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

Andrew Wagner, John Wright, Arvind Ganesh, Zihan Zhou, Yi Ma
University of Illinois at Urbana-Champaign, 1308 W. Main st. Urbana, IL 61801
{awagner, jnwright, abalasu2, zzhou7, yima}@illinois.edu*

Abstract

Most contemporary face recognition algorithms work well under laboratory conditions but degrade when tested in less-controlled environments. This is mostly due to the difficulty of simultaneously handling variations in illumination, alignment, pose, and occlusion. In this paper, we propose a simple and practical face recognition system that achieves a high degree of robustness and stability to all these variations. We demonstrate how to use tools from sparse representation to align a test face image with a set of frontal training images in the presence of significant registration error and occlusion. We thoroughly characterize the region of attraction for our alignment algorithm on public face datasets such as Multi-PIE. We further study how to obtain a sufficient set of training illuminations for linearly interpolating practical lighting conditions. We have implemented a complete face recognition system, including a projector-based training acquisition system, in order to evaluate how our algorithms work under practical testing conditions. We show that our system can efficiently and effectively recognize faces under a variety of realistic conditions, using only frontal images under the proposed illuminations as training.

1. Introduction

Automatic face recognition remains one of the most active areas in computer vision. While classical algorithms [11, 2] remain popular for their speed and simplicity, they tend to fail on large-scale, practical tests, falling short of the ultimate goal of truly automating face recognition for real-world applications such as access control for facilities, computer systems and automatic teller machines. These applications are interesting both for their potential sociological impact and also because they allow the possibility of carefully controlling the acquisition of the training data, allowing more tractable and reliable solutions.[1] In this setting, one promising recent direction, set forth in [14], 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.

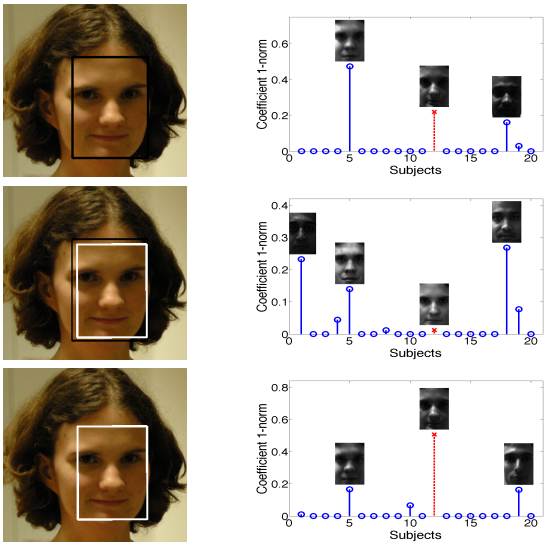


Figure 1. **Compound effect of registration and illumination**. The task is to identify the girl among 20 subjects, by computing the sparse representation of her input face with respect to the entire training set. The absolute sum of the coefficients associated with each subject is plotted on the right. We also show the faces reconstructed with each subject's training images weighted by the associated sparse coefficients. The red line (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:** Informative representation obtained by using both well-aligned input face and sufficient (all 38) illuminations in the training.

While that work achieves impressive results on public datasets taken under controlled laboratory conditions such as Extended Yale B [4], it fails to address two critical aspects of real world face recognition: significant variations in both the *image domain* and in the *image value*. We illustrate this with 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 representation is no longer informative, even if suffi-

*This work was supported by NSF IIS 08-49292, NSF ECCS 07-01676, and ONR N00014-09-1-0230 grants. John Wright was partially supported by a Microsoft Fellowship.

[1] Face recognition with less-controlled training samples taken under uncontrolled scenarios remains an active research area as well [7].

cient illuminations are present in the training as shown in Figure 1 top. In addition, in order to sufficiently interpolate the illumination of a typical indoor (or outdoor) environment, illuminations from behind the subject are also needed in the training. Otherwise, even for perfectly aligned test images, the representation will not necessarily be sparse or informative, as shown by the example in Figure 1 middle. Unfortunately, most public face databases lack images with a significant component of rear (more than 90 degrees from frontal) illumination, either for training or testing.

Contributions. In this paper, we show how the two *strongly coupled* issues of registration and illumination can be naturally addressed within the 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, scale, and pose, even when the face is only partially visible due to eyeglasses, hats, closed eyes and open mouth, sensor saturation, etc. We also propose a sufficient, if not the smallest, set of training illuminations that is capable of interpolating typical indoor and outdoor lighting, along with a practical hardware system for capturing them. Finally, we demonstrate the effectiveness of the proposed new methods with a complete face recognition system that is *simple, stable, and scalable*. The proposed algorithm performs robust automatic recognition of subjects from loosely controlled images taken both indoors and outdoors, using labeled frontal views of the subjects' faces under the proposed illuminations for training and an off-the-shelf face detector[2] to detect faces in images.

2. Handling Practical Registration Error

As demonstrated in Figure 1 top, the main limitation of the *sparse representation and classification* (SRC) algorithm of [14] 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 [14], 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 occlusion. The sparse representation can be recovered by minimizing the sum or the 1-norm[4] of $\boldsymbol{x}$ and $\boldsymbol{e}$:

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

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 [14] is no longer appropriate, as seen in Figure 1 top.

Batch and individual alignment. Notice that 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 $\boldsymbol{y} \circ \tau = A\boldsymbol{x}_0 + \boldsymbol{e}_0$. 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, by solving

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

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. In this sense, the misaligned recognition problem differs from the well-aligned version studied in [14]. 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:

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

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.

Alignment via iterative 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:

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

[2] In this paper, we use the OpenCV implementation of the Viola and Jones' face detector [12].

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

[4] The 1-norm of a vector $\boldsymbol{x}$ is the sum of absolute values of the entries.

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

Figure 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 $\boldsymbol{e}$.

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. Near the correct alignment we expect the aligned testing 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 of the registration error $\boldsymbol{e}$, in terms of its ℓ^1-norm:

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

Notice that this is different from the popular choice that minimizes the 2-norm of the registration error:

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

which is also equivalent to finding the deformation step $\Delta\tau$ by solving the least-square problem: $\min \|\boldsymbol{y} + J\Delta\tau - A_i\boldsymbol{x}\|_2$. 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 $\boldsymbol{y}_0$, iterative ℓ^1-minimization (5) is significantly better than iterative ℓ^2-minimization (6). Figure 2 shows an example.

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 expanding a single black pixel in 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 down-sampled versions of the training and testing images, using the result of one level to initialize the next.

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.

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

Input: Frontal training images $A_1, A_2, \ldots, A_K \in \mathbb{R}^{m\times n_i}$ for K subjects, a test image $\boldsymbol{y} \in \mathbb{R}^m$ and a deformation group T considered.
for each subject k,
  $\tau^{(0)} \leftarrow I$.
  do
    $\tilde{\boldsymbol{y}}(\tau) \leftarrow \frac{\boldsymbol{y}\circ\tau}{\|\boldsymbol{y}\circ\tau\|_2}$; $\quad J \leftarrow \frac{\partial}{\partial\tau}\tilde{\boldsymbol{y}}(\tau)\big|_{\tau^{(i)}}$;
    $\Delta\tau = \arg\min \|\boldsymbol{e}\|_1$ s.t. $\tilde{\boldsymbol{y}}+J\Delta\tau = A_k\boldsymbol{x}+\boldsymbol{e},\ \boldsymbol{x} \geq \boldsymbol{0}$.
    $\tau^{(i+1)} \leftarrow \tau^{(i)} + \Delta\tau$;
  while $\|\tau^{(i+1)} - \tau^{(i)}\| \geq \varepsilon$.
end
Keep the top S candidates $k_1, \ldots, k_S$ with the smallest residuals $\|\boldsymbol{e}\|_1$.
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:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x},\boldsymbol{e}} \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \ \text{ subj } \ \boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e},\ \boldsymbol{x} \geq 0.$$

Compute residuals $r_i(\boldsymbol{y}) = \|\boldsymbol{y} - A_i\,\hat{\boldsymbol{x}}_i\|_2$ for $i = k_1, \ldots, k_S$.
Output: identity$(\boldsymbol{y}) = \arg\min_i r_i(\boldsymbol{y})$.

The most important free parameter in Algorithm 1 is the class of deformations T. In our experiments, we typically use 2D similarity transformations, $T = \mathbb{SE}(2) \times \mathbb{R}_+$, for compensating error incurred by face detector, or 2D projective transformations, $T = \mathbb{GL}(3)$, for handling some pose variation. The parameter S decides how many top candidates get 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 [14]. We have implemented a fast linear program for our algorithm in C. Running on a 2.8GHz Mac Pro, alignment takes 0.65 second per subject for our database.

Simulations and experiments on region of attraction. We now perform simulations and experiments demonstrating the effectiveness of the individual alignment procedure outlined in the previous section, and clarifying its operating range. We delay large-scale recognition experiments to

Section 4, 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 [6]. We select 120 subjects in Session 2, use 11 illuminations per person from Session 2 for training, and test on one new illumination from Session 3.[5] We manually select eye corners in both training and testing as the ground truth for registration. We down-sample the images to 80×60 pixels[6] and the distance between the two outer eye corners are normalized to be 50 pixels for each person. We introduce artificial deformation to the testing image with a combination of translation or rotation. We consider a registration successful if the difference between the final registration error is within 1% of the error by manual registration. Figure 3 shows the percentage of successful registrations for the 120 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 25% change in scale.

   We have gathered the statistics of the Viola and Jones' face detector on the Multi-PIE datasets. 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 variation in scale is 17%. This falls safely inside the range of attraction for our alignment algorithm.

2. *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 the image of the subject under a typical indoor lighting condition at pose ranging from -90° to $+90^\circ$ with step size 5.625°, a total of 33 poses. We use Viola and Jones' face detector to initialize our alignment algorithm. Figure 4 shows the typical alignment results of our algorithm, working surprisingly well with poses up to $\pm 45^\circ$.

[5]The training are illuminations $\{0, 1, 3, 5, 7, 9, 11, 13, 14, 16, 18\}$ of [6], and the testing is the illumination 10.

[6]Unless otherwise stated, this will be the default resolution at which we prepare all our training and testing datasets and run all our experiments.

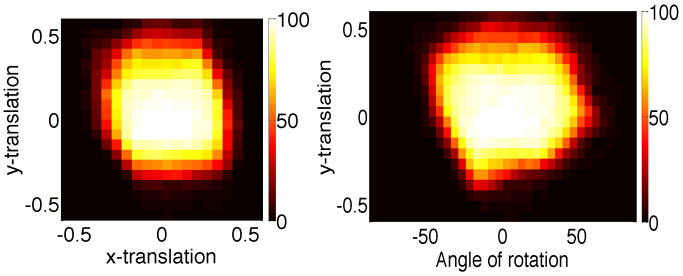


Figure 3. **Region of attraction.** Fraction of subjects for which the algorithm successfully aligns a synthetically perturbed test image. The amount of translation is expressed as a fraction of the distance between the outer eye corners, and the amount of in-plane rotation in degrees. **Left:** Simultaneous translation in x and y directions. More than 90% of the subjects were correctly aligned for any combination of x and y translations, each upto 0.2. **Right:** Simultaneous translation in y direction and in-plane rotation θ. More than 90% of the subjects were correctly aligned for any combination of y translation upto 0.2 and θ upto 25°.

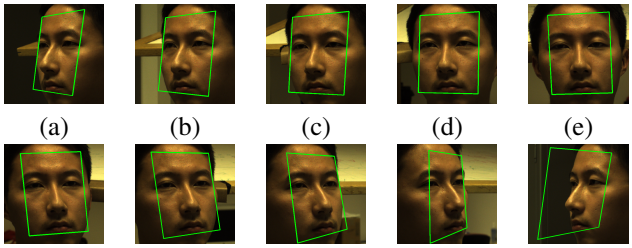


Figure 4. **Aligning different poses to frontal training images. (a) to (i):** good alignment for pose from -45° to $+45^\circ$. **(j):** a case when the algorithm fails for an extreme pose ($> 45^\circ$).

Relationship to existing work. Our modification to SRC roots solidly in the tradition of adding deformation-robustness to face recognition algorithms [3, 5, 13]. However, the only previous work to investigate face alignment in the context of sparse signal representation and SRC is the work of [8]. 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 exact experimental settings in [8]. 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. For our algorithm we down-sample all the images to 88×80 for memory reasons, whereas the work of [8] uses random pro-

jections. Our algorithm achieves the recognition rate 88.59% which is on par with the result reported in [8]. However, this special setting is disadvantageous to our algorithm: The use of cropped test images introduces boundary effects, and the presence of very extreme illuminations makes enforcing nonnegativity of $\boldsymbol{x}$ (as in Algorithm 1 less appropriate. We further discuss the justification for nonnegativity in the next section.

2. *CMU Multi-PIE.* In this experiment, we choose 160 subjects from the CMU Multi-PIE, 11 training images from Session 2 and 1 test image from Session 3 per person. The setting is exactly the same as the previous experiment on 2D deformation, except that we have more subjects. We again work with downsampled images of size 80×60. An artificial translation of 5 pixels (in both x and y directions) was induced in the test image. The algorithm of [8] achieves a recognition rate of 73.75%,[7] while ours does 90.625%.

3. Handling Practical Illumination Variation

In the above section, we have made the assumption that the test image, although taken under some arbitrary illumination, can be linearly interpolated by a finite number of training illuminations. It has been shown that for a convex Lambertian surface, one only needs about nine basis illuminations to linearly interpolate all other illuminations [1]. Although a human face is neither perfectly Lambertian nor convex, it has been observed in various empirical studies that one can often get away using a similar number of frontal illuminations to interpolate a wide range of new frontal illuminations that taken under the same laboratory conditions [4]. 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 interpolate images of a faces taken under typical indoor or outdoor conditions (see the experiment conducted in Section 4.2). The representation computed is not always sparse or informative, as shown by the example in Figure 1. Subsequently, the recognition could become inaccurate. Thus, to ensure our algorithm works in practice, we need to find a set of training illuminations that are indeed *sufficient* to linearly interpolate variety of practical indoor and outdoor illuminations.

Capturing a sufficient set of training illuminations. To this end, we have designed a system that can illuminate the subject from all directions above horizontal, while acquiring the subject's frontal images. A sketch of the system is shown in Figure 5: The illumination system consists of four projectors that display various bright patterns onto the three

[7] That algorithm has two free parameters - l and d. For this experiment we chose $l = 1$ and $d = 514$ (higher values may get a better recognition rate at the expense of higher running time).

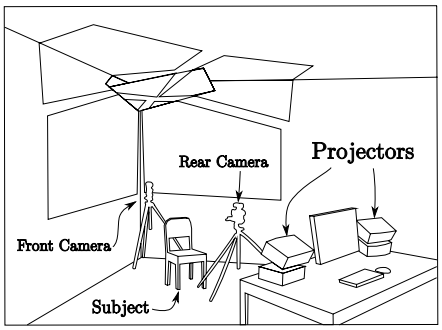


Figure 5. **Training acquisition system:** Four projectors and two cameras controlled by one computer.

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:

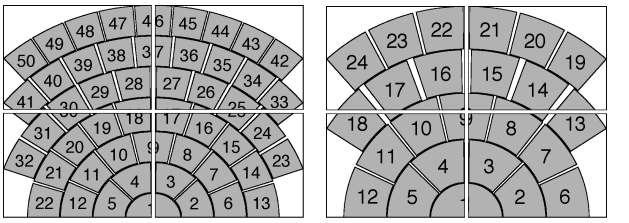
- The illuminations can be defined in software.
- It is easy to capture many different illuminations.
- There is no need to mount cameras on walls or construct a large dome.
- No custom hardware is needed for a basic system.

With our projector system, our choice of illuminations is constrained only by the need to achieve a good SNR for representing typical test images and a reasonably short total acquisition time.[8] We ran two experiments to guide our choice of illuminations for our large-scale experiments:

- *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 6 (a). The more rings of training illuminations are added, the lower the representation error becomes, with diminishing returns.
- *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

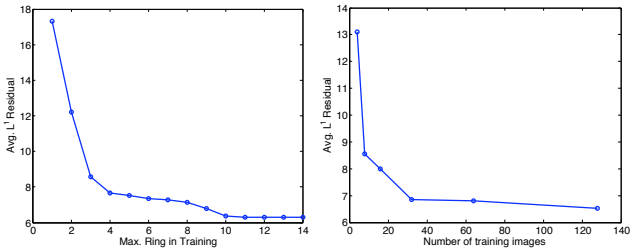
[8] Better SNR can be achieved with more illuminations but that will increase the capture time for each subject.

registration error is computed as in the coverage experiment and shown in Figure 7 (b). Again, diminishing returns are observed as more illuminations are added.



(a) Coverage Experiment (b) Chosen Illumination Patterns

Figure 6. **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

Figure 7. **Study of sufficient illuminations.** The average ℓ^1 registration residual versus different illumination training sets.

Chosen illumination patterns. In the plot for the coverage experiment, Figure 7 (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 can span majority of illumination on the face, some illuminations from the back are needed in the training to emulate the effect of ambient illumination from all directions. In the plot for the granularity experiment, Figure 7 (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 6 (b). For our large-scale experiments, we have collected those illuminations for all our subjects.[9]

[9]It is very likely that with more careful experiments, we can further reduce the number of illuminations needed. Especially some of the frontal illuminations might be redundant. But we keep those in our training anyway as the additional images do not add too much cost to our alignment and recognition algorithm.

See below for the 38 images for one subject:

The role of nonnegativity. One critical issue in linear illumination models is whether to enforce nonnegativity in the coefficients $\boldsymbol{x}$: whether to model illumination using a cone or a subspace. Nonnegative combinations are guaranteed to correspond to physically plausible illuminations, but will not be sufficient to represent all physical illuminations unless the training images actually span the boundary of the illumination cone. Because we have a flexible acquisition system, we can directly generate a set of illuminations that span most of the illumination cone, without resorting to negative coefficients and risking overfitting. Thus, in Algorithm 1, we have enforced $\boldsymbol{x}$ to be non-negative.

4. Overall System Evaluation

In this section, to verify the performance of our algorithm and system, we conduct comprehensive experiments on large-scale face databases. We first test on the largest public face database available that is suitable for testing our algorithm, the CMU Multi-PIE. The goal is to show that our algorithm can indeed be used to achieve good performance on such a dataset with test images obtained from an off-the-shelf face detector, even though we can only use a small number of, not necessarily sufficient, training illuminations. We then test our algorithm on a face dataset that is collected by our own system. The goal is 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.

4.1. Tests on public databases

CMU Multi-PIE provides the most extensive test of our algorithm 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 the 249 subjects present in Session 1 as a training set. The remaining 88 subjects are considered "outliers" or invalid images. For each of the 249 training subjects, we include frontal images of 7 frontal illuminations[10], taken with neutral expression. As suggested by the work of [4], these extreme frontal illuminations would be sufficient to interpolate other frontal illuminations, as will also be corroborated by the next experiment on our own dataset. For the test set, we use all 20 illuminations from Sessions 2-4, which were recorded at distinct times over a period of

[10]They are illuminations $\{0, 1, 7, 13, 14, 16, 18\}$ of [6]. For each directional illumination, we subtract the ambient-illuminated image 0.

several months. The dataset is challenging due to the large number of subjects, and due to natural variation in subject appearance over time. Table 8 shows the result of our algorithm on each of the 3 testing sessions. Our algorithm achieves recognition rates above 90% for all three sessions, with input *directly* obtained from the Viola and Jones' face detector – no manual intervention. We compare our result to baseline linear-projection-based algorithms, such as Nearest Neighbor (NN), Nearest Subspace (NS) [9], and Linear Discriminant Analysis (LDA) [2].[11] Since these algorithms assume pixel-accurate alignment, they are not expected to work well if the test is not well-aligned with the training. In the table of Figure 8, we report the results of those algorithms with two types of input: 1. the output of the Viola and Jones' detector, indicated by a subscript "d"; 2. the input face is aligned to the training with manually selected outer eye corners, indicated by a subscript "m". Notice that, despite careful manual registration, these baseline algorithms perform significantly worse than our algorithm, which uses input directly from the face detector. The performance of the LDA algorithm on Multi-PIE reported here seems to agree with that reported already in [6].

Subject validation. We test the algorithms' ability to reject invalid images of the 88 subjects not appearing in the training database. Figure 8 (bottom) plots the receiver operating characteristic (ROC) curve for each algorithm.[12] Similar contrasts between our algorithm and baseline algorithms were also observed for SRC in [14], though on much smaller datasets.

Cause of errors. 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. In fact, for those subjects, alignment and recognition fail on almost all test illuminations.

Pose and expression. We also run limited tests of our algorithm on images with pose and expression in Multi-PIE. Using the same training as above, we test our algorithm on images in Session 2 with 15° pose, for all 20 illuminations. The recognition rate is 77.5%. We also test our algorithm on images in Session 3 with smile. For illumination 0 (ambient), the rate is 58.5%, for illumination 10, the rate is 68.6%.

4.2. Tests on our own datasets

Using the training acquisition system that we have described in the previous section, Figure 5, we have collected

[11] We do not list results on PCA [11] as its performance is always below that of Nearest Subspace.

[12] Rejecting invalid images not in the entire database is much more difficult than deciding if two face images are the same subject. Figure 8 should not be confused with typical ROC curves for face similarity, e.g., [10].

Rec. Rates	Session 2	Session 3	Session 4
LDA_d (LDA_m)	5.1 (49.4)%	5.9 (44.3)%	4.3 (47.9)%
NN_d (NN_m)	26.4 (67.3)%	24.7 (66.2)%	21.9 (62.8)%
NS_d (NS_m)	30.8 (77.6)%	29.4 (74.3)%	24.6 (73.4)%
Algorithm 1	**91.4** %	**90.3** %	**90.2** %

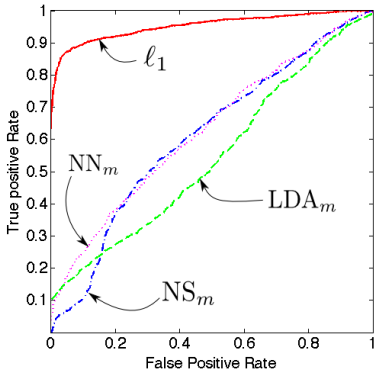


Figure 8. **Large-scale experiments on Multi PIE. Top:** Recognition rates; **Bottom:** ROC curves for our algorithm (labeled as "ℓ_1"), compared with those for NN_m, NS_m, and LDA_m.

Figure 9. **Representative examples of failed Multi-PIE subjects**. **Top:** training from Session 1; **Bottom:** test images from Session 2 – first four are frontal and the last two with 15° pose. Notice the change of hair style and facial hair, which makes alignment fail on those subjects, actually regardless of test image illuminations.

the frontal view of 74 subjects *without eyeglasses* under 38 illuminations shown in Figure 6. For testing our algorithm, we have also taken 593 images of these subjects with a different camera under a variety of practical conditions.

Limitation of frontal illuminations. To see how training illuminations affect the performance of our algorithm in practice, we now compare how well a few frontal illuminations can interpolate: 1. other frontal illuminations taken under the same laboratory conditions, and 2. typical indoor and outdoor illuminations. To this end, we select 20 subjects from 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.[13] We then test our algorithm on the remaining $24 - 7 = 17$ frontal illuminations for all the 20 subjects. The recognition rate is 99.7%, nearly perfect. We also test our algorithm on 173 frontal images of these subjects taken under a variety of indoor and outdoor conditions (in category 1 specified below), similar to

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

the one shown in Figure 1, and the recognition drops down to 93.6%. One would expect the rate to drop even further when the number of subjects increases.

Large-scale test with sufficient training illuminations. Now we use all 74 subjects and 38 illuminations in the training and test on 593 images taken under a variety of conditions. Based on the main variability in the test images, we have partitioned them into five main categories:

C1: 242 images of 47 subjects without eyeglasses, generally frontal view, under a variety of practical illuminations (indoor and outdoor) (Fig. 10, row 1).

C2: 109 images of 23 subjects with eyeglasses (Fig. 10, row 2).

C3: 19 images of 14 subjects with sunglasses (Fig. 10, row 3).

C4: 100 images of 40 subjects with noticeable expressions, poses, mild blur, and sometimes occlusion (Fig. 11, both rows).

C5: 123 images of 17 subjects with little control (out of focus, motion blur, significant pose, large occlusion, funny faces, extreme expressions) (Fig. 12, both rows).

We apply Viola and Jones' face detector on these images and directly use the detected faces as the input to our algorithm. The table below reports the performance of our algorithm on each category. The errors include failures of the face detector on some of the more challenging images.

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

5. Conclusion

We have proposed a new algorithm and system for recognizing human faces from images taken under practical conditions. The proposed system is very *simple* and hence the results are easy to reproduce. The proposed algorithm is *scalable* both in terms of computational complexity and recognition performance. The system is directly compatible with off-the-shelf face detectors and achieves extremely *stable* performance under a wide range of variations in illumination, misalignment, pose, and occlusion. We achieve very good recognition performance on large-scale tests with public datasets and our practical face images, using only frontal 2D images in the training without any explicit 3D face model. Our implementation still has plenty of room for further engineering improvements.

References

[1] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *PAMI*, 25(2):218–233, 2003.

[2] P. Belhumeur, J. Hespanda, and D. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *PAMI*, 19(7):711–720, 1997.

[3] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *PAMI*, 23(6):681–685, 2001.

Figure 10. **Representative examples of categories 1-3**. One row for each category.

Figure 11. **Representative examples of category 4**. Top row: successful examples. Bottom row: failures.

Figure 12. **Representative examples of category 5**. Top row: successful examples. Bottom row: failures.

[4] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *PAMI*, 23(6), 2001.

[5] R. Gross, I. Matthews, and S. Baker. Active appearance models with occlusion. *PAMI*, 24(6):593 – 604, 2006.

[6] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. In *FGR*, 2008.

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

[8] J. Huang, X. Huang, and D. Metaxas. Simultaneous image transformation and sparse representation recovery. In *CVPR*, 2008.

[9] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *PAMI*, 27(5):684–698, 2005.

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

[11] M. Turk and A. Pentland. Eigenfaces for recognition. In *CVPR*, 1991.

[12] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57:137 – 154, 2004.

[13] L. Wiskott, J. Fellous, N. Kuiger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *PAMI*, 19(7), 1997.

[14] J. Wright, A. Yang, A. Ganesh, Sastry, and Y. Ma. Robust face recognition via sparse representation. to appear in *PAMI*, 2008.