.2** %

images are taken under controlled laboratory lighting conditions, restricting our choice of training and testing sets to these conditions, which may not cover all typical natural illuminations. Therefore, our goal of this experiment is to simply demonstrate the effectiveness of our fully automatic system with respect to such a large number of classes. We next test on a face database collected using our own acquisition system as described in Section 1.7. The goal of that experiment is then to show that with a sufficient set of training illuminations, our system is indeed capable of performing robust face recognition with loosely controlled test images taken under practical indoor and outdoor conditions.

For the CMU Multi-PIE database, we use all the 249 subjects present in Session 1 as the training set. The remaining 88 subjects are considered as "outliers" and are used to test our system's ability to reject invalid images. To further challenge our system, we include only 7 extreme frontal illumination for each of the 249 subjects in the training, and use frontal images of all the 20 illuminations from Session 2-4 as testing, which were recorded at different times over a period of several months. Table 1.1 shows the result of our algorithm on each of the three testing sessions, as well as the results obtained using baseline linear-projection-based algorithms including Nearest Neighbor (NN), Nearest Subspace (NS)[30] and Linear Discriminant Analysis (LDA)[3]. Note that we initialize these baseline algorithms in two different ways, namely, from the output of the Viola and Jones' detector, indicated by a subscript "d", and with images which are aligned to the training with manually clicked outer eye-corners, indicated by a subscript "m". One can see in Table 1.1 that, despite careful manual registration, these baseline algorithms perform significantly worse than our system, which uses input directly from the face detector.

We further perform subject validation on Multi-PIE database, using the measure of concentration of the sparse coefficients as introduced in Section 1.2, and compare this method to the classifiers based on thresholding the error residuals of NN, NS and LDA. Figure 1.11 plots the receiver operating characteristic (ROC) curves, which are generated by sweeping the threshold through the entire range of possible values for each algorithm. We can see that our approach again significantly outperforms the other three algorithms.

For experiments on our own database, we have collected the frontal view of 74 subjects without eyeglasses under 38 illuminations as shown in Section 1.7 and use them as the training set. For testing our algorithm, we have also taken 593 images of these subjects with a different camera under a variety of indoor and outdoor conditions. Based on the main variability in the test images, we further partitioned the testing set into five categories:

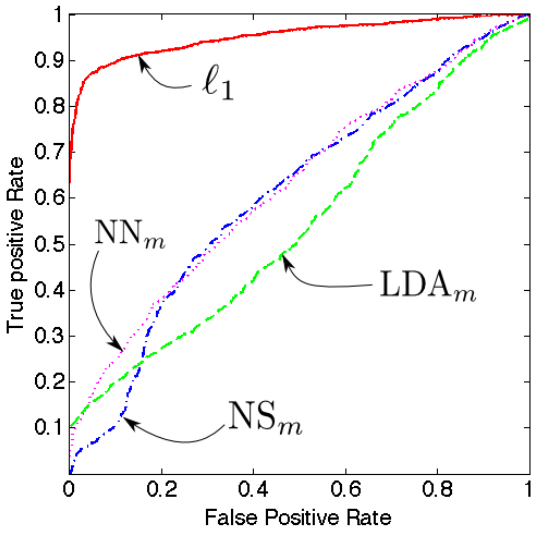

Figure 1.11 ROC curves for our algorithm (labeled as "ℓ_1"), compared with those for NN_m, NS_m, and LDA_m.

Table 1.2. Recognition rates on our own database.

Test Categories	C1	C2	C3	C4	C5
Rec. Rates (%)	95.9	91.5	63.2	73.7	53.5

C1: 242 images of 47 subjects without eyeglasses, generally frontal view, under a variety of practical illuminations (indoor and outdoor) (Figure 1.12, row 1).
C2: 109 images of 23 subjects with eyeglasses (Figure 1.12, row 2).
C3: 19 images of 14 subjects with sunglasses (Figure 1.12, row 3).
C4: 100 images of 40 subjects with noticeable expressions, poses, mild blur, and sometimes occlusion (Figure 1.13, both rows).
C5: 123 images of 17 subjects with little control (out of focus, motion blur, significant pose, large occlusion, funny faces, extreme expressions) (Figure 1.14, both rows).

Table 1.2 reports the recognition rates of our system on each category. As one can see, our system achieves recognition rates above 90% for face images with general frontal views, under a variety of practical illuminations. Our algorithm is also robust to small amounts of pose, expression and occlusion (i.e., eyeglasses).

1.9 Conclusion and Discussion

Based on the theory of sparse representation, we have proposed a comprehensive framework/system to tackle the classical problem of face recognition in computer vision. The

Figure 1.12 Representative examples of categories 1-3. One row for each category.

Figure 1.13 Representative examples of category 4. Top row: successful examples. Bottom row: failures.

Figure 1.14 Representative examples of category 5. Top row: successful examples. Bottom row: failures.

initial success of our solution relies on careful analysis of the special data structures in high-dimensional face images. Although our study has revealed new insights about face recognition, many new problems remain largely open. For instance, it is still not clear why the sparse representation based classification (SRC) is so discriminative for highly correlated face images. Indeed, since the matrix $\Phi = [\Phi_1, \Phi_2, \cdots, \Phi_C]$ has class structure, one simple alternative to SRC is to treat each class one at a time, solving a robust regression problem via the ℓ^1 norm, and then select the class with the lowest regression error. Similar to SRC, this alternative respects the physics of illumination and in-plane transformations, and leverages the ability of ℓ_1-minimization to correct sparse errors. However, we find that SRC has a consistent advantage in terms of classification percentage (about 5% on Multi-PIE [44]). One more sophisticated way to take advantage of class structure is by enforcing group sparsity on the coefficients $\boldsymbol{c}$. While this may impair the system's ability to reject invalid subjects (as in Figure 1.11), it also has the potential to improve recognition performance [39, 32].

Together with other papers that appeared in the similar time frame, this work has inspired researchers to look into a broader range of recognition problems within the framework of sparse representation. Notable examples include image super-resolution [50], object recognition [34, 41], human activity recognition [49], speech recognition

[20], 3-D motion segmentation [37, 19], and compressed learning [7]. While these promising works raise many intriguing questions, we believe the full potential of sparse representation for recognition problems remains to be better understood mathematically and carefully evaluated in practice.

References

[1] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[3] P. Belhumeur, J. Hespanda, and D. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[4] P. Belhumeur and D. Kriegman. What is the set of images of an object under all possible illumination conditions? *International Journal on Computer Vision*, 28(3):245–260, 1998.

[5] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2003.

[6] A. Bruckstein, D. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, 2009.

[7] R. Calderbank, S. Jafarpour, and R. Schapire. Compressed learning: universal sparse dimensionality reduction and learning in the measurement domain. *preprint*, 2009.

[8] E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12), 2005.

[9] E. Candès and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[10] H. Chen, H. Chang, and T. Liu. Local discriminant embedding and its variants. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[11] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[12] T. Chen, W. Yin, X. Zhou, D. Comaniciu, and T. Huang. Total variation models for variable lighting face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1519–1524, 2006.

[13] D. Donoho. Neighborly polytopes and sparse solution of underdetermined linear equations. *preprint*, 2005.

[14] D. Donoho. For most large underdetermined systems of linear equations the minimal ℓ^1-norm near solution approximates the sparest solution. *Communications on*

Pure and Applied Mathematics, 59(6):797–829, 2006.

[15] D. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[16] D. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences*, 102(27):9452–9457, 2005.

[17] D. Donoho and J. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22(1):1–53, 2009.

[18] M. Elad, J. Starck, P. Querre, and D. Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Applied and Computational Harmonic Analysis*, 19:340–358, 2005.

[19] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

[20] J. Gemmeke, H. Van Hamme, B. Cranen, and L. Boves. Compressive sensing for missing data imputation in noise robust speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):272–287, 2010.

[21] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[22] R. Glowinski and A. Marrocco. Sur l'approximation par él'ements finis d'ordre un, et la r'esolution, par p'enalisation-dualit'e d'une classe de probl'emes de dirichlet nonlin'eaires. *Revue Francaise d'Automatique, Informatique, Recherche Opérationnelle*, 9(2):41–76, 1975.

[23] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. In *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*, 2008.

[24] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, 1998.

[25] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[26] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[27] K. Jittorntrum and M. Osborne. Strong uniqueness and second order convergence in nonlinear discrete approximation. *Numerische Mathematik*, 34:439–455, 1980.

[28] N. Karmarkar. A new polynomial time algorithm for linear programming. *Combinatorica*, 4:373–395, 1984.

[29] T. Kim and J. Kittler. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):318–327, 2005.

[30] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005.

[31] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of International Joint Conference on Artificial Intelligence*, volume 3, pages 674–679, 1981.

[32] A. Majumdar and R. Ward. Improved group sparse classifier. *Pattern Recognition Letters*, 31:1959–1964, 2010.

[33] A. Martinez and R. Benavente. The AR face database. Technical report, CVC Technical Report No. 24, 1998.

[34] N. Naikal, A. Yang, and S. Sastry. Towards an efficient distributed object recognition system in wireless smart camera networks. In *Proceedings of the International Conference on Information Fusion*, 2010.

[35] M. Osborne and R. Womersley. Strong uniqueness in seequential linear programming. *Journal of the Australian Mathematical Society, Series B*, 31:379–384, 1990.

[36] L. Qiao, S. Chen, and X. Tan. Sparsity preserving projections with applications to face recognition. *Pattern Recognition*, 43(1):331–341, 2010.

[37] S. Rao, R. Tron, and R. Vidal. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1832–1845, 2010.

[38] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006.

[39] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar. C-HiLasso: A collaborative hierarchical sparse modeling framework. *(To appear) IEEE Transactions on Signal Processing*, 2011.

[40] J. Tropp and S. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98:948–958, 2010.

[41] G. Tsagkatakis and A. Savakis. A framework for object class recognition with no visual examples. In *Western New York Image Processing Workshop*, 2010.

[42] M. Turk and A. Pentland. Eigenfaces for recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 1991.

[43] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[44] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma. Toward a practical automatic face recognition system: Robust pose and illumination via sparse representation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

[45] J. Wright and Y. Ma. Dense error correction via ℓ^1-minimization. *IEEE Transactions on Information Theory*, 56(7):3540–3560, 2010.

[46] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210 – 227, 2009.

[47] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extension: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:40–51, 2007.

[48] A. Yang, A. Ganesh, Z. Zhou, S. Sastry, and Y. Ma. Fast ℓ_1-minimization algorithms for robust face recognition. *(preprint) arXiv:1007.3753*, 2011.

[49] A. Yang, R. Jafari, S. Sastry, and R. Bajcsy. Distributed recognition of human actions using wearable motion sensor networks. *Journal of Ambient Intelligence and Smart Environments*, 1(2):103–115, 2009.

[50] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution as sparse representation of raw image patches. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[51] J. Yang and Y. Zhang. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *arXiv:0912.1185*, 2009.

[52] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, pages 399–458, 2003.

[53] S. Zhou, G. Aggarwal, R. Chellappa, and D. Jacobs. Appearance characterization of linear lambertian objects, generalized photometric stereo, and illumination-invariant face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 230–245, 2007.

Index