Int J Comput Vis
DOI 10.1007/s11263-014-0749-x

Sparse Illumination Learning and Transfer for Single-Sample Face Recognition with Image Corruption and Misalignment

Liansheng Zhuang · Tsung-Han Chan · Allen Y. Yang · S. Shankar Sastry · Yi Ma

Received: 7 February 2014 / Accepted: 2 July 2014

Abstract Single-sample face recognition is one of the most challenging problems in face recognition. We propose a novel algorithm to address this problem based on a sparse representation based classification (SRC) framework. The new algorithm is robust to image misalignment and pixel corruption, and is able to reduce required gallery images to one sample per class. To compensate for the missing illumination information traditionally provided by multiple gallery images, a sparse illumination learning and transfer (SILT) technique is introduced. The illumination in SILT is learned by fitting illumination examples of auxiliary face images from one or more additional subjects with a sparsely-used illumination dictionary. By enforcing a sparse representation of the query image in the illumination dictionary, the SILT can effectively recover and transfer the illumination and pose information from the alignment stage to the recognition stage. Our extensive experiments have demonstrated that the new algorithms significantly outperform the state of the art in the single-sample regime and with less restrictions. In particular, the single-sample face alignment accuracy is comparable to that of the well-known Deformable SRC algorithm using multiple gallery images per class. Furthermore, the face recognition accuracy exceeds those of the SRC and Extended SRC algorithms using hand labeled alignment initialization.

Keywords Single-sample face recognition · Illumination dictionary learning · Sparse illumination transfer · Face alignment · Robust face recognition

Communicated by Julien Mairal, Francis Bach, and Michael Elad.

A preliminary version of the results was published in Zhuang et al. (2013).

L. Zhuang
CAS Key Laboratory of Electromagnetic Space Information, University of Science and Technology of China, Heifei, China
e-mail: lszhuang@ustc.edu.cn

T.-H. Chan
Advanced Digital Sciences Center, Singapore, Singapore
e-mail: thchan@ieee.org

A. Y. Yang (✉)· S. S. Sastry
Department of EECS, University of California, Berkeley, CA 94720, USA
e-mail: yang@eecs.berkeley.edu

S. S. Sastry
e-mail: sastry@eecs.berkeley.edu

Y. Ma
ShanghaiTech University, Shanghai, China
e-mail: mayi@shanghaitech.edu.cn

1 Introduction

Face recognition is one of the classical problems in computer vision. Given a natural image that may contain a human face, it has been known that the appearance of the face image can be easily affected by many image nuisances, including background illumination, pose, and facial corruption/disguise such as makeup, beard, and glasses. Therefore, to develop a *robust* face recognition system whose performance can be comparable to or even exceed that of human vision, the computer system needs to address at least the following three closely related problems: First, it needs to effectively model the change of illumination on the human face. Second, it needs to align the pose of the face. Third, it needs to tolerance the corruption of facial features that leads to potential gross pixel error against the gallery images.

In the literature, many well-known solutions have been studied to tackle these problems (Hager and Belhumeur

Published online: 23 July 2014

1998; Zhao et al. 2003; Ho et al. 2003; Ganesh et al. 2011), although a complete review of the field is outside the scope of this paper. More recently, a new face recognition framework called *sparse-representation based classification* (SRC) was proposed (Wright et al. 2009), which can successfully address most of the above problems. The framework is built on a subspace illumination model characterizing the distribution of a corruption-free face image sample (stacked in vector form) under a fixed pose, one subspace model per subject class (Belhumeur et al. 1997; Basri and Jacobs 2003). When an unknown query image is jointly represented by all the subspace models, only a small subset of these subspace coefficients need to be nonzero, which would primarily correspond to the subspace model of the true subject. Therefore, by optimizing the sparsity of such an overcomplete linear representation, the dominant nonzero coefficients indicate the identity of the query image. In the case of image corruption, since the corruption typically only affects a sparse set of pixel values, one can concurrently optimize a sparse error term in the image space to compensate for the corrupted pixel values.

In practice, a face image may appear at any image location with random background. Therefore, a face detection and registration step is typically first used to detect the face image. Most of the methods in face detection would learn a class of local image features/patches that are sensitive to the appearance of key facial features (Yan et al. 2003; Viola and Jones 2004; Liang et al. 2008). Using either an active shape model (Cootes et al. 1995) or an active appearance model (Cootes et al. 1998), the location of the face can be detected even when the expression of the face is not neutral or some facial features are occluded (Saragih et al. 2009; Gu and Kanade 2008). However, using these face registration algorithms *alone* is not sufficient to align a query image to gallery images in SRC. The main reasons are two-fold: First, except for some fast detectors such as Viola-Jones (Viola and Jones 2004), more sophisticated detectors are expensive to run and require learning prior distribution of the shape model from meticulously hand-labeled gallery images. More importantly, these detectors would register the pixel values of the query image with respect to the *average* shape model learned from all the gallery images, but they typically cannot align the pixel values of the query image to the gallery images for the purpose of recognition, as required in SRC.

Following the sparse representation framework in Wright et al. (2009), Wagner et al. (2012), we propose a novel algorithm to effectively extend SRC for face alignment and recognition in the small-sample-set scenario. We observe that in addition to the aforementioned image nuisances, one of the outstanding challenges in face recognition is indeed the small sample set problem. For instance, in many biometric, surveillance, and Internet applications, there may be only a few gallery examples that are collected for a subject of interest, and the subject may not be able to undergo a comprehensive image collection session in a laboratory.[1]

Unfortunately, most of the existing SRC-based alignment and recognition algorithms would fail in such scenarios. For starters, the original SRC algorithm (Wright et al. 2009) assumes a plurality of gallery samples from each class must sufficiently span its illumination subspace. The algorithm performs poorly in the single sample regime, as we will later shown in our experiment. In Wagner et al. (2012), in order to guarantee that the gallery images contain sufficient illumination patterns, the test subjects must further go through a non-trivial passport-style image collection process in a dark room in order to be entered into the gallery database. More recently, another development in the SRC framework is simultaneous face alignment and recognition methods (Yan et al. 2010; Huang et al. 2008; Yang et al. 2012). Nevertheless, these methods did not go beyond the basic assumption used in SRC and other prior art that the face illumination model is measured by multiple gallery samples for each class. Furthermore, as shown in Wagner et al. (2012), robust face alignment and recognition can be solved separately as a two-step process, as long as the recovered image transformation can be carried over from the alignment stage to the recognition stage. Therefore, simultaneous face alignment and recognition could make the already expensive sparse optimization problem even more difficult to solve.

1.1 Contributions

Single-sample face alignment and recognition represents an important step towards practical face recognition solutions using images collected in the wild or on the Internet. We contend that the problem can be solved quite effectively by an elegant algorithm. The key observation is that one sample per class mainly deprives the algorithm of an illumination subspace model for individual classes. We show that an *illumination dictionary* can be learned from additional subject classes to compensate for the lack of the illumination information in the gallery set.

Due to the fact that the variations of human faces are usually smaller than illumination changes of the same face, we propose a dictionary learning method to decompose the face images as vectors into two components: a low-rank matrix encodes the subject identities while a sparsely-used matrix (or dictionary) represents the possible illumination variations. The auxiliary illumination images can be selected outside the set of gallery subjects. Since most of the informa-

[1] In this paper, we use Viola-Jones face detector to initialize the face image location. As a result, we do not consider scenarios where the face may contain a large 3D transformation or large expression change. These more severe conditions can be addressed in the face detection stage using more sophisticated face models as we previously mentioned.

tion associated with the subject identities is contained in the rank-constrained matrix, the sparsely-used illumination dictionary is expected to be subject-invariant. Finally, we show that the other image nuisances, including pose variation and image corruption, can be readily corrected by a single gallery image of *arbitrary illumination condition* combined with the illumination dictionary. The algorithm also does not need to know the information of any possible facial corruption for the algorithm to be robust. The new method is called *sparse illumination learning and transfer* (SILT). Similarly, the illumination dictionary defined in the method will be referred to as the SILT dictionary.

Preliminary results of this work were first reported in our conference paper (Zhuang et al. 2013). To the best of our knowledge, the paper (Zhuang et al. 2013) was the first to propose a solution to perform small-sample-set facial alignment and recognition via a sparse illumination transfer. However, the construction of the illumination dictionary in Zhuang et al. (2013) was largely ad hoc via a simple concatenation of the auxiliary illumination samples. It was suggested in Zhuang et al. (2013) that a sparse illumination representation can be found to compensate for the missing illumination model in single gallery images. In this paper, we propose a new illumination dictionary model to specifically learn the dictionary from the auxiliary images. We also study efficient optimization algorithms to solve the dictionary learning problem numerically. Finally, more comprehensive experiments are conducted, especially on the case when the number of available illumination learning subjects grows from one to many. In the largest scale, we employ all the 38 available subjects in the Extended YaleB database (Lee et al. 2005) as the auxiliary illumination samples. The new results show improved recognition results than those in Zhuang et al. (2013).

In terms of the algorithm complexity, learning the SILT dictionary contains two successive procedures; one is principal component analysis (PCA)-like solution while the other involves solving a sequence of linear programs. The learning algorithm is almost parameter-free, only dependent on the dictionary size. Applying the SILT dictionary in the alignment and recognition stages potentially can significantly improve the speed of SRC-type algorithms, because a sparse optimization solver such as those in Yang et al. (2013) is now faced with much smaller linear systems that only involves a single sample per class plus a small learned illumination dictionary.

This paper bears resemblance to the work called Extended SRC (Deng et al. 2012), whereby an intraclass variant dictionary was similarly added to be a part of the SRC objective function for recognition. Our work differs from Deng et al. (2012) in that the proposed SILT dictionary is automatically learned from a selection of independent subject(s), whereas in Deng et al. (2012), the dictionary is simply hand-crafted. Yet, the subject classes used to learn the SILT dictionary is also impartial to the gallery classes. Furthermore, by transferring both the pose and illumination from the alignment stage to the recognition stage, our algorithm can handle insufficient illumination and misalignment at the same time, and allows for the single reference images to have arbitrary illumination conditions. Finally, our algorithm is also robust to moderate amounts of image pixel corruption, even though we do not need to include any image corruption examples in the SILT dictionary, while in Deng et al. (2012) the intraclass variant dictionary uses both normal and corrupted face samples. We also compare our performance with Deng et al. (2012) in Sect. 5.

More recently, the problem of single-sample face recognition was considered in another work (Yang et al. 2013), called *sparse variation dictionary learning* (SVDL). The work proposed an alternative method to learn a sparse variation dictionary that amends the SRC framework with single samples. The main difference between the two dictionary learning algorithms is that in SVDL, both the illumination learning images and the gallery images are involved in the dictionary learning algorithm. The authors argued that jointly considering the illumination samples and the gallery samples helps to generate a very compact, adaptive dictionary that exploits the correlation between the illumination learning set and the gallery set. While in this paper, the learning of the SILT dictionary is independent of the gallery set and the alignment and recognition tasks. Therefore, the learned dictionary can be estimated off-line and without costing any computational penalty when the gallery images are presented. Furthermore, the SILT framework addresses both face alignment and recognition problems, and is capable of transferring both the illumination *and* pose information from the alignment stage to the recognition stage. In contrast, SVDL in Yang et al. (2013) only concerns face recognition with a frontal position, and its complexity would grow substantially when its adaptive dictionary needs to be re-computed under varying poses of the query image. We will show in Sect. 5 that, without considering this pose-related computational penalty for SVDL, the SILT framework outperforms SVDL in both recognition accuracy and robustness to pixel corruption.

2 Sparse Representation-based Classification

In this section, we first briefly review the SRC framework. Assume a face image $\boldsymbol{b} \in \mathbb{R}^d$ in grayscale can be written in vector form by stacking its pixels. Given L subject classes, assume n_i well-aligned gallery images $A_i = [\boldsymbol{a}_{i,1}, \boldsymbol{a}_{i,2}, \ldots, \boldsymbol{a}_{i,n_i}] \in \mathbb{R}^{d \times n_i}$ of the same dimension as $\boldsymbol{b}$ are sampled for the i-th class under the frontal position and various illumination conditions. These gallery images are further aligned in terms of the coordinates of some salient facial features, e.g., eye corners and/or mouth corners. For brevity,

the gallery images under such conditions are said to be in the *neutral position*. Furthermore, we do not explicitly model the variation of facial expression in this paper. Based on the illumination subspace assumption, if $\boldsymbol{b}$ belongs to the i-th class, then $\boldsymbol{b}$ lies in the low-dimensional subspace spanned by the gallery images in A_i, namely,

$$\boldsymbol{b} = A_i \boldsymbol{x}_i. \tag{1}$$

When the query image $\boldsymbol{b}$ is captured in practice, it may contain an unknown 3D pose that is different from the neutral position. In image registration literature (Lucas and Kanade 1981; Hager and Belhumeur 1998; Wagner et al. 2012), the effect of the 3D pose can be modeled as an image transformation $\tau \in T$, where T is a finite-dimensional group of transformations, such as translation, similarity transform, affine transform, and homography. The goal of face alignment is to recover the transformation τ, such that the unwarped query image $\boldsymbol{b}_0$ in the neutral position remains in the same illumination subspace: $\boldsymbol{b}_0 \doteq \boldsymbol{b} \circ \tau = A_i \boldsymbol{x}_i$.

In robust face alignment, the issue is often further exacerbated by the cascade of complex illumination patterns and moderate image pixel corruption and occlusion. In the SRC framework (Wright et al. 2009; Wagner et al. 2012), the combined effect of image misalignment and sparse corruption is modeled by

$$\hat{\tau}_i = \arg \min_{\boldsymbol{x}_i, \boldsymbol{e}, \tau_i} \|\boldsymbol{e}\|_1 \quad \text{subj. to} \quad \boldsymbol{b} \circ \tau_i = A_i \boldsymbol{x}_i + \boldsymbol{e}, \tag{2}$$

where the alignment is achieved on a per-class basis for each A_i, and $\boldsymbol{e} \in \mathbb{R}^d$ is the sparse alignment error. After linearizing the potentially nonlinear image transformation function τ, (2) can be solved iteratively by a standard ℓ_1-minimization solver. In Wagner et al. (2012), it was shown that the alignment based on (2) can tolerate translation shift up to 20% of the between-eye distance and up to 30° in-plane rotation, which is typically sufficient to compensate moderate misalignment caused by a good face detector.

Once the optimal transformation τ_i is recovered for each class i, the transformation is carried over to the recognition algorithm, where the gallery images in each A_i are transformed by τ_i^{-1} to align with the query image $\boldsymbol{b}$. Finally, a global sparse representation $\boldsymbol{x}$ with respect to the transformed gallery images is sought by solving the following sparse optimization problem:

$$\begin{aligned} \boldsymbol{x}^* &= \arg\min_{\boldsymbol{x}, \boldsymbol{e}} \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \\ \text{subj. to} \quad \boldsymbol{b} &= \left[A_1 \circ \tau_1^{-1}, \ldots, A_L \circ \tau_L^{-1}\right] \boldsymbol{x} + \boldsymbol{e}. \end{aligned} \tag{3}$$

One can further show that when the correlation of the face samples in A is sufficiently tight in the high-dimensional image space, solving (3) via ℓ_1-minimization guarantees to recover both the sparse coefficients $\boldsymbol{x}$ and very dense randomly signed error $\boldsymbol{e}$ (Wright and Ma 2010).

3 Sparse Illumination Learning and Transfer

In this section, we propose a novel face alignment algorithm that is effective even when a very small number of training images are provided per class, called *sparse illumination learning and transfer* (SILT). In the extreme case, we specifically consider the *single-sample face alignment problem* where only one training image $\boldsymbol{a}_i$ of *arbitrary illumination* is available from class i. The same algorithm easily extends to the case when multiple training images are provided. In Sect. 4.2, we will show how to integrate the estimation of SILT in robust single-sample face recognition. In Sect. 5, we further show in our experiment that SILT is also complementary and useful in other existing face recognition methods as an image pre-processing step.

3.1 Illumination Dictionary Learning

To mitigate the scarcity of the training images, something has to give to recover the missing illumination model under which the image appearance of a human face can be affected. Motivated by the idea of transfer learning (Do and Ng 2005; Quattoni et al. 2008; Lampert et al. 2009), we stipulate that one can obtain the illumination information for both alignment *and* recognition from a set of additional subject classes, called the *illumination dictionary*. The auxiliary face images for learning the illumination dictionary have the same frontal pose as the gallery images, and can be collected offline and different from the query classes $A = [A_1, \ldots, A_L]$. In other words, no matter how scarce the gallery images are, one can always obtain a potentially large set of auxiliary face images from other unrelated subjects who may have similar face shapes as the query subjects and may provide sufficient illumination examples.

Suppose that we are given face images of sufficient illumination patterns for additional p subjects $D = [D_1, \ldots, D_p] \in \mathbb{R}^{d \times (np)}$, and assume without loss of generality that each subject contains n face images, i.e., $D_i \in \mathbb{R}^{d \times n}$ for subject i, and each image has the same dimension as the gallery images.

Our hope is that D can be expressed by a superposition of a rank-constrained matrix and a sparsely-used matrix:

$$D = V \otimes \mathbf{1}^T + CS, \tag{4}$$

where $V \in \mathbb{R}^{d \times p}$ is a matrix where each column vector represents a subject class from 1 to p, $\mathbf{1} \in \mathbb{R}^n$, $C \in \mathbb{R}^{d \times k}$ is a learned illumination dictionary, and $S \in \mathbb{R}^{k \times np}$ is a sparse matrix. Here, $\otimes$ denotes the Kronecker product, and hence

the first term $V \otimes \mathbf{1}^T \in \mathbb{R}^{d\times(np)}$ in (4) is clearly low rank. We also assume that $k \leq \min\{d, np\}$ for C to prevent model over-fitting.

One can better understand the roles of the different matrices in (4) as follows: $V\otimes\mathbf{1}^T$ describes the inter-class variation associated with the p different subject identifies, C describes the common intra-class variation associated with the illumination change, and S operates like a sparse representation of illumination patterns that compensate the singular subject images in V. Considering other possible face variations, we may further add a small error term $E \in \mathbb{R}^{d\times np}$ in (4) as

$$D = V \otimes \mathbf{1}^T + CS + E. \tag{5}$$

To encourage sparsity of S and minimum fitting error E, we formulate the illumination dictionary learning problem as an optimization problem

$$\min_{V,C,S,E} \|S\|_0 + \|E\|_F \text{ subj. to } D = V \otimes \mathbf{1}^T + CS + E, \tag{6}$$

where $\|\cdot\|_0$ denotes the matrix ℓ_0-norm and $\|\cdot\|_F$ is the Frobenius norm. Note that in the SRC framework such as (2) and (3), the image corruption has been traditionally estimated by minimizing a sparse error term $\|E\|_0$. The reason we can model a dense error term using $\|E\|_F$ is that the selection of the auxiliary illumination examples is conducted manually and offline. Therefore, it is reasonable to assume that the face images in D do not contain significant facial disguise and pixel corruption. This assumption also simplifies the complexity of the optimization problem in (6).

3.2 Numerical Implementation

Solving (6) is a challenging problem, mainly because it has a non-convex objective function *and* a non-convex, non-linear constraint. In optimization, the standard procedure to relax the non-convex objective function is to find a good convex surrogate. However, the second problem about how to handle the non-convex, non-linear constraint is less understood. Although the well-known alternating direction method (Gabay and Mercier 1976; Tseng 1991; Boyd et al. 2011) can be applied, the solution may not converge to the global optimum.

In the following, we will reformulate the constraint in (6) and propose a successive optimization algorithm. The algorithm can be shown numerically to recover V, C, S, E exactly if S is sufficiently sparse.

First, we reformulate the constraint of (6) as follows:

$$D = \underbrace{(V - CF)}_{\overline{V}} \otimes \mathbf{1}^T + \underbrace{CW}_{\overline{C}} \underbrace{W^{-1}(S + F \otimes \mathbf{1}^T)}_{H} + E, \tag{7}$$

where $F \in \mathbb{R}^{k\times p}$ measures the possible ambiguity between the first two terms of the right hand side, and $W \in \mathbb{R}^{k\times k}$ is a non-singular transformation such that $\overline{C}^T\overline{C} = I$, where I is the identity matrix of proper dimension.

From (7), we have

$$S = WH - F \otimes \mathbf{1}^T. \tag{8}$$

Hence, problem (6) can be written as:

$$\min_{\substack{\text{rank}(W)=k\\ F}} \left[\|WH - F \otimes \mathbf{1}^T\|_0 + \left(\min_{\substack{D=\overline{V}\otimes\mathbf{1}^T+\overline{C}H+E\\ \overline{C}^T\overline{C}=I}} \|E\|_F \right) \right]. \tag{9}$$

The new formulation in (9) allows us to apply a successive optimization strategy. In this case, successive optimization exploits the successive structure of (9) to recursively approximate problem (6). Although it is a heuristic, but it can have promising performance in practice.

More specifically, we approximate problem (9) by decoupling it into two successively processed stages:

$$\{\overline{V}^*, \overline{C}^*, H^*, E^*\} = \arg\min \|E\|_F \quad \text{subj. to } D = \overline{V} \otimes \mathbf{1}^T + \overline{C}H + E,\ \overline{C}^T\overline{C} = I, \tag{10}$$

$$\{W^*, F^*\} = \arg\min \|WH^* - F \otimes \mathbf{1}^T\|_0 \quad \text{subj. to } \text{rank}(W) = k. \tag{11}$$

Suppose that $\{\overline{V}^*, \overline{C}^*, H^*, E^*, W^*, F^*\}$ are found, then the solutions of the other variables in problem (6) are given by

$$C^* = \overline{C}^*(W^*)^{-1}, \tag{12a}$$

$$V^* = \overline{V}^* + C^*F^*, \tag{12b}$$

$$S^* = W^*H^* - F^* \otimes \mathbf{1}^T. \tag{12c}$$

In what follows, we describe how to solve problem (10) and (11).

3.2.1 Solving Problem (10)

Problem (10) is a difficult non-convex problem. Fortunately we can prove that it has a closed-form solution, as stated in the following theorem:

Theorem 1 *Suppose that $D = [D_1, D_2, ..., D_p]$ where D_i is the training set associated with subject i, and assume each subject has n images. Problem (10) has the following closed-form solution:*

$$\begin{aligned}
\overline{V}^* &= [\ \tfrac{1}{n} D_1 \mathbf{1}, \tfrac{1}{n} D_2 \mathbf{1}, \ldots, \tfrac{1}{n} D_p \mathbf{1}\ ], \\
\overline{C}^* &= [\ \boldsymbol{q}_1(UU^T), \boldsymbol{q}_2(UU^T), \ldots, \boldsymbol{q}_k(UU^T)\ ], \\
H^* &= (\overline{C}^*)^T (D - \overline{V}^* \otimes \mathbf{1}^T), \\
E^* &= D - \overline{V}^* \otimes \mathbf{1}^T - \overline{C}^* H^*.
\end{aligned} \tag{13}$$

where $U = D - \overline{V}^* \otimes \mathbf{1}^T$ *and* $\boldsymbol{q}_i(Z)$ *is the eigenvector associated with the* i*th principal eigenvalue of the square matrix* Z.

The proof of Theorem 1 is given in Appendix. To better understand the closed-form solution, we can see that each column of $\overline{V}^*$ represents the mean vector of a training set D_i. Therefore, U represents a normalized data matrix when the mean vectors are removed from D. Since the column vectors of $\overline{C}$ are the first k orthonomal basis vectors that maximizes the inter-class variance, it can be thought of as a variant of principal component analysis (PCA).

3.2.2 Solving Problem (11)

We now turn our attention to problem (11), which is more difficult than (10). The problem is very similar to a conventional sparse dictionary learning problem, where the goal is to learn a basis that most compactly represents the face images D. While many heuristics have been proposed before (*e.g.*, see Aharon et al. (2006) and the references therein), because of its combinatorial nature, this problem is difficult to solve efficiently.

Our solution of (11) is largely inspired by a recent paper (Spielman et al. 2012), which shows that the inverse problem can be well-defined, and there exist efficient and provably correct algorithms to solve the inverse problem. The only difference lies in that our problem has an additional unknown matrix F here. Hence, we propose to solve problem (11) by solving the following linear programs sequentially; that is, for i from 1 to k, we solve

$$\begin{aligned}
\{\hat{\boldsymbol{w}}_i, \hat{\boldsymbol{f}}_i\} = \arg \min_{\boldsymbol{w} \in \mathbb{R}^k, \boldsymbol{f} \in \mathbb{R}^p} & \|\boldsymbol{w}^T H^* - \boldsymbol{f}^T \otimes \mathbf{1}^T\|_1, \\
\text{subj. to } & \boldsymbol{w}^T P^{\perp}_{\hat{W}_{i-1}} \mathbf{r} = 1,
\end{aligned} \tag{14}$$

where $\hat{\boldsymbol{w}}_i^T$ and $\hat{\boldsymbol{f}}_i^T$ denote the estimates of the ith row vector of W and F, respectively, $\hat{W}_{i-1} = [\hat{\boldsymbol{w}}_1, ..., \hat{\boldsymbol{w}}_{i-1}] \in \mathbb{R}^{k \times (i-1)}$ denotes a matrix comprising previously found solutions, $P^{\perp}_{\hat{W}_{i-1}}$ is the orthogonal complement projector of $\hat{W}_{i-1}$, and $\mathbf{r} \in \mathbb{R}^k$ is an analysis filter. Note that the constraint in (14) is to ensure $P^{\perp}_{\hat{W}_{i-1}} \boldsymbol{w} \neq \mathbf{0},\ \forall\, i$, and so the rank of the final solution W^* is equal to k.

The intuition behind (14) is to use a sequence of ℓ_1-minimization (or linear programs) to approximate the non-convex ℓ_0 minimization problem (11). While the problem addressed in Spielman et al. (2012) slightly differs from (14), their theoretical results may suggest us how to choose the analysis filter $\mathbf{r}$. Applying their results to our problem, we select $\mathbf{r}$ to be a column of H^* and choose the solution to be the one that results in minimum cardinality.

The details of the successive optimization for problem (6) are summarized in Algorithm 1. Here, $[\cdot]_i$ denotes the ith column of a matrix. Note that the proposed method only has the number of atoms k to tune. Therefore, it generates consistent results for a given dataset and k.

Algorithm 1: Successive optimization for (6).

input : Data matrix D, and number of atoms k.
initialize $\hat{W} = \mathbf{0}$ and $\hat{F} = \mathbf{0}$.
compute $\overline{V}^*$, $\overline{C}^*$, H^*, and E^* by (13).
for $i = 1, ..., k$ **do**
  for $j = 1, ..., np$ **do**
    choose $\mathbf{r} = [H^*]_j$.
    compute $\{\hat{\boldsymbol{w}}_{ij}, \hat{\boldsymbol{f}}_{ij}\}$ by (14).
  end
  compute $\ell \in \arg\min_j \|\hat{\boldsymbol{w}}_{ij}^T H^* - \hat{\boldsymbol{f}}_{ij}^T \otimes \mathbf{1}^T\|_0$.
  update $(\hat{\boldsymbol{w}}_i, \hat{\boldsymbol{f}}_i) = (\hat{\boldsymbol{w}}_{i\ell}, \hat{\boldsymbol{f}}_{i\ell})$.
  update $[\hat{W}]_i = \hat{\boldsymbol{w}}_i$ and $[\hat{F}]_i = \hat{\boldsymbol{f}}_i$.
end
update $(W^*, F^*) = (\hat{W}^T, \hat{F}^T)$.
compute C^*, V^*, S^* by (12).
output: solution (V^*, C^*, S^*, E^*).

Example 1 To illustrate the illumination dictionary model in (5), we conduct a simple experiment on Extended YaleB database (Lee et al. 2005). Only the frontal images of the 38 subjects in the database are included. Figure 1 illustrates the learned identity vectors of the first 10 subjects in V and the first 10 atoms in the illumination dictionary C.

In the next section, we will propose an extension of the SRC framework using SILT, which is aimed at addressing both alignment and recognition with small gallery samples. In particular, among the estimates from Algorithm 1, only the illumination dictionary C will be used in the subsequent sparse illumination transfer process. We should emphasize here that in the literature, there are several other algorithms

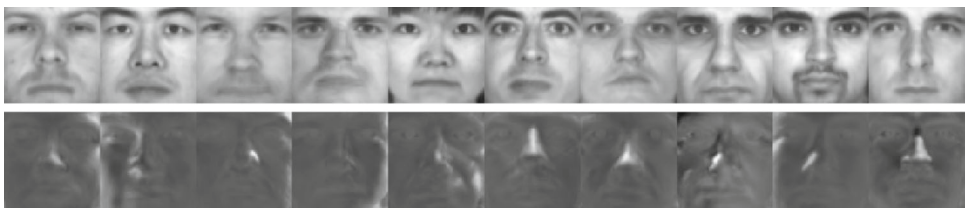

Fig. 1 *Top:* First 10 columns of V unstacked as subject identity images. *Bottom:* First 10 columns of C unstacked as illumination images

Fig. 2 Single-sample alignment results on Multi-PIE. The solid red boxes are the initial face locations provided by a face detector. The *dash green boxes* show the alignment results. The subject image on the right has 30 % of the face pixels corrupted by random noise

that deal with illumination transfer functions, such as the quotient image (Shashua and Riklin-Raviv 2001; Peers et al. 2007) and edge-preserving filters (Chen et al. 2011). The focus of this paper is to learn an illumination dictionary for single-sample alignment and recognition in the SRC framework. The approach of adding an auxiliary dictionary to help recognition was also considered in Deng et al. (2012); Yang et al. (2013). However, most of these illumination transfer methods are only for recognition but not alignment.

4 Robust Single-Sample Face Alignment and Recognition using SILT

4.1 Robust Single-Sample Alignment

Without loss of generality, we assume each gallery class only contains one sample $A_i = \boldsymbol{a}_i$. It is important to note that in our problem setting, each $\boldsymbol{a}_i$ can be sampled from an arbitrary lighting condition, and we do not assume the gallery images to share the same illumination pattern. In the alignment stage, given a query image $\boldsymbol{b}$, we estimate an image transformation τ_i applied in the 2-D image coordinates of $\boldsymbol{b}$ to align it with $\boldsymbol{a}_i$. Clearly, if one were to directly apply the standard SRC solution (2), the so-defined alignment error $\boldsymbol{e} = \boldsymbol{b} \circ \tau_i - \boldsymbol{a}_i x_i$ may not be sparse. More specifically, the different illumination conditions between $\boldsymbol{b}$ and $\boldsymbol{a}_i$ may introduce a *dense* alignment error even when the two images are perfectly aligned. Although an alignment error can still be minimized with respect to an ℓ_1-norm or ℓ_2-norm penalty, the algorithm would lose its robustness when concurrently handling sparse image corruption and facial disguise.

The SILT algorithm mitigates the problem by using the sparsely-used illumination dictionary C to compensate the illumination difference between $\boldsymbol{b}$ and $\boldsymbol{a}_i$. More specifically, SILT alignment solves the following problem:

$$(\hat{\tau}_i, \hat{x}_i, \hat{\boldsymbol{y}}_i) = \arg\min_{\tau_i, x_i, \boldsymbol{y}_i, \boldsymbol{e}} \|\boldsymbol{y}_i\|_1 + \lambda \|\boldsymbol{e}\|_1 \quad \text{subj. to} \quad \boldsymbol{b} \circ \tau_i = \boldsymbol{a}_i x_i + C\boldsymbol{y}_i + \boldsymbol{e}. \tag{15}$$

In (15), $\lambda > 0$ is a parameter that balances the weight of $\boldsymbol{y}_i$ and $\boldsymbol{e}$, which can be chosen empirically. C is the SILT dictionary learned in Algorithm 1. Finally, the objective function (15) can be solved efficiently using ℓ_1-minimization techniques such as those discussed in Wagner et al. (2012); Yang et al. (2013). Figure 2 shows two examples of the SILT alignment results.

4.2 Robust Single-Sample Recognition

Next, we propose a novel face recognition algorithm that extends the SRC framework to the single-sample case. Similar to the above alignment algorithm, the algorithm also applies trivially when multiple gallery samples are available per class.

In the previous SRC framework (3), once the transformation τ_i is recovered for each class A_i, the transformation is carried over to the recognition stage, where the gallery images in A_i are transformed by τ_i^{-1} to align with the query image $\boldsymbol{b}$. In the single-sample case, the sparse representation model in (3) will not be satisfied due to two reasons. First, as the A matrix only contains one sample per class, even when $\boldsymbol{b}$ is a valid query image with no gross image corruption or facial disguise, the equality constraint $\boldsymbol{b} = A\boldsymbol{x}$ typically will not hold true. As a result, it becomes difficult to classify $\boldsymbol{b}$ based on the sparse coefficients of $\boldsymbol{x}$ as suggested in SRC. Second, as the illumination condition of $\boldsymbol{b}$ may not be fully expressed by the linear combination $A\boldsymbol{x}$, it causes the error $\boldsymbol{e} = \boldsymbol{b} - A\boldsymbol{x}$ to be dense, mostly to compensate the difference in their illumination conditions. The problem reduces the effectiveness of SRC to compensate gross image corruption by minimizing the sparsity of $\boldsymbol{e}$.

In the SILT framework, we have seen in (15) that if an auxiliary illumination dictionary C is provided, it can be used to compensate the missing illumination information in single gallery images $\boldsymbol{a}_i$. Therefore, in the recognition stage, one may consider transfer both the illumination information $C\hat{\boldsymbol{y}}_i$ and alignment $\hat{\tau}_i$ to compensate each $\boldsymbol{a}_i$:

$$\tilde{\boldsymbol{a}}_i = (\boldsymbol{a}_i \hat{x}_i + C\hat{\boldsymbol{y}}_i) \circ \hat{\tau}_i^{-1}. \tag{16}$$

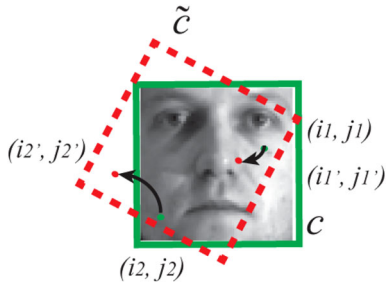

Fig. 3 Warping a cropped auxiliary image by τ_i^{-1} may result in copying some pixel values that are out of bound. The values of these out-of-bound pixels are not available in (16). In this example, the pixel with the coordinates (i_1', j_1') after transformation τ_i^{-1} remains within the original bounding box in green color, but (i_2', j_2') is outside the original bounding box. Pixel coordinates such as (i_2, j_2) should be removed from the support set Ω

The collection of all the warped gallery images is defined as $\tilde{A} = [\tilde{\boldsymbol{a}}_1, \tilde{\boldsymbol{a}}_2, \ldots, \tilde{\boldsymbol{a}}_L]$.

Unfortunately, a careful examination of this proposal (16) reveals a rather subtle issue that prevents us to directly apply the warped gallery images $\tilde{A}$ in the recognition stage. The problem lies in the fact that the illumination dictionary C is learned from the auxiliary face images with the frontal position that are typically cropped and normalized. As a result, the atoms of the dictionary C cannot be simply warped by an image transformation τ_i^{-1}. An exact solution to update the pose of the illumination dictionary C would require the algorithm to first warp the auxiliary images themselves in D, and then retrain the illumination dictionary $C_{\tau_i^{-1}}$ for each transformation τ_i. Clearly, this task is prohibitively expensive.[2]

In addition, applying (16) that warps auxiliary images and gallery images to the query image sometimes can be undesirable in practice. Figure 3 illustrates the problem. In many cases, the auxiliary and gallery images are provided only within a cropped face region. Therefore, any pixel outside the original bounding box may not have a valid value. In some other cases, even when those pixels are available, still the original pixels within the training bounding box are typically well chosen to best represent the appearance of the face. As a result, using pixel values outside the bounding box may negatively affect the accuracy of the recognition.

In this paper, we propose a more efficient solution to address the problem. The key idea is to constrain the sparse representation-based classification on a subset of pixels whose pixel values remain valid after the alignment compensation (16).

Without loss of generality, we assume each auxiliary image in D is of dimension $w \times h$, i.e., $d = wh$. In the SILT recognition step, given an estimated transformation from the alignment stage τ_i^{-1} for the gallery image $\boldsymbol{a}_i$, we apply the transformation τ_i^{-1} on each pixel within the face image $(i, j) \in [1, w] \times [1, h]$. Define the support set for the transformation τ_i^{-1}:

$$\Omega_i \doteq \{(i, j) | \tau_i^{-1}(i, j) \in [1, w] \times [1, h]\}. \tag{17}$$

Given all the collection of all the transformations $\tau_1, \tau_2, \ldots, \tau_L$, we define the total support set Ω as the intersection

$$\Omega = \bigcap_{i=1}^{L} \Omega_i, \tag{18}$$

that is, each element in Ω corresponds to a valid pixel in the auxiliary images and C after the transformations τ_i^{-1} are applied for all $i = 1, \ldots, L$. The projection of an image in vector form $\boldsymbol{b}$ onto a support set Ω is denoted as $\mathcal{P}_\Omega(\boldsymbol{b}) \in \mathbb{R}^{|\Omega|}$.

The effect of applying a mask defined by a support set Ω is illustrated in Fig. 4. Initially, the input query images in the first column and the gallery images of the same subjects have very different illumination conditions and poses. In the third column, an illumination transfer pattern is estimated for each gallery image. For example, in the second subject example, the left side of $\boldsymbol{b}$ is brighter than that of $\boldsymbol{a}$. This is reflected by having a brighter illumination pattern in its $C\hat{\boldsymbol{y}}$. Finally, the gallery images are further warped based on the estimated poses τ^{-1}, and the masks of their support sets Ω are applied to both the warped gallery images $\mathcal{P}_\Omega(\tilde{\boldsymbol{a}})$ and the query images $\mathcal{P}_\Omega(\boldsymbol{b})$). We can see that, compared to the input image pairs $(\boldsymbol{a}, \boldsymbol{b})$, the processed image pairs in the SILT algorithm $(\mathcal{P}_\Omega(\tilde{\boldsymbol{a}}), \mathcal{P}_\Omega(\boldsymbol{b}))$ have closer illumination conditions and similar poses.

The remaining SILT algorithm involves solving a sparse representation $\boldsymbol{x}$ in the presence of a possible sparse error $\boldsymbol{e}$ constrained on the support set Ω, namely,

$$\begin{aligned} (\boldsymbol{x}^*, \boldsymbol{e}^*) &= \arg\min_{\boldsymbol{x}, \boldsymbol{e}} \|\boldsymbol{x}\|_1 + \lambda \|\boldsymbol{e}\|_1 \\ \text{subj. to} \quad & \mathcal{P}_\Omega(\boldsymbol{b}) = \mathcal{P}_\Omega(\tilde{A})\boldsymbol{x} + \boldsymbol{e}, \end{aligned} \tag{19}$$

where the operation $\mathcal{P}_\Omega(\tilde{A})$ applies pixel selection on each column of $\tilde{A}$ based on the support set Ω. Similar to the previous formulations, the parameter λ is chosen empirically via cross validation.

Using the sparse representation $\boldsymbol{x}$ in (19), the final decision rule to classify $\boldsymbol{b}$ can be simplified from the original SRC algorithm in Wright et al. (2009) where the reconstruction

[2] In our previous work (Zhuang et al. 2013), this simple extension was in fact used as the solution to transfer both the alignment and illumination information from the alignment stage to the recognition stage. However, the assumption was valid because the illumination dictionary used in Zhuang et al. (2013) was constructed by concatenating the auxiliary images themselves, namely, D in this paper. Therefore, the problem of warping a learned dictionary was mitigated.

Fig. 4 Examples of warping a gallery image $\tilde{\boldsymbol{a}}$ and applying a mask Ω on both the query image $\boldsymbol{b}$ and the warped gallery image $\tilde{\boldsymbol{a}}$. **a** Query images $\boldsymbol{b}$. **b** Gallery images $\boldsymbol{a}$. **c** Illumination transfer information $C\hat{\boldsymbol{y}}$. **d** Warped gallery images $\tilde{\boldsymbol{a}}$ under a mask Ω. **e** Applying the same masks Ω on $\boldsymbol{b}$

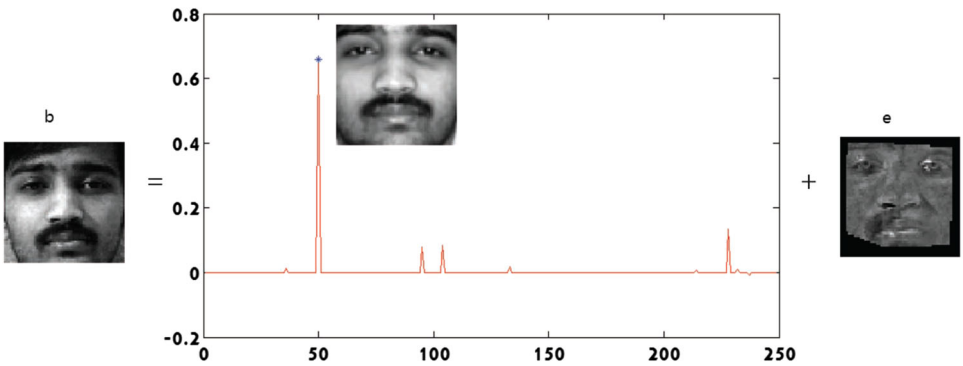

Fig. 5 Illustration of SILT recognition. *Left:* Query image $\boldsymbol{b}$ with unknown pose and illumination. *Right:* Sparse representation $\boldsymbol{x}$ with the correct gallery image $\boldsymbol{a}_i$ superimposed and sparse error $\boldsymbol{e}$. The effect of pose alignment between $\boldsymbol{b}$ and all the 250 gallery images is illustrated by the mask Ω shown in $\boldsymbol{e}$

residual was used. In SILT, since there is only one sample per each subject class in A, the class with the largest coefficient magnitude in $\boldsymbol{x}$ is the estimated class of the query image $\boldsymbol{b}$. We note that this simplified strategy does not compromise the generality of the SILT method, as one can still estimate the objective function of the reconstruction residual when each class contains one or more gallery images. Figure 5 shows an example of the SILT recognition and its estimated sparse representation.

Before we move on to examine the performance of the new recognition algorithm (19), one may question the efficacy of enforcing a sparse representation in the constraint (19). The question may arise because in the original SRC framework, the data matrix $A = [A_1, \ldots, A_L]$ is a collection of highly correlated image samples that span the L illumination subspaces. Therefore, it makes sense to enforce a sparse representation as also validated by several followup studies (Wright and Ma 2010; Elhamifar and Vidal 2012; Zhang et al. 2012). However, in single-sample recognition, only one sample $\boldsymbol{a}_i$ is provided per class. Therefore, one would think that the best recognition performance can only be achieved by the nearest-neighbor algorithm.

There are at least two arguments to justify the use of sparse representation in (19). One one hand, as discussed in Wright et al. (2009), in the case where $\boldsymbol{e}$ and $C\boldsymbol{y}$ represent a small error and the nearest-neighbor solution corresponds to a one-sparse binary vector $\boldsymbol{x}_0 = [\cdots, 0, 1, 0 \cdots]^T$ in the formulation (19), then solving (19) via ℓ_1-minimization can also recover the sparsest solution, namely, $\boldsymbol{x}^* \approx \boldsymbol{x}_0$. On the other hand, in the case where $C\boldsymbol{y}$ represents a large illumination change and $\boldsymbol{e}$ represents additional gross image corruption, as long as the elements of A in (19) remain tightly correlated in the image space, the ℓ_1-minimization algorithm can compensate the dense error in the query image $\boldsymbol{b}$ (Wright and Ma 2010). This is a unique advantage over nearest-neighbor type algorithms.

5 Experiment

In this section, we present a comprehensive experiment to demonstrate the performance of our illumination learning, face alignment, and recognition algorithms.

The illumination dictionary is constructed from Extended YaleB database (Lee et al. 2005). The Extended YaleB contains 21888 face image of 38 subjects under 9 poses and 64 illumination conditions. For every subject in a particular pose, an image with ambient (background) illumination was also captured. In this paper, only the frontal images of the 38 subjects are used as the auxiliary images.

For the gallery and query subjects, we choose images from a much larger CMU Multi-PIE database (Gross et al. 2008). Except for Sect. 5.4, 166 shared subject classes from Session 1 and Session 2 are selected for testing. In Session 1, we randomly select one frontal image per class with arbitrary illumination as the gallery image. Then we randomly select two different frontal images from Session 1 or Session 2 for testing. The outer eye corners of both training and query images are manually marked as the ground truth for registration. All the training face images are manually cropped into 60×60 pixels based on the locations of eyes out-corner points, and the distance between the two outer eye corners is normalized to be 50 pixels for each person. We again emphasize that our experimental setting is more practical than those used in some other publications, as we allow the training images to have arbitrary illumination and not necessarily just the ambient illumination.

We compare our algorithms with several state-of-the-art face alignment and recognition algorithms under the SRC framework. To conduct a fair comparison, it is important to separate those algorithms that were originally proposed to handle only the recognition problem versus those that can handle both face alignment and recognition. The original SRC algorithm (Wright et al. 2009), the Extended SRC (ESRC) (Deng et al. 2012), and SVDL (Yang et al. 2013) belong to the first case, while Deformable SRC (DSRC) (Wagner et al. 2012), misalignment robust representation (MRR) (Yang et al. 2012), and SILT proposed in this paper belong to the second case.

Finally, as the SILT algorithm relies on an auxiliary illumination dictionary C, another variability we need to investigate further is how the choice of C may affect the performance of SILT. Our investigation on this issue will be divided in three steps. First, in Sect. 5.1, we validate in an ideal, noise-free simulation that the proposed dictionary learning algorithm can successfully recover the subject identity matrix V and the illumination dictionary C in (6). We further utilize Extended YaleB database to construct an illumination dictionary from the real face images. Second, in Sect. 5.4, we will compare the recognition rates of SILT using different illumination dictionaries. The experiment further shows the SILT framework significantly outperforms DSRC and MRR in single-sample face recognition with misalignment and pixel corruption. Finally, in Sect. 5.5, we again use Extended YaleB database to illustrate how the variation in the atom size and the training subjects of the auxiliary data affects the performance of the SILT algorithm.

5.1 Learning Illumination Dictionaries

In this experiment, we validate the performance of the illumination dictionary learning algorithm in Algorithm 1. First, we use noise-free synthetic data to evaluate the success rate for the algorithm to recover a subject-identity matrix V and a sparsely-used dictionary C as in (4). Specifically, the elements in the $V \in \mathbb{R}^{d\times p}$ and $C \in \mathbb{R}^{d\times k}$ matrices are generated from independent and identically distributed (i.i.d.) Gaussian distributions. The columns of the sparse coefficient matrix $S \in \mathbb{R}^{k\times np}$ are assumed to be t-sparse, where each column has exactly t non-zero coefficients, where $n \doteq k \log_e k$ is the number of samples from each class and varies with the atom size k. These synthesized ground-truth matrices then generate the data matrices $D_1, D_2, \ldots, D_p \in \mathbb{R}^{d\times n}$.

In the experiment, we set $d = 100$, $p = 5$, and let k vary between 10 and 50 and t between 1 and 10. In addition, to resolve the potential ambiguity in the permutation of the estimated dictionary atoms, we adopt the following relative error metric to a performance index:

$$\phi(Z^*, Z) = \min_{\Pi,\Lambda} \|Z^*\Pi\Lambda - Z\|_F / \|Z\|_F \tag{20}$$

where Π is a permutation matrix, and Λ is a diagonal scaling matrix.

Figure 6 shows the simulation result. The average relative error (20) for both V and C is reported in grayscale, where the white blocks indicate zero error, and the darker blocks indicate larger relative error. We can clearly see that when the dictionary size k is sufficiently large and when the sparsity t sufficiently small, Algorithm 1 perfectly recovers the two matrices. The algorithm only fails when $k = 10, t < 3$ and $k = 20, t = 10$. Furthermore, the phase transition from failed recovery settings to perfect recovery settings is quite sharp.

Next, we apply Algorithm 1 to learn the illumination dictionary C from Extended YaleB database. For the experimental purpose in the subsequent sections, we construct two dictionaries with very different settings:

1. *Ad-Hoc Dictionary*: We choose the very first subject in Extended YaleB database with 65 aligned frontal images (1 ambient + 64 illuminations). The dictionary C is directly constructed by subtracting the ambient image from the other 64 illumination images, and no additional learning algorithm is involved. This dictionary is

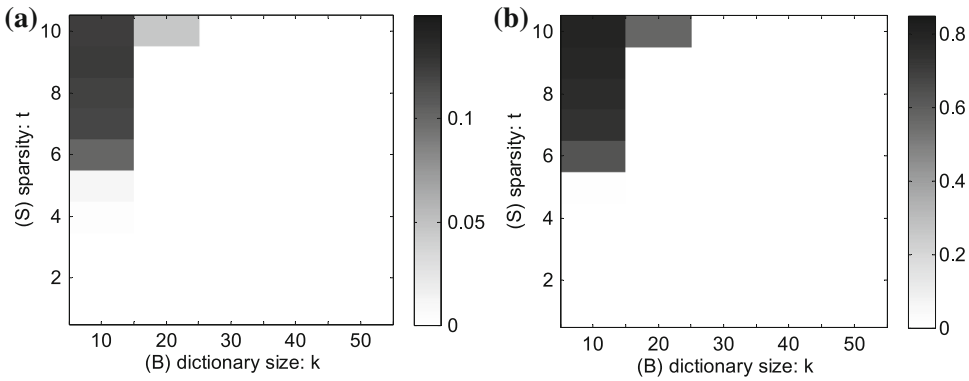

Fig. 6 Mean relative errors over 5 trials, with varying support t and basis size k for **a** V and **b** C estimated by Algorithm 1

Fig. 7 Illustration of the first 10 atoms of the illumination dictionary C. *Top:* Ad-Hoc Dictionary constructed from the first subject of Extended YaleB database. *Bottom:* Yale Dictionary learned from all the 38 subjects

identical to the one used in our previous work (Zhuang et al. 2013).

2. *Yale Dictionary*: We employ all the 38 subjects in Extended YaleB database to learn an illumination dictionary using Algorithm 1.

Some atoms from the above two dictionaries are shown as Fig. 7. The atom size of Yale Dictionary in this illustration is fixed at 80. In Sect. 5.4 and 5.5, we will compare the performance of different dictionaries.

5.2 Simulation on 2D Alignment

In this experiment, we demonstrate the performance of the SILT alignment algorithm (15). The performance is measured using simulated 2D deformation on the face image, including translation, rotation and scaling. Without loss of generality, we will only use Yale Dictionary as our illumination dictionary. The added deformation is introduced to the query images based on the ground truth coordinates of eye corners. The translation ranges from [-12, 12] pixels with a step size of 2 pixels.

Similar to Wagner et al. 2012, we use the estimated alignment error $\|\boldsymbol{e}\|_1$ as an indicator of success. More specifically, let $\boldsymbol{e}_0$ be the alignment error obtained by aligning a query image from the manually labeled position to the training images. We consider the alignment successful if $|\|\boldsymbol{e}\|_1 - \|\boldsymbol{e}_0\|_1| \leq 0.01\|\boldsymbol{e}_0\|_1$.

We compare our method with DSRC and MRR. As DSRC and MRR would require to have multiple reference images per class, to provide a fair comparison, we evaluate both algorithms under two settings: Firstly, seven reference images are provided per class to DSRC.[3] We denote this case as DSRC-7. Secondly, one randomly chosen image per class as the same setting as in the SILT algorithm. We denote this case as DSRC-1 and MRR-1, respectively.

We draw the following observations from the alignment results shown in Fig. 8:

1. SILT works well under a broad range of 2D deformation, particularly when the translation in x or y direction is less than 20% of the eye distance (10 pixels) and when the in-plane rotation is less than 30 degrees.
2. Clearly, SILT outperforms both DSRC-1 and MRR-1 when the same setting is used, namely, one sample per class. The obvious reason is that DSRC and MRR were not designed to handle the single-sample alignment scenario.
3. The accuracy of SILT and DSRC-7 is generally comparable across the board in all the simulations. However, since DSRC-7 has access to seven gallery images of different illumination conditions, the result shows the power of using the new illumination dictionary in (15), where SILT only works with a single gallery image.

5.3 Single-Sample Recognition

In this subsection, we evaluate the SILT recognition algorithm based on single reference images of the 166 subject classes shared in Multi-PIE Sessions 1 and 2. We compare its performance with SRC (Wright et al. 2009), ESRC (Deng et al. 2012), DSRC (Wagner et al. 2012), MRR (Yang et al.

[3] The training are illuminations {0,1,7,13,14,16,18} in Multi-PIE Session 1.

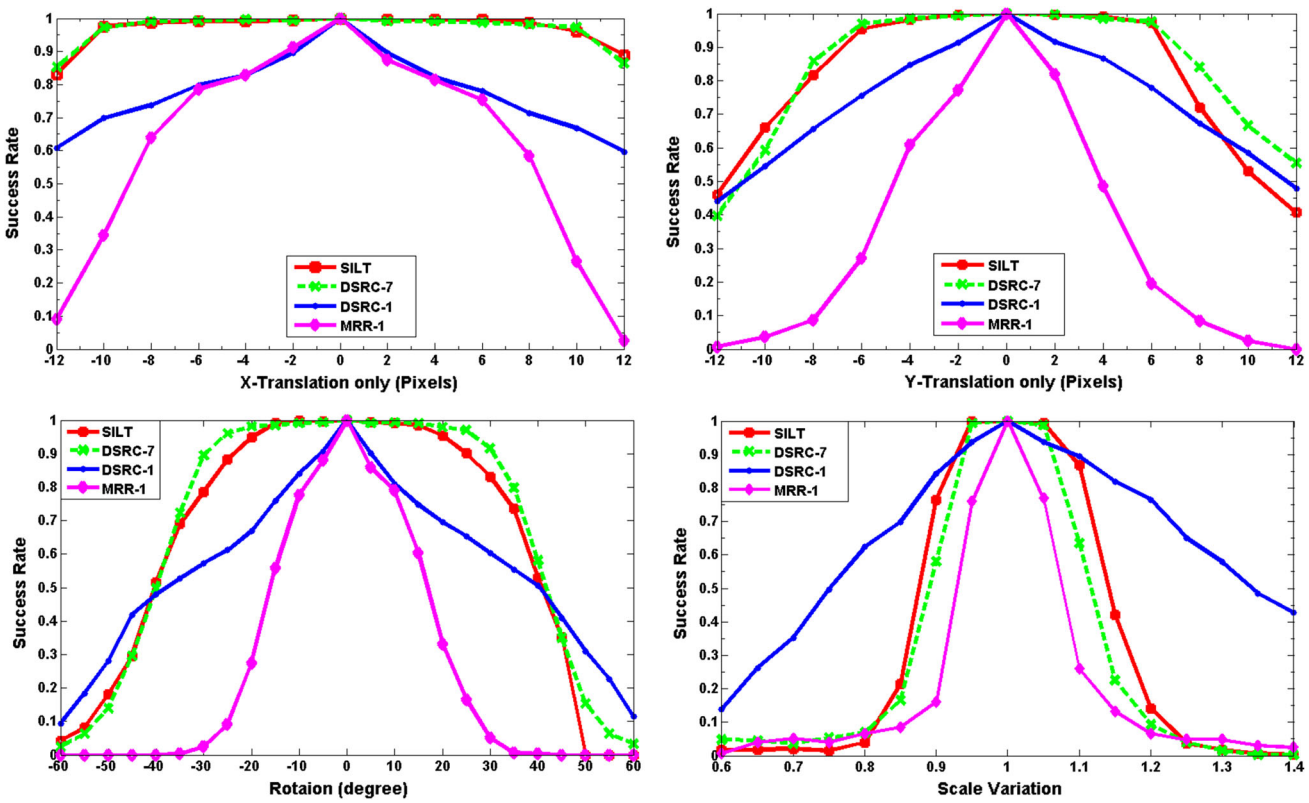

Fig. 8 Success rate of face alignment under four types of 2D deformation: x-translation, y-translation, rotation, and scaling. The amount of translation is expressed in pixels, and the in-plane rotation is expressed in degrees

Table 1 Single-sample recognition accuracy via manual alignment. The atom size is fixed to 80

Method	Session 1 (%)	Session 2 (%)
SRC_M	88.0	53.6
ESRC_M	89.6	56.6
SILT + SRC_M	92.8	59.0
SILT + ESRC_M	**93.2**	**59.3**
SVDL_M	70.3	41.6

2012), and SVDL (Yang et al. 2013). The illumination dictionary used in these experiments is Yale Dictionary.

First, we note that the new SILT framework and the existing sparse representation algorithms are *not* mutually exclusive. In particular, the illumination transfer (16) can be easily adopted by the other algorithms to improve the illumination condition of the training images, especially in the single-sample setting. In the first experiment, we demonstrate the improvement of SRC and ESRC with the illumination transfer. Since both algorithms do not address the alignment problem, manual labels of the face location are assumed to be the aligned face location. The comparison is presented in Table 1.

Since the gallery images are selected from Session 1, there is no surprise that the average recognition rate of Session 1 is significantly higher than that of Session 2. The comparison further shows that adding the illumination transfer information to the existing SRC and ESRC algorithms meaningfully improves their performance by 3% – 5%.

In Table 1, the performance of the SVDL algorithm is also shown.[4] Interestingly, in our setting of single-sample recognition, SVDL performs worse than SRC and ESRC. A possible explanation is that the SVDL algorithm expects all the gallery images to have the same uniform lighting condition, while in this paper, the illumination condition of the gallery images is randomly selected. Our experimental setting is more challenging but more similar to the single-sample face recognition problem in practice. Furthermore, one can consider combining the SILT framework and the illumination dictionary of SVDL. This variation will be considered in Sect. 5.4.

Second, we compare DSRC, MRR, and SILT in the full pipeline of alignment plus recognition shown in Table 2. The initial positions of the face images are automatically detected by Viola-Jones detector.

Compared with the past reported results of DSRC and MRR, their recognition accuracy decreases significantly

[4] The implementation of SVDL was provided by their authors at: http://www4.comp.polyu.edu.hk/~cslzhang/code/SVDL.zip.

Table 2 Single-sample alignment + recognition accuracy

Method	Session 1 (%)	Session 2 (%)
DSRC	36.1	35.7
MRR	46.2	34.6
SILT	**76.7**	**61.6**

when only one training image is available per class. It demonstrates that these algorithm were not designed to perform well in the single-sample regime. In both Session 1 and Session 2, SILT outperforms both algorithms by more the 30%. It is more interesting to compare the recognition rates of different algorithms on Session 2 in Table 1 and Table 2. SILT that relies on an auxiliary illumination dictionary to automatically alignment the query images achieves 61.6%, which is even higher than the ESRC rate of 59.3% with manual alignment.

5.4 Robustness under Random Corruption

In this subsection, we further compare the robustness of the SILT recognition algorithm to random pixel corruption. We compare the overall recognition rate of SILT with DSRC, and MRR, the two most relevant algorithms. For the SILT algorithm, in addition to using the two previous illumination dictionaries, namely, Ad-Hoc and Yale, we also demonstrate the performance using the SVDL dictionary (Yang et al. 2013).

To benchmark the recognition under different corruption percentage, it is important that the query images and the gallery images have close facial appearance, otherwise different facial features would also contribute to facial corruption or disguise, such as glasses, beard, or different hair styles. To limit this variability, in this experiment, we select both query and gallery images from Multi-PIE Session 1, although the images should never overlap. We use all the subjects in Session 1. For each subject, we randomly select one frontal image with arbitrary illumination for testing. Various levels of image corruption from 10% to 40% are randomly generated in the face region. Similar to the previous experiments, the face regions are detected by Viola-Jones detector. The performance is shown in Table 3.

Table 3 Recognition rates (%) under various percentage of random pixel corruption. The atom size is fixed to 80

Corruption	10%	20%	30%	40%
DSRC	32.9	31.7	28.9	24.1
MRR	24.9	14.5	11.7	9.2
SILT(Ad-Hoc)	66.2	59.8	49.6	44.7
SILT(Yale)	**73.3**	**68.7**	**67.3**	**49.0**
SILT(SVDL)	60.0	56.1	52.3	41.1

The comparison is more illustrative than Table 2. First of all, all three SILT implementations based on very different illumination dictionaries significantly outperform DSRC and MMR. For instance, with 40% pixel corruption, SILT still maintains 49% accuracy; with 10% corruption, SILT outperforms DSRC and MRR by more than 40%.

Second, we note that in the presence of pixel corruption, the illumination dictionary learned by SVDL does not perform as well as Ad-Hoc and Yale dictionaries. It shows that our proposed dictionary learning method is more suited for estimating auxiliary illumination dictionaries in the SILT framework.

5.5 Influence of Atom Size and Subject Number

In this section, we discuss how the efficacy of an SILT dictionary may be affected by the choice of the atom size and the subject number. More specifically, We learn illumination dictionaries using Algorithm 1 from Extended YaleB database with varying number of the auxiliary subjects and atom size of the dictionary. Then, we measure the accuracy of face recognition under the frameworks of "SILT+ESRC$_M$" and "SILT+SRC$_M$" with manual alignment. The settings is the same as Sect. 5.3, namely, gallery and query images are chosen from Session 1 of Multi-PIE database. The results are shown in Table 4 and Table 5.

First, we notice that there is no data point taken at 200 atom size when the subject number is one. This is due to the fact that each subject in Extended YaleB database only provides 65 frontal images. When one tries to solve for more atoms in the corresponding illumination dictionary in (6), the problem becomes ill-conditioned. This issue can be first observed by examining the recognition rates for one subject and atom sizes greater than 60, namely, 80 and 120. In these two settings, the recognition rates are either identical

Table 4 Recognition rates (%) under the SILT+ESRC$_M$ implementation with manual alignment

atom size	40	60	80	120	200
subject # = 1	89.6	89.2	89.2	89.2	-
subject # = 10	90.0	92.8	92.8	94.0	94.8
subject # = 38	90.8	93.2	93.2	95.2	**96.8**

Table 5 Recognition rates (%) under the SILT+SRC$_M$ implementation with manual alignment

atom size	40	60	80	120	200
subject # = 1	86.8	88.0	87.2	87.2	-
subject # = 10	87.2	91.2	92.4	90.8	92.8
subject # = 38	91.2	93.2	92.8	94.8	**95.6**

or slightly worse than those at atom size 60 in both Table 4 and Table 5. At atom size 200, through visual inspection, we discover that the illumination patterns in the estimated C matrices are close to random noise, and do not contain useful illumination information for the SILT algorithm. Therefore, their performance is ignored.

Second, when the subject number is higher than one, increasing the atom size of the illumination dictionary clearly improves the recognition rate. For example, using all the 38 subjects and the SILT+ESRC$_M$ algorithm, the recognition rate using a 40-atom illumination dictionary is 90.8%. The rate is raised to 96.8% when the atom size increases to 200. It is worth emphasizing that this recognition rate represents one of the best accuracy on Multi-PIE database when only single gallery images of random illumination are available, to the best of our knowledge.

Finally, it comes as no surprise that if we fix the size of the illumination dictionary in each column of Table 4 and Table 5, including more subjects in the illumination database also improves the recognition. This phenomenon can be explained by considering the well-known Lambertian model of the human face. It states that the image appearance of a face is determined not only by the illumination of the environment, but also by the shape of the face and its surface albedo pertaining to individual subjects. Therefore, having more subjects would help to generalize the distribution of the illumination patterns under different face shape and albedo. Then, the use of sparse representation in the alignment and recognition algorithms can effectively select a sparse subset of these illumination patterns that are most similar to the illumination, shape, and albedo condition of the query image.

6 Conclusion and Discussion

In this paper, we have presented a novel face recognition algorithm specifically designed for single-sample alignment and recognition. To compensate for the missing illumination information traditionally provided by multiple gallery images, we have proposed a novel dictionary learning algorithm to estimate an illumination dictionary from auxiliary training images. We have further proposed an illumination transfer technique to transfer the estimate illumination compensation and pose information from the face alignment stage to the recognition stage. The overall algorithm is called *sparse illumination learning and transfer* (SILT). The extensive experiment has validated that not only the standalone SILT algorithm outperforms the state of the art in single-sample face recognition by a significant margin, the illumination learning and transfer technique is also complementary to many existing algorithms as a pre-processing step to improve the image condition due to misalignment and pixel corruption.

Although we have provided some exciting results that represent a meaningful step forward towards a real-world face recognition system in this paper, one of the open problems remains to be how to improve illumination transfer in complex real-world conditions and with minimal training data. Although the current way of constructing the illumination dictionary is efficient, the method is not able to separate the effect of surface albedo, shape, and illumination completely from face images. Therefore, we believe a more sophisticated illumination transfer algorithm could lead to better overall performance.

Acknowledgments The work was supported in part by ARO 63092-MA-II, DARPA FA8650-11-1-7153, ONR N00014-09-1-0230, NSF CCF09-64215, NSFC No. 61103134 and 61371192, and the Science Foundation for Outstanding Young Talent of Anhui Province (BJ2101020001).

Appendix

We proof Theorem 1 in this Appendix. First, eliminating the variable E of problem (10) with

$$E = D - \overline{V} \otimes \mathbf{1}^T - \overline{C}H, \tag{21}$$

Problem (10) can then be equivalently written as

$$\min_{\overline{V},\overline{C},H} \|D - \overline{V} \otimes \mathbf{1}^T - \overline{C}H\|_F^2, \text{ s.t. } \overline{C}^T\overline{C} = I. \tag{22}$$

As a basic result in least squares (Horn and Johnson 1985), the optimal H can be written as

$$H^* = \overline{C}^T(D - \overline{V} \otimes \mathbf{1}^T), \tag{23}$$

for any $\overline{V} \in \mathbb{R}^{m\times p}$ and any $\overline{C} \in \mathbb{R}^{m\times k}$ such that $\overline{C}^T\overline{C} = I$. Substituting H^* into (22) yields

$$\min_{\overline{V},\overline{C}} \|P_{\overline{C}}^{\perp}(D - \overline{V} \otimes \mathbf{1}^T)\|_F^2, \text{ s.t. } \overline{C}^T\overline{C} = I, \tag{24}$$

where $P_{\overline{C}}^{\perp} = I - \overline{C}\overline{C}^T$ denotes the orthogonal complement projector of $\overline{C}$. It is also easy to show from (24) that a solution of $\overline{V}$ is

$$[\overline{V}^*]_i = \frac{1}{n}D_i\mathbf{1},\ i = 1, ..., p. \tag{25}$$

Note that the solution $[\overline{V}^*]_i$ presents the mean vector of the data matrix D_i corresponding to subject i. Furthermore, by letting $U = D - \overline{V}^* \otimes \mathbf{1}^T$, problem (24) becomes $\min_{\overline{C}^T\overline{C}=I} \text{trace}(U^T P_{\overline{C}}^{\perp} U)$, and it is equivalent to

$$\overline{C}^* = \arg \max_{\overline{C}^T \overline{C} = I} \text{trace}(\overline{C}^T U U^T \overline{C}). \quad (26)$$

By Horn and Johnson (1985), an optimal solution $\overline{C}^*$ is known to be the k principal eigenvector matrix of UU^T; i.e.,

$$\overline{C}^* = [\, \boldsymbol{q}_1(UU^T), \boldsymbol{q}_2(UU^T), \ldots, \boldsymbol{q}_k(UU^T) \,]. \quad (27)$$

Hence, the problem solution (13) simply follows from (21), (23) (25), and (27). □

References

Aharon, M., Elad, M., & Bruckstein, A. (2006). The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, *54*(11), 4311–4322.

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

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

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122.

Chen, X., Chen, M., Jin, X., & Zhao, Q. (2011). Face illumination transfer through edge-preserving filters. In *Proceedings of the IEEE international conference on computer vision and pattern recognition*.

Cootes, T., Edwards, G., & Taylor, C. (1998). Active appearance models. In *Proceedings of the European conference on computer vision*.

Cootes, T., Taylor, C., & Graham, J. (1995). Active shape models—Their training and application. *Computer Vision and Image Understanding*, *61*, 38–59.

Deng, W., Hu, J., & Guo, J. (2012). Extended SRC: Undersampled face recognition via intraclass variant dictionary. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*, 1864–1870.

Do, C., & Ng, A. (2005). Transfer learning for text classification. In *Proceedings of NIPS*.

Elhamifar, E., & Vidal, R. (2012). Block-sparse recovery via convex optimization. *IEEE Transactions on Signal Processing*, *60*, 4094–4107.

Gabay, D., & Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. *Computers and Mathematics with Applications*, *2*, 17–40.

Ganesh, A., Wagner, A., Wright, J., Yang, A., Zhou, Z., & Ma, Y. (2011). Face recognition by sparse representation. In *Compressed Sensing: Theory and Applications*. Cambridge University Press.

Gross, R., Mathews, I., Cohn, J., Kanade, T., & Baker, S. (2008). Multi-PIE. In *Proceedings of the eighth IEEE international conference on automatic face and gesture recognition*.

Gu, L., & Kanade, T. (2008). A generative shape regularization model for robust face alignment. In *Proceedings of the European conference on computer vision*.

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

Ho, J., Yang, M., Lim, J., Lee, K., & Kriegman, D. (2003). Clustering appearances of objects under varying illumination conditions. In *Proceedings of the IEEE international conference on computer vision and pattern recognition*.

Horn, R. A., & Johnson, C. R. (1985). *Matrix analysis*. New York: Cambridge University Press.

Huang, J., Huang, X., & Metaxas, D. (2008). Simultaneous image transformation and sparse representation recovery. In *Proceedings of the IEEE international conference on computer vision and pattern recognition*.

Lampert, C., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE international conference on computer vision and pattern recognition*.

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

Liang, L., Xiao, R., Wen, F., & Sun, J. (2008). Face alignment via component-based discriminative search. In *Proceedings of the European conference on computer vision*.

Lucas, B., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of international joint conference on artificial intelligence*.

Peers, P., Tamura, N., Matusik, W., Debevec, P. (2007). Post-production facial performance relighting using reflectance transfer. In *Proceedings of ACM SIGGRAPH*.

Quattoni, A., Collins, M., & Darrell, T. (2008). Transfer learning for image classification with sparse prototype representations. In *Proceedings of the IEEE international conference on computer vision and pattern recognition*.

Saragih, J., Lucey, S., & Cohn, J. (2009). Face alignment through subspace constrained mean-shifts. In *Proceedings of the IEEE international conference on computer vision*.

Shashua, A., & Riklin-Raviv, T. (2001). The quotient image: Class-based re-rendering and recognition with varying illuminations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(2), 129–139.

Spielman, D. A., Wang, H., & Wright, J. (2012). Exact recovery of sparsely-used dictionaries. *Journal of Machine Learning Research*, *23*(18), 1–37.

Tseng, P. (1991). Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM Journal on Control and Optimization*, *29*, 119–138.

Viola, P., & Jones, J. (2004). Robust real-time face detection. *International Journal on Computer Vision*, *57*, 137–154.

Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Mobahi, H., & Ma, Y. (2012). Toward a practical face recognition: Robust pose and illumination via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(2), 372–386.

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

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

Yan, S., Liu, C., Li, S., Zhang, H., Shum, H., & Cheng, Q. (2003). Face alignment using texture-constrained active shape models. *Image and Vision Computing*, *21*, 69–75.

Yan, S., Wang, H., Liu, J., Tang, X., & Huang, T. (2010). Misalignment-robust face recognition. *IEEE Transactions on Image Processing*, *19*, 1087–1096.

Yang, A., Zhou, Z., Ganesh, A., Sastry, S., & Ma, Y. (2013). Fast ℓ_1-minimization algorithms for robust face recognition. *IEEE Transactions on Image Processing*, *22*(8), 3234–3246.

Yang, M., Gool, L.V., & Zhang, L. (2013). Sparse variation dictionary learning for face recognition with a single training sample per person. In *Proceedings of the IEEE international conference on computer vision*.

Yang, M., Zhang, L., & Zhang, D. (2012). Efficient misalignment-robust representation for real-time face recognition. In *Proceedings of the European conference on computer vision*.

Zhang, L., Feng, M. Y. X., Ma, Y., & Feng, X. (2012). Collaborative representation based classification for face recognition. Technical report, arXiv:1204.2358.

Zhao, W., Chellappa, R., Phillips, J., & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys, 35*(4), 399–458.

Zhuang, L., Yang, A. Y., Zhang, Z., Sastry, S., & Ma, Y. (2013). Single-sample face recognition with image corruption and misalignment via sparse illumination transfer. In *Proceedings of the IEEE international conference on computer vision and pattern recognition*, Portland, Oregon.