Towards a Practical Face Recognition System: Robust Alignment and Illumination by Sparse Representation

Andrew Wagner, *Student Member, IEEE,* John Wright, *Member, IEEE,*
Arvind Ganesh, *Student Member, IEEE,* Zihan Zhou, *Student Member, IEEE,* Hossein Mobahi,
and Yi Ma, *Senior Member, IEEE*

Abstract—Many classic and contemporary face recognition algorithms work well on public data sets, but degrade sharply when they are used in a real recognition system. This is mostly due to the difficulty of simultaneously handling variations in illumination, image misalignment, and occlusion in the test image. We consider a scenario where the training images are well controlled, and test images are only loosely controlled. We propose a conceptually simple face recognition system that achieves a high degree of robustness and stability to illumination variation, image misalignment, and partial occlusion. The system uses tools from sparse representation to align a test face image to a set of frontal training images. The region of attraction of our alignment algorithm is computed empirically for public face datasets such as Multi-PIE. We demonstrate how to capture a set of training images with enough illumination variation that they span test images taken under uncontrolled illumination. In order to evaluate how our algorithms work under practical testing conditions, we have implemented a complete face recognition system, including a projector-based training acquisition system. Our system can efficiently and effectively recognize faces under a variety of realistic conditions, using only frontal images under the proposed illuminations as training.

Index Terms—Face Recognition, Face Alignment, Illumination Variation, Occlusion and Corruption, Sparse Representation, Error Correction, Validation and Outlier Rejection.

1 Introduction

Face recognition applications to date have fallen into roughly two categories. Face recognition has recently seen a lot of success in a family of less-demanding applications such as online image search and family photo album organization (e.g. Google Picassa, Microsoft Photo Gallery, and Apple iPhoto). At the other end of the tractability spectrum there are the terrorist watchlist and mass surveillance applications that have for the most part dominated the field of face recognition research. However, there are many face recognition applications that fall roughly between these extremes, where very high recognition performance is desired, but the users in the gallery are still allies of the system rather than adversaries. These applications include access control for secure facilities (e.g., prisons, office buildings), computer systems, automobiles, or automatic teller machines, where controlled gallery images can be obtained in advance. These applications are very interesting due to their potential sociological impact. Since the gallery subjects are allies, rather than opponents, of the recognition system, this creates the possibility of carefully controlling the acquisition of the training data. While the same can be said for other biometrics such as fingerprints and iris recognition, face recognition has the potential of working with test data that is much less controlled, allowing the access control system to be made less intrusive to the users of the system. To some extent, the goal of this paper is to show how a reliable face recognition system can be built for this restricted, but still important, scenario.

Very few recognition systems specifically target applications where many well- controlled training images are available. Of these, the classical holistic subspace-based face recognition methods [1], [2] are well known for their speed and simplicity, as well as for their natural extension to linear illumination models. However, their performance has been shown to be extremely brittle not only to alignment variation, but to even minor occlusions caused by, say, a wisp of hair, a blinked eye, or mouth that is slightly open. One promising recent direction, set forth in [3], casts the recognition problem as one of finding a sparse representation of the test image in terms of the training set as a whole, up to some sparse error due to occlusion. A *sparse representation-based classification* (SRC) method is then proposed for recognition. The main idea is that the sparse nonzero coefficients should concentrate on the training samples with the same class label as the test sample. SRC has demonstrated striking recognition performance despite severe occlusion or corruption by solving a simple convex program.

Unfortunately, while the work [3] achieves impressive results, it does not deal with misalignment between the test and training images, and it requires a rich set of illuminations in the gallery images for good performance. We illustrate the compounded effect of both alignment and illumination with

A. Wagner, A. Ganesh, Z. Zhou, and Y. Ma are with the Dept. of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. H. Mobahi is with the Computer Science Dept. at the University of Illinois at Urbana-Champaign. J. Wright and Y. Ma are with Microsoft Research Asia. Corresponding author: Andrew Wagner, awagner@illinois.edu, 1308 W. Main st. Urbana, IL 61801, (312) 343-1380.

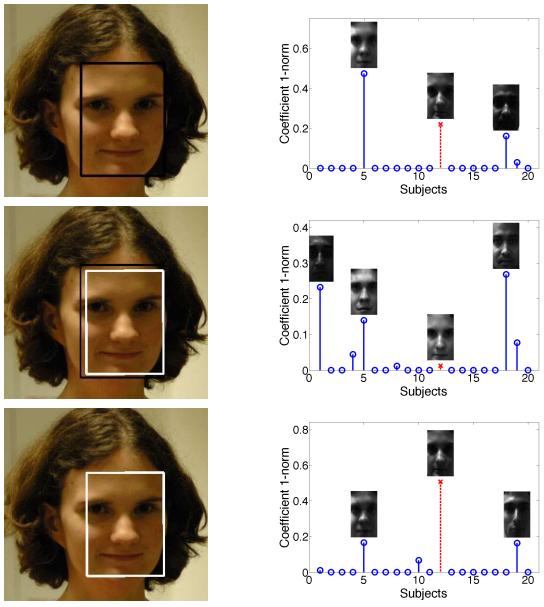


Fig. 1. **Effects of registration and illumination on Recognition**. In this example we identify the girl among 20 subjects, by computing the sparse representation of her input face with respect to the entire training set. The absolute sum of the coefficients associated with each subject is plotted on the right. We also show the faces reconstructed with each subject's training images weighted by the associated sparse coefficients. The red line (cross) corresponds to her true identity, subject 12. **Top:** The input face is from Viola and Jones' face detector (the black box) and all 38 illuminations specified in Section 3 are used in the training. **Middle:** The input face is well-aligned (the white box) with the training by our algorithm specified in Section 2 but only 24 frontal illuminations are used in the training for recognition (see Section 3). **Bottom:** The input face is well aligned and a sufficient set (all 38) of illuminations are used in the training. Both are necessary for correct recognition using SRC.

an example in Figure 1. The task is to identify the girl among 20 subjects. If the test face image, say obtained from an off-the-shelf face detector, has even a small amount of registration error against the training images (caused by mild pose, scale, or misalignment), the sparse representation obtained using the method of [3] is no longer informative, even if sufficient illuminations are present in the training, as shown in Figure 1(top). Additionally, in order to span the illuminations of a typical indoor (or outdoor) environment, illuminations from behind the subject are needed in the training set. Otherwise, even for perfectly aligned test images, the sparse representation obtained using [3] will not necessarily be sparse or informative, as shown by the example in Figure 1(middle). Clearly, both good alignment, as well as sufficient training images are needed to ensure success of the sparsity-based recognition method proposed by [3]. In this paper, we examine how to handle alignment and illumination simultaneously in the sparse representation framework, bringing the method proposed in [3] closer to practical use.

1.1 Related Work

We briefly review existing techniques for recognition, image registration, and handling of illumination variation. Our system is based purely on 2D techniques. This fact immediately distinguishes our approach from systems that either require a 3D data as an input, or attempt to estimate a 3D model from 2D input [4], [5]. While these techniques can achieve better robustness to pose variation given a sufficiently accurate 3D model, for access control applications where only moderate pose variation is present, the proposed method will be more than sufficient. Note that 2D images of faces under varying illuminations already contain 3D shape-related information,[1] and this information can be leveraged by 2D algorithms for alignment and recognition even if shape is not reconstructed explicitly.

In holistic recognition algorithms, correspondence between points in the test image and in the training must be achieved. A long line of research exists on using Active Appearance Models [6], and the closely related Active Shape Models [7] to register images against a relatively high-dimensional model of plausible face appearances, often leveraging face-specific contours. While these model-based techniques have advantages in dealing with variations in expression and pose, they may add unnecessary complexity to applications where subjects normally present a neutral face or only have moderate expression. We prefer to focus on deformations with far fewer degrees of freedom, i.e. similarity transformations, and to use the training images themselves as the appearance model. Iterative registration in this spirit dates at least back to the Lucas-Kanade algorithm [8].

Whereas much of the early work on image registration is aimed at the problem of registering nearly identical images, say by minimizing a sum of squared distances or maximizing normalized correlation, here we must confront several physical factors simultaneously: misalignment, illumination variations, and corrupted pixels. As we discuss further below, illumination variation can be dealt by expressing the test image as a linear combination of an appropriate set of training images. Similar representations have been exploited in illumination-robust tracking (e.g., [9], [10]). For robustness to gross errors, the ℓ^1-norm of the residual is a more appropriate objective function than the classical ℓ^2-norm. Its use here is loosely motivated by theoretical results due to Candès and Tao [11] (see also [12]). These two observations lead us to pose the registration problem as the search for a set of transformations and illumination coefficients that minimize the ℓ^1-norm of the representation error. We solve this problem using a generalized Gauss-Newton method which solves a sequence of affine-constrained ℓ^1-norm minimization problems [13], [14]. Each of these problems can also be solved efficiently using recently developed first-order techniques for ℓ^1-minimization, which are reviewed in [15].

Researchers have tried various techniques to deal with illumination variation. In almost all recognition algorithms where only a single gallery image is available per individual, illumination effects are regarded as a nuisance that must be removed before the algorithm can continue. This is typically done by making statistical assumptions about how illumination affects the image, and using those assumptions to extract a new representation that is claimed to be illumination invariant. Recent examples include [16] and [17]. However, despite these efforts, truly illumination-invariant features are difficult

1. In principle, one can recover the 3D shape of the face from multiple illuminations using photometric stereo.

to obtain from a single input image. We argue that if one has the luxury of designing the acquisition system and the application demands a high recognition rate, it is then unwise to limit the gallery to a single image per person. We therefore take the strategy of sampling many gallery images of each individual under varying illuminations. These images are used as the basis for either a convex cone model [18], [19], or a subspace model [20]. Images are captured using a simple-to-construct projector based light stage. While similar systems have been used for other applications, to our knowledge, we are the first to use projectors to indirectly illuminate a subject's face for the purpose of face recognition.

1.2 Contributions

In this paper, we show how registration and illumination can be simultaneously addressed within a robust sparse representation framework. We show that face registration, a challenging nonlinear problem, can be solved by a series of linear programs that iteratively minimize the sparsity of the registration error. This leads to an efficient and effective alignment algorithm for face images that works for a large range of variation in translation, rotation, and scale, even when the face is only partially visible due to eyeglasses, closed eyes and open mouth, sensor saturation, etc. We also propose a sufficient set of training illuminations that is capable of linearly representing typical indoor and outdoor lighting, along with a practical hardware system for capturing them.

We then demonstrate the effectiveness of the proposed new methods with a complete face recognition system that is *simple, stable, and scalable*. The proposed system performs robust automatic recognition of subjects from loosely controlled probe images taken both indoors and outdoors, using a gallery of frontal views of the subjects' faces under the proposed illuminations. An off-the-shelf face detector[2] is used to detect faces in the test images.

We conduct extensive experiments on the proposed system with both public databases and a face database that is collected by our own acquisition system. Our experimental results on large-scale public face databases show that our algorithm indeed achieves very good performance on these databases, exceeding or competing with the state-of-the-art algorithms. Additionally, our experimental results on our own database clearly demonstrate that our system not only works well with images taken under controlled laboratory conditions, but is capable of handling practical indoor and outdoor illuminations as well.

Organization of this paper: In Section 2, we derive our robust registration and recognition algorithm within the sparse representation framework. We elaborate on algorithmic implementation issues, conduct region of attraction experiments with respect to both 2D in-plane deformation and 3D pose variation, and discuss its relationship to existing work. Section 3 is dedicated to our training acquisition system. Using this system, we investigate empirically how many training illuminations are required to handle practical illumination variations, and suggest a sufficient set of 38 training illuminations. Extensive experiments on a large-scale public database and on our own database are conducted in Section 4 and Section 5, respectively, to verify the proposed system. Section 6 concludes our work with discussion of promising future directions.

2 Robust Alignment

As demonstrated in Figure 1(top), the main limitation of the *sparse representation and classification* (SRC) algorithm of [3] is the assumption of pixel-accurate alignment between the test image and the training set. This leads to brittleness under pose and misalignment, making it inappropriate for deployment outside a laboratory setting. In this section, we show how this weakness can be rectified while still preserving the conceptual simplicity and good recognition performance of SRC.

SRC assumes access to a database of multiple registered training images per subject, taken under varying illuminations. The images of subject i, stacked as vectors, form a matrix $A_i \in \mathbb{R}^{m \times n_i}$. Taken together, all of the images form a large matrix $A = [A_1 \mid A_2 \mid \cdots \mid A_K] \in \mathbb{R}^{m \times n}$. As argued in [3], a well-aligned test image $\boldsymbol{y}_0$ can be represented as a sparse linear combination $A\boldsymbol{x}_0$ of all of the images in the database,[3] plus a sparse error $\boldsymbol{e}_0$ due to corrupted pixels. The sparse representation can be recovered by minimizing the ℓ^1-norm[4] of $\boldsymbol{x}$ and $\boldsymbol{e}$:

$$\min_{\boldsymbol{x},\boldsymbol{e}} \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subj to} \quad \boldsymbol{y}_0 = A\boldsymbol{x} + \boldsymbol{e}. \tag{1}$$

Now suppose that $\boldsymbol{y}_0$ is subject to some pose or misalignment, so that instead of observing $\boldsymbol{y}_0$, we observe the warped image $\boldsymbol{y} = \boldsymbol{y}_0 \circ \tau^{-1}$, for some transformation $\tau \in T$ where T is a finite-dimensional group of transformations acting on the image domain. The transformed image $\boldsymbol{y}$ no longer has a sparse representation of the form $\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{e}_0$, and naively applying the algorithm of [3] is no longer appropriate, as seen in Figure 1(top).

2.1 Batch and Individual Alignment

If the true deformation τ^{-1} can be found, then we can apply its inverse τ to the test image and it again becomes possible to find a sparse representation of the resulting image, as $\boldsymbol{y} \circ \tau = A\boldsymbol{x}_0 + \boldsymbol{e}_0$.[5] This sparsity provides a strong cue for finding the correct deformation τ: conceptually, one would like to seek a transformation τ that allows the sparsest representation, via

$$\hat{\tau} = \arg\min_{\boldsymbol{x},\boldsymbol{e},\tau \in T} \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subj to} \quad \boldsymbol{y} \circ \tau = A\boldsymbol{x} + \boldsymbol{e}. \tag{2}$$

For fixed τ, this problem is jointly convex in $\boldsymbol{x}$ and $\boldsymbol{e}$. However, as a simultaneous optimization over the coefficients $\boldsymbol{x}$, error representation $\boldsymbol{e}$, and transformation τ, it is a difficult, nonconvex optimization problem. One source of difficulty is the presence of multiple faces in the matrix A: (2) has many local minima that correspond to aligning $\boldsymbol{y}$ to different subjects.

2. We use the OpenCV implementation of the Viola and Jones' face detector [21].

3. We assume the training illuminations are sufficient. We will address how to ensure illumination sufficiency in the next section.

4. The ℓ^1-norm of a vector, denoted by $\|\cdot\|_1$, is the sum of absolute values of its entries.

5. In the terminology of [22], this formulation is "Forward Additive".

In this sense, the misaligned recognition problem differs from the well-aligned version studied in [3]. For the well-aligned case, it is possible to directly solve for a global representation, with no concern for local minima. With possible misalignment, it is more appropriate to seek the best alignment of the test face with each subject i:

$$\hat{\tau}_i = \arg\min_{\boldsymbol{x},\boldsymbol{e},\tau_i \in T} \|\boldsymbol{e}\|_1 \quad \text{subj to} \quad \boldsymbol{y} \circ \tau_i = A_i \boldsymbol{x} + \boldsymbol{e}. \tag{3}$$

We no longer penalize $\|\boldsymbol{x}\|_1$, since A_i consists of only images of subject i and so $\boldsymbol{x}$ is no longer expected to be sparse.

2.2 Alignment via Sequential ℓ^1-Minimization

While the problem (3) is still nonconvex, for cases of practical interest in face recognition, a good initial guess for the transformation is available, e.g., from the output of a face detector. We can refine this initialization to an estimate of the true transformation by repeatedly linearizing about the current estimate of τ, and seeking representations of the form:

$$\boldsymbol{y} \circ \tau + J\Delta\tau = A_i \boldsymbol{x} + \boldsymbol{e}. \tag{4}$$

Here, $J = \frac{\partial}{\partial \tau} \boldsymbol{y} \circ \tau$ is the Jacobian of $\boldsymbol{y} \circ \tau$ with respect to the transformation parameters τ, and $\Delta\tau$ is the step in τ. The above equation is underdetermined if we allow the registration error $\boldsymbol{e}$ to be arbitrary. At the correct alignment we expect the test image to differ from $A_i \boldsymbol{x}$ only for the minority of the pixels corrupted by occlusions. Thus, we seek a deformation step $\Delta\tau$ that best sparsifies the registration error $\boldsymbol{e}$, in terms of its ℓ^1-norm:

$$\Delta\hat{\tau}_1 = \arg\min_{\boldsymbol{x},\boldsymbol{e},\Delta\tau \in T} \|\boldsymbol{e}\|_1 \quad \text{subj to} \quad \boldsymbol{y} \circ \tau + J\Delta\tau = A_i \boldsymbol{x} + \boldsymbol{e}. \tag{5}$$

This is different from the popular choice that minimizes the ℓ^2-norm of the registration error:

$$\Delta\hat{\tau}_2 = \arg\min_{\boldsymbol{x},\boldsymbol{e},\Delta\tau \in T} \|\boldsymbol{e}\|_2 \quad \text{subj to} \quad \boldsymbol{y} \circ \tau + J\Delta\tau = A_i \boldsymbol{x} + \boldsymbol{e}, \tag{6}$$

which is also equivalent to finding the deformation step $\Delta\tau$ by solving the least-square problem: $\min_{\boldsymbol{x},\Delta\tau} \|\boldsymbol{y} \circ \tau + J\Delta\tau - A_i \boldsymbol{x}\|_2$. Empirically, we find that if there is only small noise between $\boldsymbol{y}_0$ and $A_i \boldsymbol{x}$, both (5) and (6) have similar performance. However, if there are occlusions in $\boldsymbol{y}_0$, sequential ℓ^1-minimization (5) is significantly better than sequential ℓ^2-minimization (6). Figure 2 shows an example.

The scheme (5) can be viewed 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. Such algorithms date at least back to the 1970's [23], [14], and continue to attract attention today [24]. While space precludes a detailed discussion of their properties, we should mention that the scheme (5) is known to converge quadratically in the neighborhood of any local optimum of the ℓ^1-norm. In practice, this means that ≈ 10 to 15 iterations suffice to reach the desired solution. We refer the interested reader to [14], [13] and the references therein.

In addition to normalizing the training images (which is done once), it is important to normalize the warped testing image $\boldsymbol{y} \circ \tau$ as the algorithm runs. Without normalization, the algorithm may fall into a degenerate global minimum corresponding to zooming in on a dark region of the test image. Normalization is done by replacing the linearization of $\boldsymbol{y} \circ \tau$ with a linearization of the normalized version $\tilde{\boldsymbol{y}}(\tau) = \frac{\boldsymbol{y} \circ \tau}{\|\boldsymbol{y} \circ \tau\|_2}$. The proposed alignment algorithm can be easily extended to work in a *multiscale* fashion, with benefits both in convergence behavior and computational cost. The alignment algorithm is simply run to completion on progressively less downsampled versions of the training and testing images, using the result of one level to initialize the next.

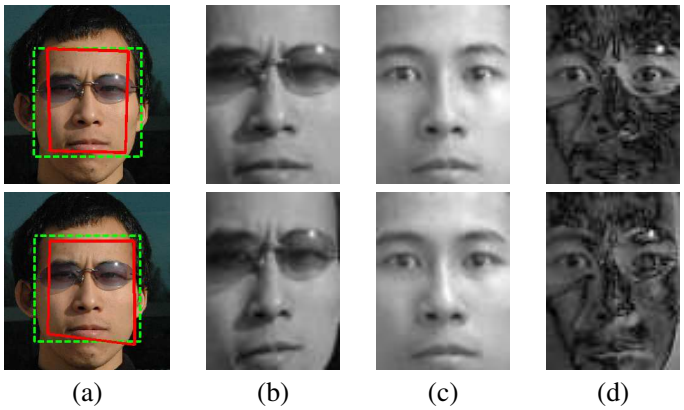


Fig. 2. **Comparing alignment of a subject wearing sunglasses by ℓ^1 and ℓ^2 minimization. Top:** alignment result of minimizing $\|\boldsymbol{e}\|_1$; **Bottom:** result of minimizing $\|\boldsymbol{e}\|_2$. (a) *Green (dotted):* Initial face boundary given by the face detector, *Red (solid):* Alignment result shown on the same face; (b) warped testing image using the estimated transformation $\boldsymbol{y}_0$; (c) reconstructed face $A_i \boldsymbol{x}$ using the training; (d) image of error $\boldsymbol{e}$.

2.3 Robust Recognition by Sparse Representation

Once the best transformation τ_i has been computed for each subject i, the training sets A_i can be aligned to $\boldsymbol{y}$, and a global sparse representation problem of the form (1) can be solved to obtain a discriminative representation in terms of the entire training set. Moreover, the per-subject alignment residuals $\|\boldsymbol{e}\|_1$ can be used to prune unpromising candidates from the global optimization, leaving a much smaller and more efficiently solvable problem. The complete optimization procedure is summarized as Algorithm 1. The parameter S in our algorithm is the number of subjects considered together to provide a sparse representation for the test image. If $S = 1$, the algorithm reduces to classification by registration error; but considering the test image might be an invalid subject, we typically choose $S = 10$. Since valid images have a sparse representation in terms of this larger set, we can reject invalid test images using the *sparsity concentration index* proposed in [3]. The function $\delta_i(\boldsymbol{x})$ in Algorithm 1 selects coefficients from the vector $\boldsymbol{x}$ corresponding to subject i.

Another important free parameter in Algorithm 1 is the class of deformations T. In our experiments, we typically use 2D similarity transformations, $T = \mathbb{SE}(2) \times \mathbb{R}_+$[6], for removing alignment error incurred by face detector, or 2D

6. Here, SE stands for Special Euclidean, i.e., 2D rigid transformations. The $\mathbb{R}_+$ accounts for the scale.

projective transformations, $T = \mathbb{GL}(3)$[7], for handling some pose variation.

In Algorithm 1, we also implement a simple heuristic which improves the performance of our system, based on the observation that the face detector output may be poorly centered on the face, and may contain a significant amount of the background. Therefore, before the recognition stage, instead of aligning the training sets to the original $\boldsymbol{y}$ directly obtained from the face detector, we compute an average transformation $\bar{\tau}$ from $\tau_{k_1}, \tau_{k_2}, \ldots, \tau_{k_S}$ of the top S classes, which is believed to be better centered, and update $\boldsymbol{y}$ according to $\bar{\tau}$. For the 2D similarity transformations, which are used in our system when initialized by the face detector, a transformation τ can be parameterized as $\tau = (\tau^1, \tau^2, \tau^3, \tau^4)$, where τ^1 and τ^2 represent the translations in x- and y-axis, τ^3 represents the rotation angle and τ^4 represents the scale. Then the average transformation is simply obtained by taking the componentwise mean:

$$\bar{\tau}^i = (\tau^i_{k_1} + \tau^i_{k_2} + \cdots + \tau^i_{k_S})/S, i = 1, 2, 3, 4.$$

Finally, the training sets are aligned to the new $\boldsymbol{y}$.

Algorithm 1 (Deformable Sparse Recovery and Classification for Face Recognition)

Input: Training images $\{A_i \in \mathbb{R}^{m \times n_i}\}_{i=1}^K$ for K subjects, a test image $\boldsymbol{y} \in \mathbb{R}^m$ and a deformation group T.
for each subject i,
  $\tau^{(0)} \leftarrow I$.
  while not converged $(j = 1, 2, \ldots)$ **do**
    $\tilde{\boldsymbol{y}}(\tau) \leftarrow \frac{\boldsymbol{y} \circ \tau}{\|\boldsymbol{y} \circ \tau\|_2}$; $J \leftarrow \frac{\partial}{\partial \tau} \tilde{\boldsymbol{y}}(\tau)\big|_{\tau^{(j)}}$;
    $\Delta\tau = \arg\min \|\boldsymbol{e}\|_1$ subj to $\tilde{\boldsymbol{y}} + J\Delta\tau = A_i\boldsymbol{x} + \boldsymbol{e}$.
    $\tau^{(j+1)} \leftarrow \tau^{(j)} + \Delta\tau$;
  end while
end
Keep the top S candidates $k_1, \ldots, k_S$ with the smallest residuals $\|\boldsymbol{e}\|_1$.
Compute an average transformation $\bar{\tau}$ from $\tau_{k_1}, \tau_{k_2}, \ldots, \tau_{k_S}$.
Update $\boldsymbol{y} \leftarrow \boldsymbol{y} \circ \bar{\tau}$ and $\tau_i \leftarrow \tau_i \cdot \bar{\tau}^{-1}$ for $i = k_1, \ldots, k_S$.
Set $A \leftarrow \left[A_{k_1} \circ \tau_{k_1}^{-1} \mid A_{k_2} \circ \tau_{k_2}^{-1} \mid \cdots \mid A_{k_S} \circ \tau_{k_S}^{-1}\right]$.
Solve the ℓ^1-minimization problem: $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x},\boldsymbol{e}} \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1$ subj to $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$.
Compute residuals $r_i(\boldsymbol{y}) = \|\boldsymbol{y} - A_i\,\delta_i(\hat{\boldsymbol{x}})\|_2$ for $i = k_1, \ldots, k_S$.
Output: identity$(\boldsymbol{y}) = \arg\min_i r_i(\boldsymbol{y})$.

The transformation τ defines a mapping between the coordinates of pixels in the large original image and a smaller (un)warped image. The pixels of the small image are stacked into a vector. To prevent aliasing artifacts in the downsampled image, one should apply a smoothing filter to the original image. For a simple implementation, a rectangular window with regular sampling can used, but in general, the small image need not be regularly sampled in pixel coordinates. For example, the sample locations could be arbitrarily selected from within a "face shaped" area. We will discuss how choosing different windows can affect the performance of our algorithm in Section 4.

7. Here, GL stands for General Linear. This class of transformations is able to represent distortion in a perspective image of a planar object.

2.4 System Implementation

The runtime of Algorithm 1 is dominated by the time spent solving two qualitatively similar ℓ_1 minimization problems. We have developed custom solvers for this purpose based on *Augmented Lagrange Multiplier* (ALM) algorithm. We have selected this algorithm because it strikes the best balance between speed, accuracy, and scalability for our problem out of many algorithms that we have tested. We refer the reader to our supplementary materials for a more in-depth discussion of our solvers. For a more detailed discussion of competing approaches, we refer the interested reader to [15]. On a Mac Pro with Dual-Core 2.66GHz Xeon processors and 4GB memory, running on our database containing images size 80×60 pixels from 109 subjects under 38 illuminations, our C implementation of Algorithm 1 takes about 0.60 seconds per subject for alignment and about 2.0 seconds for global recognition. Compared to the highly customized interior point method used in the conference version of this paper [25], this new algorithm is only slightly faster for per subject alignment. However, it is much simpler to implement and it achieves a *speedup of more than a factor of 10* for global recognition!

2.5 Experiments on Region of Attraction

We will now present three experimental results demonstrating the effectiveness of the individual alignment procedure outlined in the previous section. They show the sufficiency of the region of attraction, verify effectiveness of the multiscale extension, and show stability to small pose variations. We delay large-scale recognition experiments to Sections 4 and 5, after we have discussed the issue of illumination in the next section.

1) *2D Deformation.* We first verify the effectiveness of our alignment algorithm with images from the CMU Multi-PIE Database [26]. We select all the subjects in Session 1, use 7 illuminations per person from Session 1 for training, and test on one new illumination from Session 2.[8] We manually select eye corners in both training and testing as the ground truth for registration. We downsample the images to 80×60 pixels[9] and the distance between the two outer eye corners is normalized to be 50 pixels for each person. We introduce artificial deformation to the testing image with a combination of translation, rotation and scaling. We further use the alignment error $\|\boldsymbol{e}\|_1$ as an indicator of success. Let r_0 be the alignment error obtained by aligning a test image to the training images without any artificial perturbation. When the test image is artificially perturbed and aligned, resulting in an alignment error r, we consider the alignment successful if $|r - r_0| \leq 0.01 r_0$. Figure 3 shows the percentage of successful registrations for all subjects for each artificial deformation. The results suggest that our algorithm works extremely well with translation up to 20% of the eye distance (or 10 pixels) in all 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.

8. The training are illuminations $\{0, 1, 7, 13, 14, 16, 18\}$ of [26], and the testing is the illumination 10.

9. Unless otherwise stated, this will be the default resolution at which we prepare all our training and testing datasets.

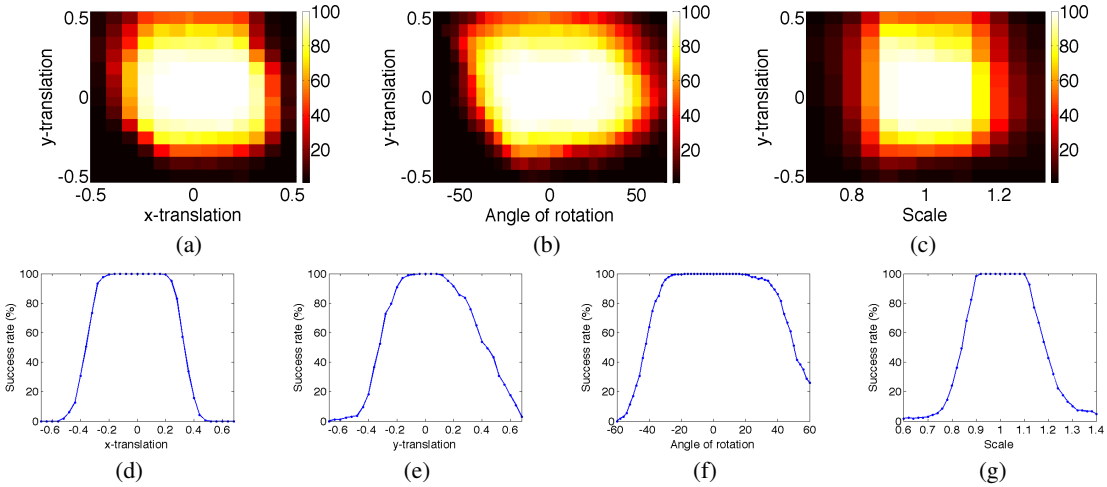


Fig. 3. **Region of attraction.** Fraction of subjects for which the algorithm successfully aligns a synthetically perturbed test image. The amount of translation is expressed as a fraction of the distance between the outer eye corners, and the amount of in-plane rotation in degrees. **Top row:** (a) Simultaneous translation in x and y directions. (b) Simultaneous translation in y direction and in-plane rotation. (c) Simultaneous translation in y direction and scale variation. **Bottom row:** (d) Translation in x direction only. (e) Translation in y direction only. (f) In-plane rotation only. (g) Scale variation only.

We have gathered the statistics of the Viola and Jones' face detector on the Multi-PIE dataset. For 4,600 frontal images of 230 subjects under 20 different illuminations, using manual registration as the ground truth, the average misalignment error of the detected faces is about 6 pixels and the average variation in scale is 8%. This falls safely inside the region of attraction for our alignment algorithm.

2) *Multiscale Implementation.* Performing alignment in a multiscale fashion has two benefits: first, it provides a larger region of attraction, and second, it reduces overall computational cost. Here, we further investigate the convergence behavior of the algorithm as a function of the standard deviation σ of the Gaussian smoothing filter and the number of scales considered. We use the same 7 illuminations in Session 1 as training, and all 20 illuminations in the same session as testing. We introduce artificial deformation in both x and y directions up to 16 pixels in the 80×60 frame, with a step size of 4 pixels, i.e., $(\Delta x, \Delta y) \in \{-16, -12, \ldots, 12, 16\} \times \{-16, -12, \ldots, 12, 16\}$. We consider an alignment successful if the estimated coordinates of the eye-corners are within 1 pixel from the ground truth in the original image. In Figure 4, we report the alignment success rate, averaged over the artificially perturbed initial deformations, as a function of the standard deviation of the Gaussian kernel σ, for three choices of the number of scales. As one can see, using multiscale indeed improves the performance, and when 3 scales are used, a smaller convolution kernel can achieve a similar performance compared to a much larger kernel when only 2 scales are used.

3) *3D Pose Variation.* As densely sampled pose and illumination face images are not available in any of the public databases, including Multi-PIE, we have collected our own dataset using our own system (to be introduced in the next section). We use frontal face images of a subject under the 38 illuminations proposed in the next section as training. For testing, we collect images of the subject under a typical indoor

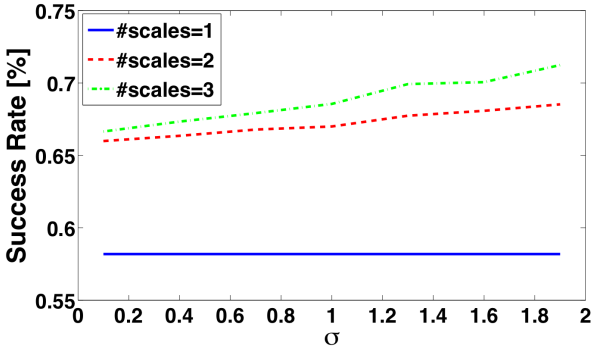


Fig. 4. **Multiscale alignment.** This figure shows the average success rate of alignment over all possible perturbations. A smaller blur kernel can be applied to achieve certain level of performance when more scales are used.

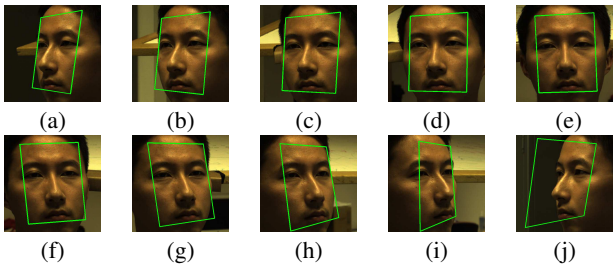


Fig. 5. **2D Alignment of test images with different poses to frontal training images. (a) to (i):** plausible alignment for pose from $-45°$ to $+45°$. **(j):** a case when the algorithm fails for an extreme pose ($> 45°$).

lighting condition at pose ranging from $-90°$ to $+90°$ with step size $5.625°$, a total of 33 poses. We use Viola and Jones' face detector to initialize our alignment algorithm. Figure 5 shows that our algorithm works reasonably well with poses up to $\pm 45°$. Note that this level of out-of-plane pose variation is beyond what we intend to handle with our formulation.

2.6 Comparison with Related Work

Our modification to SRC roots solidly in the tradition of adding deformation-robustness to face recognition algorithms

[6], [27], [28]. However, the only previous work to investigate face alignment in the context of sparse signal representation and SRC is the work of [29]. They consider the case where the training images themselves are misaligned and allow one deformation per training image. They linearize the training rather than the test, which is computationally more costly as it effectively triples the size of the training set. In addition, as they align the test image to all subjects simultaneously, it potentially is more prone to local minima when the number of subjects increases, as we will see in the following experimental comparisons.

1) *Extended Yale B.* In this experiment, we have used the same experimental settings as in [29]. 20 subjects are selected and each has 32 frontal images (selected at random) as training and another 32 for testing. An artificial translation of 10 pixels (in both x and y directions) is introduced to the test image. For our algorithm we downsample all the images to 88×80 for memory reasons, whereas the work of [29] uses random projections. Note that the use of cropped images in this experiment introduces image boundary effects. Our algorithm achieves the recognition rate 93.7%, compared to 89.1% recognition rate reported in [29].
2) *CMU Multi-PIE.* In this experiment, we choose all subjects from the CMU Multi-PIE database, 7 training images from Session 1 and 1 test image from Session 2 per person. The setting is exactly the same as the previous experiment on 2D deformation. We again work with downsampled images of size 80×60 pixels. An artificial translation of 5 pixels (in both x and y directions) was induced in the test image. The algorithm of [29] achieves a recognition rate of 67.5%,[10] while ours achieves 92.2%.

3 Handling Illumination Variation

In the above section, we have made the assumption that the test image, although taken under some arbitrary illumination, can be linearly represented by a finite number of training illuminations. Under what conditions is this a reasonable assumption to make? What can we say from first principles about how the training images should be chosen?

3.1 The Illumination Model

The strongest theoretical results so far regarding the relationship between illumination and the resulting sets of images is due to Basri and Jacobs [20]. The main result of this paper is that for convex Lambertian objects, distant illuminations, and fixed pose, all images of the object can be well approximated by linear combinations of nine (properly chosen) basis images. The basis images have mixed sign, and their illuminations consist of the lowest frequency spherical harmonics. While this is a very important result for understanding the image formation process, the direct application of this result in most practical systems is misguided for several reasons. Specularities, self-shadowing, and inter-reflections all dramatically affect the appearance of face images, and they all do so in a way that violates the modeling assumptions of the Basri analysis.

Fortunately, even with these effects, for most materials the relationship between illumination and image is still linear,[11] provided the sensor has a linear response curve.[12] For a more in-depth study of the relationship between illumination and images, we refer the reader to [19]. While the relationship between illuminations and images is linear, only positive weights are allowed; the space of all images of an object with fixed pose and varying illumination is a convex cone lying in the positive orthant. The question becomes, how many images does it take to do a good job of representing images sampled from this cone?

It has been observed in various empirical studies that one can get away with using a small number of frontal illuminations to linearly represent a wide range of new frontal illuminations, when they are all taken under the same laboratory conditions [18]. This is the case for many public face datasets, including AR, ORL, PIE, and Multi-PIE. Unfortunately, we have found that in practice, a training database consisting purely of frontal illuminations is not sufficient to linearly represent images of a faces taken under typical indoor or outdoor conditions (see the experiment conducted in Section 5). As illustrated by the example in Figure 1, an insufficient number of training illuminations can result in recognition failure. To ensure our algorithm works in practice, we need to find a set of training illuminations that are indeed *sufficient* to linearly represent a wide variety of practical indoor and outdoor illuminations.

3.2 Projector-based Illumination System

We have designed a system that can acquire frontal images of a subject while simultaneously illuminating the subject from all directions above horizontal. A sketch of the system is shown in Figure 6: The illumination system consists of four projectors that display various bright patterns onto the three white walls in the corner of a dark room. The light reflects off of the walls and illuminates the user's head indirectly. After taking the frontal illuminations we rotate the chair by 180 degrees and take pictures from the opposite direction. Having two cameras speeds the process since only the chair needs to be moved in between frontal and rear illuminations. Our projector-based system has several advantages over flash-based illumination systems for face recognition:

- The illuminations can be modified in software, rather than hardware.
- It is easy to capture many different illuminations quickly.

10. That algorithm has two free parameters - l and d, which govern the tradeoff between accuracy and run-time. For this experiment we chose $l = 1$ and $d = 514$.

11. Materials that break this assumption include fluorescent materials and the photochromic ("Transition") lenses in some eyeglasses. Most materials emit light in proportion to their incident light.

12. Proper handling of gamma encoding is an important consideration for practitioners. Most cameras apply a non-linear and often undocumented response curve to captured images. A slight degradation of performance will occur if gamma compressed images are treated as if they were linear. We recommend the use of cameras with well documented response curves that can be inverted when the image file is loaded.

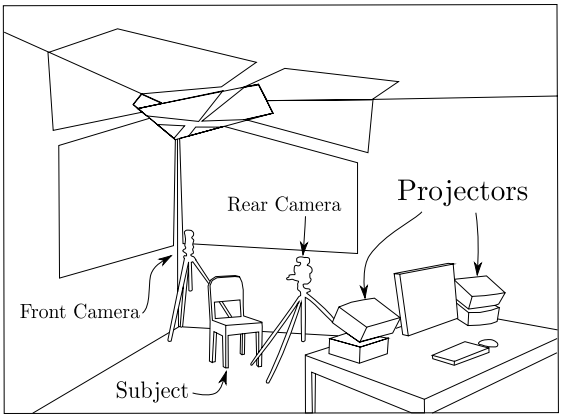


Fig. 6. **Training acquisition system:** Four projectors and two cameras controlled by one computer.

- Good coverage and distant illumination can be achieved simultaneously.
- There is no need to mount anything on the walls or construct a large dome.
- The system can be assembled from off-the-shelf hardware.

With our projector system, our choice of illuminations is constrained only by the need to achieve a good SNR,[13] avoid saturation, and achieve a reasonably short acquisition time. Two simplifying assumptions that we make are that every pixel is either turned fully on or off in every illumination, and that the illuminated regions do not overlap.

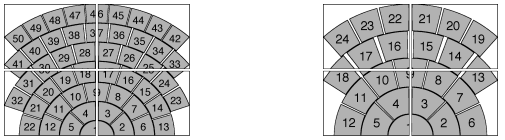
Assuming that each pixel is fully on or off enables us to guarantee that each illumination image has the same overall intensity, merely by guaranteeing that we illuminate the same number of pixels in each image.[14] Since our algorithm depends only on the linearity between the illuminations and the images, and not on the relative intensities of the illuminations, the designer has the freedom to choose the overall intensity of the illuminations to prevent saturation or low SNR, in a sort of offline exposure control.

Assuming that the sequentially illuminated regions do not overlap results in a set of training images that span a larger cone than a similar number of overlapping regions. This results in training images that require fewer negative coefficients in $\boldsymbol{x}$ to represent test images under natural illuminations. The effect of negative coefficients in $\boldsymbol{x}$ appears to depend partly on how the test images are taken and is still under study.

Relationship to existing work: Most light stages used for face recognition have been constructed for the purpose of creating public data sets to study illumination invariance [18], [26]. Many other light stages have been used for computer graphics purposes [30], [31]. The light source can be moved around manually [32], but this may result in poor consistency of illuminations between users. Structured light applications use projectors to directly illuminate the face (or other object) [33] for 3D reconstruction, but this is very disturbing to the user. Y. Schechner [34] studies techniques for multiplexing illumination that can dramatically reduce the noise of the demultiplexed images for certain classes of objects and cameras. While these techniques have not been incorporated into the current system, they fit elegantly into our framework and will likely be used in future implementations. We stress that use of this multiplexing technique is independent from the choice of original (directional) illuminations.

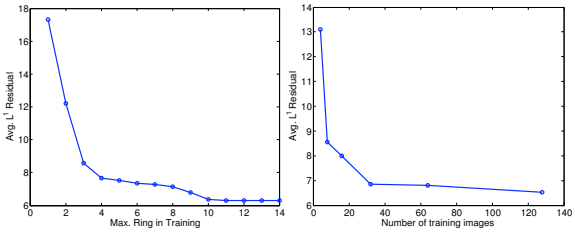
13. Since illuminations with more pixels illuminated will have a better SNR (provided they don't saturate), there is an engineering tradeoff between the SNR and the number of training images.

14. Since DLP projectors may have dramatically different response curves depending on the mode they are in, it is not advisable to simply normalize each illumination image by its mean.



(a) Coverage Experiment (b) Chosen Illumination Patterns

Fig. 7. **Illumination Patterns.** The cells are illuminated in sequence. For rear illuminations the sequence is reversed. In the chosen pattern's rear illumination, the cells 1-5 and 7-11 are omitted for a total of 38 illuminations. The four rectangular regions correspond to the four projectors.



(a) Coverage (b) Granularity

Fig. 8. **Study of sufficient illuminations.** The average ℓ^1 registration residual versus different illumination training sets.

3.3 Choice of Illumination Patterns

We ran two experiments to guide our choice of illuminations for our large-scale experiments:

1) *Coverage Experiment.* In the first experiment we attempt to determine what coverage of the sphere is required to achieve good interpolation for test images. The subject was illuminated by 100 (50 front, 50 back) illuminations arranged in concentric rings centered at the front camera. Subsets of the training images were chosen, starting at the front camera and adding a ring at a time. Each time a ring was added to the training illumination set, the average ℓ^1 registration error (residual) for a set of test images taken under sunlight was computed and plotted in Figure 7(a). The more rings of training illuminations are added, the lower the representation error becomes, with diminishing returns.
2) *Granularity Experiment.* In the second experiment we attempt to determine how finely divided the illumination sphere should be. At the first granularity level, the projectors illuminate the covered area uniformly. At each subsequent granularity level each illuminated cell is divided in two along its longer side but intensity doubled. For each granularity level the average ℓ^1 registration error is computed as in the coverage experiment and shown in Figure 8(b). Again, diminishing returns are observed as more illuminations are added.

In the plot for the coverage experiment, Figure 8(a), we clearly see two plateau regions: one is after 4 rings and one is after 10 rings. The first four rings represent the typical frontal illuminations, which are present in most public face datasets; however, we see that the residual stabilizes after 10 rings which include some illuminations from the back of the subject. This suggests that although the frontal illuminations account for most of the illumination on the face, some illuminations from the back are needed in the training set to represent images with illumination coming from all directions. In the plot for the granularity experiment, Figure 8(b), we observe that the residual reaches a plateau after four divisions, corresponding to a total of 32 illuminations. Based on the results from both experiments, we decide to partition the area covered by the first 10 rings into a total of 38 cells, whose layout is explained in Figure 7(b). For our large-scale experiments, we have collected those illuminations for all our subjects.[15]

See below for the 38 training images of one subject:

4 Tests on Public Databases

In this section and the next section, we conduct comprehensive experiments on large-scale face databases to verify the performance of our algorithm and system. We first test on the largest public face database available that is suitable for testing our algorithm, the CMU Multi-PIE. One shortcoming of the CMU Multi-PIE database for our purposes is that there is no separate set of test images taken under natural illuminations; we are left to choose which sets of images to use for testing and training. To challenge our algorithm, we choose only a small set of illuminations for the training set, yet we include all illuminations in the testing set. In the following section, we will test our algorithm on a face dataset that is collected by our own system. The goal for that experiment will be to show that with a sufficient set of training illuminations for each subject, our algorithm indeed works stably and robustly with practical illumination, misalignment, pose, and occlusion, as already indicated by our experiment shown in Figure 1(bottom).

CMU Multi-PIE provides the most extensive test set among public datasets. This database contains images of 337 subjects across simultaneous variation in pose, expression, and illumination. Of these 337 subjects, we use all of the 249 subjects present in Session 1 as the training set. The remaining 88 subjects are treated as "impostors", or invalid images. For each of the 249 training subjects, we include frontal images of 7 frontal illuminations,[16] taken with neutral expression. As suggested by the work of [18], we choose these extreme frontal illuminations in the hope that they would linearly represent other frontal illuminations well. For the test set, we use all 20 illuminations from Sessions 2-4, which were recorded over a period of several months. The dataset is challenging due to the large number of subjects, and due to natural variation in subject appearance over time. Table 1 shows the result of our algorithm on each of the 3 testing sessions. Our algorithm achieves recognition rates above 90% for all three sessions. For the test images, our iterative alignment was initialized automatically via the Viola and Jones' face detector. To demonstrate that the sparse representation based recognition step is indeed beneficial even when there are no impostors, we include results for recognition based only on the alignment error residuals (i.e. $S = 1$), shown in row 1.

15. It is possible that with further experimentation a reduced set of illuminations can be found that performs as well or better.

16. They are illuminations $\{0, 1, 7, 13, 14, 16, 18\}$ of [26]. For each directional illumination, we subtract the ambient-illuminated image 0.

TABLE 1
Recognition rates on the Multi-PIE database for Algorithm 1 and [35]

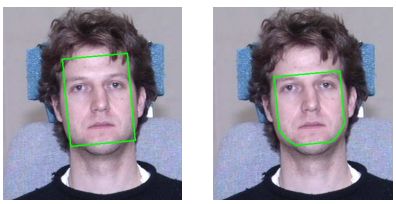
Recognition rate	Session 2	Session 3	Session 4
Alg. 1, $S = 1$	90.7%	89.6%	87.5%
Alg. 1	93.9%	93.8%	92.3%
Alg. 1 with improved window	95.0%	**96.3**%	**97.3**%
[35]	**95.2**%	93.4%	95.1%

(a) (b) (c) (d) (e) (f)

Fig. 9. **Representative failures from Multi-PIE**. **Top:** training from Session 1; **Bottom:** test images from Session 2. Due to changes in hair, glasses, beard, or pose, our alignment fails on these subjects regardless of test image illumination.

4.1 Improving the Sampling Window

Our algorithm's errors are mostly caused by a few subjects who significantly change their appearances between sessions (such as hair, facial hair, and eyeglasses). Some representative examples are shown in Figure 9. For those subjects, alignment and recognition fail on almost all test illuminations. Meanwhile, this observation also suggests that we might be able to improve the performance of our method by carefully choosing a face region which is less affected by the above factors for recognition. In particular, since the forehead region is likely to be affected by the change of hair style, we try replacing the previous 80×60 canonical frame with a new window that better excludes the forehead. We adjust the resolution of the window to keep m approximately constant. In addition, we cut off two lower corners of the 80×60 canonical frame, motivated by the observation that in many cases the



Default window. Proposed window.

Fig. 10. **Choosing different sampling windows.**

TABLE 2
Recognition rates on the Multi-PIE database for different pairings of alignment and recognition stages.

Rec. \ Align.	Face Detector			Manual			Iterative Alignment		
Session →	2	3	4	2	3	4	2	3	4
NS	30.8%	29.4%	24.6%	77.6%	74.3%	73.4%	84.5%	82.3%	81.4%
NN	26.4%	24.7%	21.9%	67.3%	66.2%	62.8%	73.5%	69.6%	69.3%
LDA	5.1%	5.9%	4.3%	49.4%	44.3%	47.9%	91.0%	89.9%	88.1%
LBP	39.9%	38.1%	33.9%	93.3%	91.2%	92.9%	**95.2%**	**94.7%**	**93.5%**
SRC	–	–	–	–	–	–	93.9%	93.8%	92.3%

corners actually contain background. An example of the new window is shown in Figure 10.

Table 1 shows that the recognition rates on Multi-PIE indeed increase with this new window. In addition, Figure 9(a), (b), and (c) illustrate three representative subjects for which the recognition rates of our algorithm are significantly boosted with the new window. However, we should mention that the best choice of the window is problem-specific and there is not a simple guideline to follow. For example, although the new window performs better on Multi-PIE, the same window does not help at all on our own database, which will be introduced in the next section. This is because most of the training and testing images in our database are taken on the same day so the variation in hair style is very small. Hence, excluding the forehead part may actually result in loss of useful discriminative information.

4.2 Comparison to Existing Work

We first compare our result to the recent work [35]. Notice that in [35], the initial registration is obtained from manually selected outer eye corners. Then, a supervised hierarchical sparse coding model based on local image descriptors is trained, which enjoys certain translation invariant properties. With the same training and testing sets, [35] is able to handle the remaining misalignment and achieves state-of-the-art performance on the CMU Multi-PIE database. Table 1 shows that our algorithm achieves similar or better performance on different sessions of Multi-PIE.

To better examine the effectiveness of our iterative alignment algorithm, we next compare our result to baseline linear-projection-based algorithms, such as Nearest Neighbor (NN), Nearest Subspace (NS) [36], and Linear Discriminant Analysis (LDA) [2].[17] Since these algorithms assume pixel-accurate alignment, they are not expected to work well if the test image is not well aligned with the training. In Table 2, we report the results of these classical algorithms with three types of testing image alignment: 1. alignment from the Viola and Jones' detector, 2. alignment via manually selected outer eye corners,[18] and 3. the output of our iterative alignment algorithm. The performance drop of the LDA algorithm on Multi-PIE reported here seems to agree with that reported already in [26]. All of the classical algorithms benefit greatly from being paired with our iterative alignment algorithm.

17. We do not list results on PCA [1] as its performance is always below that of Nearest Subspace.

18. Two manually clicked points are sufficient to define a similarity transformation. All of the experiments in this section are carried out with similarity transformations.

We also compare our result to Local Binary Patterns (LBP) [37], a local appearance descriptor which is able to capture fine details of facial appearance and texture. Due to its robustness to variations in illumination, facial expression, aging and other changes, LBP has achieved the state-of-the-art face recognition performance in the scenario when only one sample per person is used for training [38]. In this paper, we follow the same steps as in [37] to construct an LBP descriptor for each training and testing sample. The 80×60 face region is first divided into a regular 10×10 grid of cells, each of size 8×6 pixels. Within each cell, the histogram of 59 uniform binary patterns is then computed, where the patterns are generated by thresholding 8 neighboring pixels in a circle of radius 2 using the central pixel value. Finally, the local histograms are concatenated to produce the global descriptor vector. As suggested in [37], the recognition is performed using a nearest neighbor classifier with Chi square distance as the distance measure and we report the recognition rates with the same three types of input as before.

As shown in Table 2, although LBP achieves competitive recognition rates given manually aligned training and testing samples, demonstrating its robustness to moderate misalignment, it still benefits from using the output of our iterative alignment algorithm as the input. In addition, like the other classical algorithms, the performance of LBP degrades dramatically if it is applied directly to the output of a face detector. This is notable given that LBP is often applied without any special alignment in practice. Finally, we attribute the improvement in performance of LBP over SRC in this experiment to its robustness to illumination components that cannot be linearly interpolated by the training set.

In addition, although our algorithm is not designed for recognition when there is only a single gallery image per user, we compare its performance with LBP within this setting for completeness. For this experiment, we use the FERET dataset [39], which contains five standard partitions: 'fa' is the gallery containing 1196 frontal images of 1196 subjects, and 'fb', 'fc', 'dup1' and 'dup2' are four sets of probe images. The testing sets differ from the training in facial expression ('fb'), illumination ('fc'), aging ('dup1') and long aging ('dup2'). In fact, except for 'fb', we notice significant changes of illumination in all the other three test sets. For the training, we again crop and normalize the face region from each original image to an 80×60 window using manually marked eye coordinates [40]. In Table 3, we report the performance of our algorithm on the four test sets, with input directly obtained from the Viola and Jones' detector. We also report the performance of LBP with

TABLE 3
Performance on single gallery image FERET dataset

Recognition rate %	fb	fc	dup1	dup2
LBP_d	54.8	10.3	29.8	19.8
LBP_m	**96.6**	**58.8**	**71.6**	**61.5**
LBP_i	94.5	42.8	46.5	21.1
Alg. 1	95.2	28.4	46.1	20.3

the same three types of input as before: we use letters "d", "m", and "i" to indicate face detector, manual alignment, and our iterative alignment algorithm, respectively.

As expected, our algorithm does not perform well except for 'fb', in which the illumination is similar to the training and the mere variation in facial expression is handled well by the sparse error model. For the other three test sets, our algorithm fails because the illumination changes and other variations seriously violate the assumptions of our method. This also explains why LBP performs worse with our iterative alignment algorithm, compared to manual alignment. On the other hand, while LBP achieves the best recognition rates given manually aligned training and testing samples, its performance degrades drastically when the input is obtained directly from the face detector. It is also worth noting that similar poor performance of LBP, as well as other descriptors, has been observed on the Labeled Face in the Wild (LFW) database, where the training is uncontrolled and limited and the input is directly obtained from the face detector [41].

All of these experimental results confirm that both illumination and alignment need to be simultaneously handled well in order to achieve accurate face recognition, even when there is no obvious occlusion or corruption in the test.

4.3 Subject Validation

We test the algorithms' ability to reject invalid images of the 88 subjects not appearing in the training database. As mentioned before, the *sparsity concentration index* (SCI) is used as the outlier rejection rule. Given the sparse representation $\boldsymbol{x}$ of a test image with respect to K training classes, the SCI measures how concentrated the coefficients are on a single class in the dataset and is defined as in [3]:

$$\mathrm{SCI}(\boldsymbol{x}) \doteq \frac{K \cdot \max_i \|\delta_i(\boldsymbol{x})\|_1 / \|\boldsymbol{x}\|_1 - 1}{K - 1} \in [0, 1].$$

It is easy to see that if $\mathrm{SCI}(\boldsymbol{x}) = 1$, the test image is represented using images from one single subject class; if $\mathrm{SCI}(\boldsymbol{x}) = 0$, the coefficients are spread evenly over all classes. Thus, we can choose a threshold $t \in [0, 1]$ for the proposed method and accept a test image as valid if $\mathrm{SCI}(\boldsymbol{x}) \geq t$, and otherwise reject it as invalid. We compare this classifier to classifiers based on thresholding the error residuals of NN, NS, LDA, and LBP.

Figure 11 plots the receiver operating characteristic (ROC) curves, which are generated by sweeping the threshold t through the entire range of possible values for each algorithm.[19] On the left we can see that the SCI based recognition approach significantly outperforms the other algorithms, including LBP, even when all algorithms are coupled with our proposed iterative alignment. In the right plot we again see that classical algorithms, and even LBP, are very sensitive to alignment. Similar contrasts between our algorithm and baseline algorithms were also observed for SRC in [3], though on much smaller datasets.

19. Rejecting invalid images not in the entire database is much more difficult than deciding if two face images are the same subject. Figure 11 should not be confused with typical ROC curves for face similarity, e.g., [42].

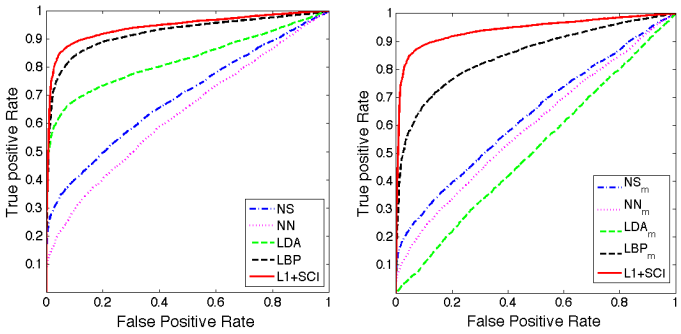


Fig. 11. **ROC curves** for subject validation on Multi-PIE database, (a) for all algorithms with iterative alignment, and (b) for the classical algorithms with manual alignment (indicated by a subscript "m").

Fig. 12. **Recognition under varying level of random block occlusion.** The above row shows examples of occluded test images with occlusion level from 10% to 50%. Our method maintains high recognition rates up to 30% occlusion:

Percent occluded	10%	20%	30%	40%	50%
Recognition rate	99.6%	94.9%	79.6%	46.5%	19.8%

4.4 Recognition with Synthetic Random Block Occlusion

We further test the robustness of our ℓ^1-norm based algorithm to synthetic occlusion. We simulate various levels of occlusion from 10% to 50% by replacing a randomly located block of the face image with an image of a baboon, as shown in Figure 12. In this experiment, to avoid any other factors that may contribute to extra occlusion of the face (such as the change of hair style), we choose illumination 10 from Session 1[20] as testing. The rest of the experimental setting remains unchanged. The table in Figure 12 shows that our algorithm is indeed capable of handling a moderate amount of occlusion. For example, at 20% occlusion, our algorithm still achieves 94.9% recognition rate.

4.5 Recognition with Pose and Expression

We now run tests of our algorithm on a subset of the images from Multi-PIE with pose and expression variation in the test set, although we do not model these variations explicitly. Using the same training set as above, we test our algorithm on images in Session 2 with 15° pose, for all 20 illuminations. As expected, the recognition rate drops to 78.0%. We also test our algorithm on images in Session 3 with smile, again for all

20. This is the same session as the training set.

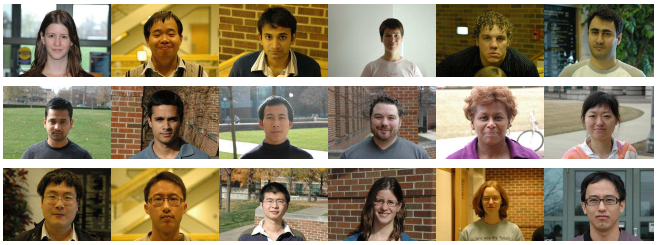


Fig. 13. **Representative examples of categories C1-C3**. One row for each category.

Fig. 14. **Representative examples of category C4**. Top row: successful examples with our method using overlapping blocks. Bottom row: failures with our method using overlapping blocks.

20 illuminations. The recognition rate is 64.8%. Of course, it is reasonable to expect that the performance of our method will be significantly improved if pose and expression data are available in the training.

5 Tests on Our Own Database

Using the training acquisition system we described in Section 3, and shown in Figure 6, we have collected the frontal view of 109 subjects *without eyeglasses* under 38 illuminations shown in Figure 7. For testing our algorithm, we have also taken 935 images of these subjects with a different camera under a variety of practical conditions.

5.1 Necessity of Rear Illuminations

To see how training illuminations affect the performance of our algorithm in practice, we now compare how well a few frontal illuminations can linearly represent: 1. other frontal illuminations taken under the same laboratory conditions, and 2. typical indoor and outdoor illuminations. To this end, we use the face database acquired by our system and use 7 illuminations per subject as training. The illuminations are chosen to be similar to the 7 illuminations used in the previous experiment on Multi-PIE.[21] We then test our algorithm on the remaining $24 - 7 = 17$ frontal illuminations for all the subjects. The recognition rate is 99.8%, nearly perfect. We also test our algorithm on 310 indoor images and 168 outdoor images of these subjects taken under a variety of lighting conditions (category 1 and 2 specified below), similar to the one shown in Figure 1, and the recognition rates for indoor and outdoor images drop down to 94.2% and 89.2%, respectively. This is a strong indication that frontal illuminations taken under laboratory conditions are insufficient for representing test images under typical indoor and outdoor illuminations.

5.2 Large-Scale Test with Sufficient Training Illuminations

Now we use all 109 subjects and 38 illuminations in the training and test on 935 images taken under a variety of practical illuminations and conditions. We have manually partitioned the test images into four main categories:

- C1: 310 *indoor* images of 72 subjects without eyeglasses, frontal view (Fig. 13, row 1).
- C2: 168 *outdoor* images of 48 subjects without eyeglasses, frontal view (Fig. 13, row 2).
- C3: 211 images of 32 subjects with *eyeglasses* (Fig. 13, row 3).
- C4: 246 images of 56 subjects with *sunglasses* (Fig. 14).

We apply Viola and Jones' face detector on these images and directly use the detected faces as the input to our algorithm. Table 4 reports the performance of our algorithm on each category. Since our focus is on face recognition, the errors do not include failures of the face detector on some of the more challenging images. As one can see, our algorithm achieves higher than 95% recognition rates on categories 1-3. Furthermore, using the full set of 38 illuminations indeed improves the performance of our system under practical illumination conditions compared to only using a small subset of 7 illuminations. However, the performance dramatically drops when the faces are occluded by various types of sunglasses, which could cover up to 40% of the entire face. Given the previous experimental results on synthetic random block occlusions, and given that the illuminations are more challenging, the result is not surprising. In the next subsection, we will show how additional assumptions can be used to improve the recognition performance.

TABLE 4
Recognition rates on our own database.

Test Category	C1	C2	C3	C4
Recognition Rate	98.4%	95.8%	95.1%	40.9%

5.3 Improving the Performance with Occlusion using Overlapping Blocks

A traditional approach to improve the performance of face recognition under severe occlusion is to use subregions instead the entire face as a whole. This idea has been explored in many earlier works; see [43], [3] for examples. Since in most real world cases the occlusion is contiguous, it is reasonable to argue that a minority of the subregions are likely to be affected by the occlusion. In this paper, we adopt the same idea and partition the face into four overlapping blocks to better handle sunglasses. This scheme is illustrated in Figure 15. Notice that in this example three out of the four blocks are partially or almost completely occluded. In our experiment, each block is of size 90×48 and covers about two-fifths of the entire face. The testing and training sets are partitioned in the same way. We then independently apply Algorithm 1 and compute a sparse representation after registration for each block independently with respect to the training set. The recognition results for individual blocks are then aggregated by voting.

In this experiment, we found that the using the *sparsity concentration index* (SCI) scores for voting achieves higher

21. We use the illuminations $\{6, 9, 12, 13, 18, 21, 22\}$ shown in Figure 7(b) to mimic the illuminations $\{0, 1, 6, 7, 13, 14, 18\}$ in Multi-PIE.

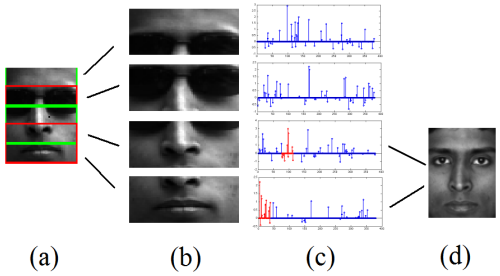


Fig. 15. **Using overlapping blocks to tackle contiguous occlusion.** (a) The test image, occluded by sunglasses. (b) The four overlapping blocks. (c) The sparse representation is calculated after alignment for each block independently. The red lines correspond to his true identity. (d) The true identity is successfully recovered by voting based on the SCI scores.

recognition rate than the residual measure used in Algorithm 1, on category 4 (sunglasses) of our database. The recognition rate is increased to 78.3%, compared to 40.9% obtained without this partition scheme. This is another evidence of the superior ability of SCI on subject validation, since a heavily occluded block can be regarded as an outlier for recognition and should be rejected while voting.

However, we should point out that a major problem with this approach is that occlusion cannot always be expected to fall within any fixed partition of the face image. Therefore, the proposed scheme should only be viewed as an example which shows that the performance under occlusion can be boosted by leveraging local information of a face as well as global information. We leave the investigation of more general models (e.g., MRF [44]) for face recognition with both misalignment and occlusion as an interesting future work.

6 Conclusion

Using a well-though-out combination of existing ideas (iterative image alignment, ℓ_1-error function, SRC, using projectors for illumination), we have proposed a system for recognizing human faces from images taken under practical conditions that is conceptually simple, well motivated, and competitive with state-of-the-art recognition systems for access control scenarios.

The system achieves extremely stable performance under a wide range of variations in illumination, misalignment, and even under small amounts of pose and occlusion. We achieve very good recognition performance on large-scale tests with public datasets as well as our practical face images, while using only frontal 2D images in the gallery and no explicit 3D face model. Our system could potentially be extended to better handle large pose and expression, either by incorporating training images with different poses or expressions or by explicitly modeling and compensating the associated deformations in the alignment stage.

Another important direction for future investigation is to extend the alignment algorithm to better tackle contiguous occlusion. We have demonstrated that misalignment can be naturally handled within the sparse representation framework. More complicated models for spatial continuity, such as Markov random fields, have also been successfully integrated into the computation of a sparse representation of well-aligned test images [45], [44]. A unified approach for face alignment and recognition in the presence of contiguous occlusion remains an open problem.

Acknowledgment

This work was supported by grants NSF IIS 08-49292, NSF ECCS 07-01676, and ONR N00014-09-1-0230. JW thanks Allen Yang of UC Berkeley EECS and Robert Fossum of UIUC Mathematics for discussions related to this work, and acknowledges support from a Microsoft Fellowship and the Lemelson-Illinois Student Prize.

References

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1991.
[2] P. Belhumeur, J. Hespanda, and D. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
[3] 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*, vol. 31, no. 2, pp. 210–227, 2009.
[4] B. Amberg, A. Blake, A. Fitzgibbon, S. Romdhani, and T. Vetter, "Reconstructing high quality face-surfaces using model based stereo," in *Proceedings of IEEE International Conference on Computer Vision*, 2007, pp. 1–8.
[5] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, 2003.
[6] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.
[7] T. Cootes and C. Taylor, "Active shape models – 'smart snakes'," in *Proceedings of British Machine Vision Conference*, 1992.
[8] 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*, vol. 3, 1981, pp. 674–679.
[9] P. Belhumeur and G. Hager, "Tracking in 3D: Image variability decomposition for recovering object pose and illumination," *Pattern Analysis and Applications*, vol. 2, pp. 82–91, 1999.
[10] H. Murase and S. Nayar, "Visual learning and recognition of 3D objects from appearance," *International Journal of Computer Vision*, vol. 14, pp. 5–24, 1995.
[11] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, 2005.
[12] J. Wright and Y. Ma, "Dense error correction via ℓ^1-minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 7, 2010.
[13] M. Osborne and R. Womersley, "Strong uniqueness in seequential linear programming," *Journal of the Australian Mathematical Society, Series B*, vol. 31, pp. 379–384, 1990.
[14] K. Jittorntrum and M. Osborne, "Strong uniqueness and second order convergence in nonlinear discrete approximation," *Numerische Mathematik*, vol. 34, pp. 439–455, 1980.
[15] A. Y. Yang, A. Ganesh, Z. Zhou, S. Sastry, and Y. Ma, "Fast ℓ_1-minimization algorithms and application in robust face recognition," *preprint*, 2010.
[16] 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*, pp. 1519–1524, 2006.
[17] 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*, pp. 230–245, 2007.
[18] 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*, vol. 23, no. 6, pp. 643–660, 2001.
[19] P. Belhumeur and D. Kriegman, "What is the set of images of an object under all possible illumination conditions?" *International Journal of Computer Vision*, vol. 28, no. 3, pp. 245–260, 1998.
[20] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218–233, 2003.

[21] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, 2004.

[22] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework: Part 1: The quantity approximated, the warp update rule, and the gradient descent approximation," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[23] L. Cromme, "Strong uniqueness: A far-reaching criterion for the convergence analysis of iterative procedures," *Numerishe Mathematik*, vol. 29, pp. 179–193, 1978.

[24] A. Lewis and S. Wright, "A proximal method for composite minimization," *Technical Report, University of Wisconsin*, 2008.

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

[26] 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.

[27] R. Gross, I. Matthews, and S. Baker, "Active appearance models with occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 593–604, 2006.

[28] L. Wiskott, J. Fellous, N. Kuiger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, 1997.

[29] J. Huang, X. Huang, and D. Metaxas, "Simultaneous image transformation and sparse representation recovery," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[30] P. Debevec, T. Hawkins, C. Tchou, H. Duiker, W. Sarokin, and M. Sagar, "Acquiring the reflectance field of a human face," in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, 2000, pp. 145–156.

[31] A. Jones, A. Gardner, M. Bolas, I. McDowall, and P. Debevec, "Performance geometry capture for spatially varying relighting," in *ACM SIGGRAPH 2005 Sketches*, 2005, p. 74.

[32] V. Masselus, P. Dutré, and F. Anrys, "The free-form light stage," in *SIGGRAPH*, 2002, p. 262.

[33] L. Zhang, B. Curless, and S. Seitz, "Rapid shape acquisition using color structured light and multi-pass dynamic programming," in *Proceedings of the 1st IEEE International Symposium on 3D Data Processing, Visualization, and Transmission*, 2002, pp. 24–36.

[34] Y. Schechner, S. Nayar, and P. Belhumeur, "Multiplexing for optimal lighting," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 1339–1354, 2007.

[35] J. Yang, K. Yu, and T. Huang, "Supervised translation-invariant sparse coding," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1–8.

[36] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[37] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[38] X. Tan, S. Chen, Z. hua Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognition*, vol. 39, pp. 1725–1745, 2006.

[39] P. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[40] W. Deng, J. Hu, J. Guo, W. Cai, and D. Feng, "Robust, accurate and efficient face recognition from a single training image: A uniform pursuit approach," *Pattern Recognition*, vol. 43, pp. 1748–1762, 2010.

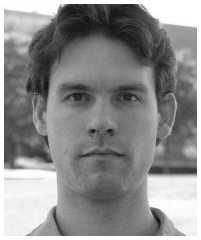
[41] L. Wolf, T. Hassner, and Y. Taigman, "Descriptor based methods in the wild," in *Faces in Real-Life Images Workshop in ECCV*, 2008.

[42] P. Phillips, W. Scruggs, A. O'Tools, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale results," NIST, Tech. Rep. NISTIR 7408, 2007.

[43] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

[44] Z. Zhou, A. Wagner, J. Wright, H. Mobahi, and Y. Ma, "Face recognition with contiguous occlusion using markov random fields," in *Proceedings of IEEE International Conference on Computer Vision*, 2009, pp. 1–8.

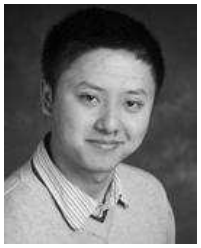
[45] V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk, "Sparse signal recovery using markov random fields," in *Proceedings of Neural Information and Processing Systems*, 2008.



Andrew Wagner received his Bachelor's degree in General Engineering in 2003, and his Master's degree in Electrical Engineering in 2006, from the University of Illinois at Urbana-Champaign, where he is currently a Ph.D. candidate in Electrical & Computer Engineering. His research interests include robotics, computer vision, and optimal control. His recent work focuses on parallel algorithms for sparsity based face recognition.

John Wright received his PhD in Electrical Engineering from the University of Illinois at Urbana-Champaign in October 2009. He is currently a researcher in the Visual Computing group at Microsoft Research Asia. His research focuses on developing provably correct and efficient tools for recovering low-dimensional structure in corrupted high-dimensional datasets. His work has received a number of awards, including the 2009 Lemelson-Illinois Prize for Innovation, the 2009 UIUC Martin Award for Excellence in Graduate Research, a 2008-2010 Microsoft Research Fellowship, a Carver fellowship, and a UIUC Bronze Tablet award.

Arvind Ganesh received his Bachelor's and Master's degrees, both in Electrical Engineering, from the Indian Institute of Technology, Madras, India in 2006. He is currently a PhD candidate in the Electrical & Computer Engineering Department at the University of Illinois, Urbana-Champaign. His research interests include compressed sensing, computer vision, and machine learning. His recent work focuses on low-rank matrix recovery techniques for batch image alignment and texture rectification.



Zihan Zhou received the bachelor's degree in automation from Tsinghua University, China, in 2007. Since then, he has been with University of Illinois at Urbana-Champaign, where he received the MS degree in electrical and computer engineering and where he is now working toward the PhD degree. During the summer of 2009, he was a research intern with Microsoft Research Asia, Beijing. His research interests include computer vision, signal processing and machine learning.

Hossein Mobahi is currently a PhD candidate in the Computer Science Department at the University of Illinois, Urbana-Champaign. His research interests include pattern classification, clustering and optimization. His recent research focuses on iterative smoothing for image alignment.

Yi Ma received two Bachelors degree in Automation and Applied Mathematics from Tsinghua University, Beijing, China, in 1995. He received an Master degree in Electrical Engineering and Computer Sciences (EECS) in 1997, a second Master degree in Mathematics in 2000, and the Ph.D. degree in EECS in 2000, all from the University of California at Berkeley. He is currently an associate professor (with tenure) at the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, and since January 2009 has also served as research manager for the Visual Computing Group at Microsoft Research Asia, Beijing, China.