From Few to Many: Generative Models for Recognition Under Variable Pose and Illumination*

Athinodoros S. Georghiades Peter N. Belhumeur
Departments of Electrical Engineering
and Computer Science
Yale University
New Haven, CT 06520-8267

David J. Kriegman
Beckman Institute
University of Illinois, Urbana-Champaign
Urbana, IL 61801

Abstract

Image variability due to changes in pose and illumination can seriously impair object recognition. This paper presents appearance-based methods which, unlike previous appearance-based approaches, require only a small set of training images to generate a rich representation that models this variability. Specifically, from as few as three images of an object in fixed pose seen under slightly varying but unknown lighting, a surface and an albedo map are reconstructed. These are then used to generate synthetic images with large variations in pose and illumination and thus build a representation useful for object recognition. Our methods have been tested within the domain of face recognition on a subset of the Yale Face Database B containing 4050 images of 10 faces seen under variable pose and illumination. This database was specifically gathered for testing these generative methods. Their performance is shown to exceed that of popular existing methods.

1 Introduction

An object can appear strikingly different due to changes in pose and illumination (see Figure 1). To handle this image variability, object recognition systems usually use one of the following approaches: (a) control viewing conditions, (b) employ a representation that is invariant to the viewing conditions, or (c) directly model this variability. For example, there is a long tradition of performing edge detection at an early stage since the presence of an edge at an image location is thought to be largely independent of lighting. It has been observed, however, that methods for face recognition based on finding local image features and using their geometric relation are generally ineffective [4].

Here, we consider issues in modeling the effects of both pose and illumination variability rather than trying to achieve invariance to these viewing conditions. We show how these models can be exploited for reconstructing the 3-D geometry of objects and used to significantly increase the performance of appearance-based recognition systems. We demonstrate the use of these models within the context of face recognition, but believe that they have much broader applicability.

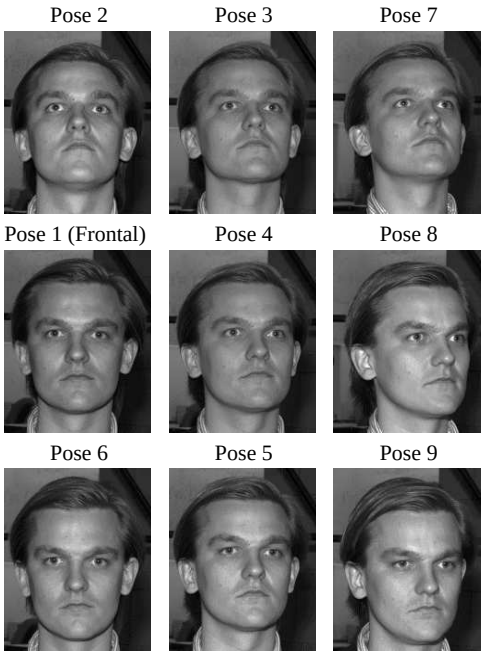
Methods have recently been introduced which use low-dimensional representations of images of objects to perform recognition, see for example [8, 13, 19]. These methods, often termed appearance-based methods, differ from feature-based methods in that their low-dimensional representation is, in a least-squares sense, faithful to the original image. Systems such as SLAM [13] and Eigenfaces [19] have demonstrated the power of appearance-based methods both in ease of implementation and in accuracy.

Yet, these methods suffer from an important drawback: recognition of an object under a particular pose and lighting can be performed reliably *provided the object has been previously seen under similar circumstances.* In other words, these methods in their original form have no way of extrapolating to novel viewing conditions. Here, we consider the construction of a generative appearance model and demonstrate its usefulness for image-based rendering and recognition.

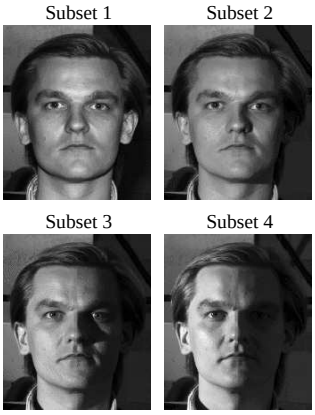
The presented approach is, in spirit, an appearance-based method for recognizing objects under large variations in pose and illumination. However, it differs substantially from previous methods in that it uses as few as three images of each object seen in fixed pose and under small but unknown changes in lighting. From these images, it generates a rich representation that models the object's image variability due to pose and illumination. One might think that pose variation is harder to handle because of occlusion or appearance of surface points and the non-linear warping of the image coordinates. Yet, as demonstrated by favorable recognition results, our approach can successfully generalize the concept of the illumination cone which models all the images of a Lambertian object in fixed pose under all variation in illumination [1].

New recognition algorithms based on these generative models have been tested on a subset of the Yale Face Database B (see Figure 1) which was specifically gathered for this purpose. This subset contained 4050 images of 10 faces each seen under 45 illumination conditions over nine poses. As we will see, these new algorithms outperform popular existing techniques.

*P. N. Belhumeur and A. S. Georghiades were supported by a Presidential Early Career Award, a NSF Career Award IRI-9703134, and an ARO grant DAAH04-95-1-0494. D. J. Kriegman was supported by NSF under NYI, IRI-9257990, and by an ARO grant DAAG55-98-1-0168.



a.



b.

Figure 1: Example images from the Yale Face Database B, showing the variability due to pose and illumination in the images of a single individual. a. An image from each of the nine different poses; b. A representative image from each illumination subset—Subset 1 (12°), Subset 2 (25°), Subset 3 (50°), Subset 4 (77°).

2 Modeling Illumination and Pose

2.1 The Illumination Cone

In earlier work, it was shown that for a convex object with Lambertian reflectance, the set of all n-pixel images under an arbitrary combination of point light sources forms a convex polyhedral cone in the image space $\mathbb{R}^n$. This cone can be built from as few as three images [1]. Here, we outline the relevant results.

Let $\mathbf{x} \in \mathbb{R}^n$ denote an image with n pixels of a convex object with a Lambertian reflectance function illuminated by a single point source at infinity. Let $B \in \mathbb{R}^{n \times 3}$ be a matrix where each row in B is the product of the albedo with the inward pointing unit normal for a point on the surface projecting to a particular pixel in the image. A point light source at infinity can be represented by $\mathbf{s} \in \mathbb{R}^3$ signifying the product of the light source intensity with a unit vector in the direction of the light source. A convex Lambertian surface with normals and albedo given by B, illuminated by $\mathbf{s}$, produces an image $\mathbf{x}$ given by

$$\mathbf{x} = \max(B\mathbf{s}, \mathbf{0}), \tag{1}$$

where $\max(B\mathbf{s}, \mathbf{0})$ sets to zero all negative components of the vector $B\mathbf{s}$. The pixels set to zero correspond to the surface points lying in an attached shadow. Convexity of the object's shape is assumed at this point to avoid cast shadows. Note that when no part of the surface is shadowed, $\mathbf{x}$ lies in the 3-D subspace $\mathcal{L}$ given by the span of the columns of B [8, 14, 16].

If an object is illuminated by k light sources at infinity, then the image is given by the superposition of the images which would have been produced by the individual light sources, i.e.,

$$\mathbf{x} = \sum_{i=1}^{k} \max(B\mathbf{s}_i, \mathbf{0}) \tag{2}$$

where $\mathbf{s}_i$ is a single light source. Due to this superposition, it follows that the set of all possible images $\mathcal{C}$ of a convex Lambertian surface created by varying the direction and strength of an arbitrary number of point light sources at infinity is a convex cone. It is also evident from Equation 2 that this convex cone is completely described by matrix B.

Furthermore, any image in the illumination cone $\mathcal{C}$ (including the boundary) can be determined as a convex combination of *extreme rays* (images) given by

$$\mathbf{x}_{ij} = \max(B\mathbf{s}_{ij}, \mathbf{0}), \tag{3}$$

where

$$\mathbf{s}_{ij} = \mathbf{b}_i \times \mathbf{b}_j. \tag{4}$$

The vectors $\mathbf{b}_i$ and $\mathbf{b}_j$ are the rows of B with $i \neq j$. It is clear that there are at most $m(m-1)$ extreme rays for $m \leq n$ independent surface normals.

2.2 Constructing the Illumination Cone

Equations 3 and 4 suggest a way to construct the illumination cone for each object: gather three or more images in fixed pose under differing but unknown illumination without shadowing and use these images to estimate a basis for the 3-D illumination subspace $\mathcal{L}$. One way of estimation is to normalize the images to be of unit length, and then use singular value decomposition (SVD) to calculate in a least-squares sense the best 3-D orthogonal basis in the form of matrix B^*. Note that even if the columns of B^* exactly span the subspace $\mathcal{L}$, they differ from those of B by an unknown linear transformation, i.e., $B = B^*A$ where $A \in GL(3)$; for any light source, $\mathbf{x} = B\mathbf{s} = (B^*A)(A^{-1}\mathbf{s})$ [10]. Nonetheless, both B^* and B define the same illumination cone $\mathcal{C}$ and represent valid illumination models.

From B^*, the extreme rays defining the illumination cone $\mathcal{C}$ can be computed using Equations 3 and 4.

Unfortunately, using SVD in the above procedure leads to an inaccurate estimate of B^*. For even a convex object whose occluding contour is visible, there is only one light source direction (the viewing direction) for which no point on the surface is in shadow. For any other light source direction, shadows will be present. If the object is non-convex, such as a face, then shadowing in the modeling images is likely to be more pronounced. When SVD is used to find B^* from images with shadows, these systematic errors bias its estimate significantly. Therefore, an alternative way is needed to find B^* that takes into account the fact that some data values are invalid and should not be used in the estimation. For the purpose of this estimation, any invalid data can be treated as missing measurements.

The technique we use here is a combination of two algorithms. A variation of [17] (see also [11, 18]) which finds a basis for the 3-D linear subspace $\mathcal{L}$ from image data with missing elements is used together with the method in [6] which enforces integrability in shape from shading. We have modified the latter method to guarantee integrability in the estimates of the basis vectors of subspace $\mathcal{L}$ from multiple images. By enforcing integrability a surface context is introduced. Namely, the vector field induced by the basis vectors is guaranteed to be a gradient field that corresponds to a surface.

Furthermore, enforcing integrability inherently leads to more accurate estimates because there are fewer parameters (or degrees of freedom) to determine. It also resolves six out of the nine parameters of $A \in GL(3)$. The other three correspond to the generalized bas-relief (GBR) transformation parameters which cannot be resolved with illumination information alone (i.e. shading and shadows) [2]. This means we cannot recover the true matrix B and its corresponding surface, $z(x, y)$. We can only find their GBR versions $\bar{B}$ and $\bar{z}(x, y)$.

Our estimation algorithm is iterative and to enforce integrability, the possibly non-integrable vector field induced by the current estimate of B^* is, in each iteration, projected down to the space of integrable vector fields, or gradient fields [6]. To begin, let us expand the surface $\bar{z}(x, y)$ using basis surfaces (functions):

$$\bar{z}(x, y; \bar{c}(\mathbf{w})) = \sum \bar{c}(\mathbf{w})\phi(x, y; \mathbf{w}) \tag{5}$$

where $\mathbf{w} = (w_x, w_y)$ is a two dimensional index, and $\{\phi(x, y; \mathbf{w})\}$ is a finite set of basis functions which are not necessarily orthogonal. We chose the discrete cosine basis so that $\{\bar{c}(\mathbf{w})\}$ is exactly the set of the 2-D discrete cosine transform (DCT) coefficients of $\bar{z}(x, y)$.

Note that the partial derivatives of $\bar{z}(x, y)$ can also be expressed in terms of this expansion, giving

$$\bar{z}_x(x, y; \bar{c}(\mathbf{w})) = \sum \bar{c}(\mathbf{w})\phi_x(x, y; \mathbf{w}) \tag{6}$$

and

$$\bar{z}_y(x, y; \bar{c}(\mathbf{w})) = \sum \bar{c}(\mathbf{w})\phi_y(x, y; \mathbf{w}). \tag{7}$$

Since the partial derivatives of the basis functions, $\phi_x(x, y; \mathbf{w})$ and $\phi_y(x, y; \mathbf{w})$, are integrable and the expansions of $\bar{z}_x(x, y)$ and $\bar{z}_y(x, y)$ share the same coefficients $\bar{c}(\mathbf{w})$, it is easy to see that $\bar{z}_{xy}(x, y) = \bar{z}_{yx}(x, y)$.

Suppose, now, we have the possibly non-integrable estimate B^* from which we can easily deduce the possibly non-integrable partial derivatives $z_x^*(x, y)$ and $z_y^*(x, y)$. These can also be expressed as a series, giving

$$z_x^*(x, y; c_1^*(\mathbf{w})) = \sum c_1^*(\mathbf{w})\phi_x(x, y; \mathbf{w}) \tag{8}$$

and

$$z_y^*(x, y; c_2^*(\mathbf{w})) = \sum c_2^*(\mathbf{w})\phi_y(x, y; \mathbf{w}). \tag{9}$$

Note that in general $c_1^*(\mathbf{w}) \neq c_2^*(\mathbf{w})$ which implies that $z_{xy}^*(x, y) \neq z_{yx}^*(x, y)$.

Let us assume that $z_x^*(x, y)$ and $z_y^*(x, y)$ are known from an estimate of B^* and we would like to find $\bar{z}_x(x, y)$ and $\bar{z}_y(x, y)$ (a set of integrable partial derivatives) which are as close as possible to $z_x^*(x, y)$ and $z_y^*(x, y)$, respectively, in a least-squares sense. The goal is to minimize the following,

$$\min_{\bar{c}} \sum_{x,y} \quad (\bar{z}_x(x, y; \bar{c}) - z_x^*(x, y; c_1^*))^2 + (\bar{z}_y(x, y; \bar{c}) - z_y^*(x, y; c_2^*))^2. \tag{10}$$

In other words, take a set of possibly non-integrable partial derivatives, $z_x^*(x, y)$ and $z_y^*(x, y)$, and "enforce" integrability by finding the least-squares fit of integrable partial derivatives $\bar{z}_x(x, y)$ and $\bar{z}_y(x, y)$. Notice that to get the GBR transformed surface $\bar{z}(x, y)$ we need only perform the inverse 2-D DCT on the coefficients $\bar{c}(\mathbf{w})$.

The above procedure is incorporated into the following algorithm. To begin, define the data matrix for k images of an individual to be $X = [\mathbf{x}_1, \ldots, \mathbf{x}_k]$. If there were no shadowing, X would be rank 3 [15] (assuming no image noise), and we could use SVD to factorize X into $X = B^*S$ where S is a $3 \times k$ matrix whose columns $\mathbf{s}_i$ are the light source directions scaled by their corresponding source intensities for all k images.

Since the images have shadows (both cast and attached), and possibly saturations, we first have to determine which data values do not satisfy the Lambertian assumption. Unlike saturations, which can be simply determined, finding shadows is more involved. In our implementation, a pixel is assigned to be in shadow if its value divided by its corresponding albedo is below a threshold. As an initial estimate of the albedo we use the average of the modeling (or training) images. A conservative threshold is then chosen to determine shadows making it almost certain no invalid data is included in the estimation process, at the small expense of throwing away a few valid measurements. After finding the invalid data, the following estimation method is used:

1. Use the average of the modeling (or training) images as an initial estimate of the albedo.

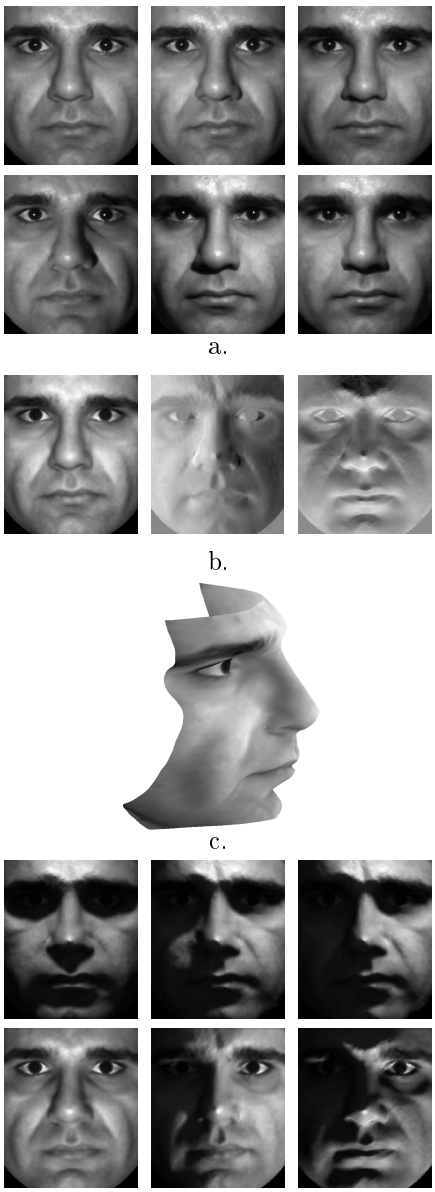


Figure 2: The process of constructing the cone $\mathcal{C}$. a. The training images; b. Images corresponding to columns of $\bar{B}$; c. Reconstruction up to a GBR transformation; d. Sample images from the illumination cone under novel lighting conditions in fixed pose.

2. Without doing any row or column permutations sift out all the full rows (with no invalid data) of matrix X to form a full sub-matrix $\tilde{X}$.
3. Perform SVD on $\tilde{X}$ to get an initial estimate of S.
4. Fix S and the albedo, and estimate a possibly non-integrable set $z_x^*(x,y)$ and $z_y^*(x,y)$ using least-squares.
5. By minimizing the cost functional in Equation 10, estimate (as functions of $\bar{c}(\mathbf{w})$) a set of integrable partial derivatives $\bar{z}_x(x,y)$ and $\bar{z}_y(x,y)$.
6. Fix S and use $\bar{z}_x(x,y)$ and $\bar{z}_y(x,y)$ to update the albedo using least-squares.
7. Use the newly calculated albedo and the partial derivatives $\bar{z}_x(x,y)$ and $\bar{z}_y(x,y)$ to construct $\bar{B}$.
8. Then, fix $\bar{B}$ and update each of the light source directions $\mathbf{s}_i$ independently using least-squares.
9. Repeat steps 4-8 until the estimates converge.
10. Perform inverse DCT on the coefficients $\bar{c}(\mathbf{w})$ to get the GBR surface $\bar{z}(x,y)$.

In our experiments, the algorithm is well behaved, provided the input data is well conditioned, and converges within 10-15 iterations.

Figure 2 demonstrates the process for constructing the illumination cone: Figure 2.a shows six of the 19 single light source images of a face used in the estimation of matrix $\bar{B}$. Note that the light source in each image moves only by a small amount ($\pm 15^o$ in either direction) about the viewing axis. Despite this, the images do exhibit some shadowing, e.g. left and right of the nose. Figure 2.b shows the basis images of the estimated matrix $\bar{B}$. These basis images encode not only the albedo (reflectance) of the face but also its surface normal field. They can be used to construct images of the face under arbitrary and quite extreme illumination conditions. Figure 2.c shows the reconstructed surface of the face $\bar{z}(x,y)$ up to a GBR transformation. The first basis image of matrix $\bar{B}$ shown in Figure 2.b has been texture-mapped on the surface.

Figure 2.d shows images of the face generated using the image formation model in Equation 1 which has been extended to account for cast shadows. To determine cast shadows, we employ ray-tracing that uses the reconstructed GBR surface of the face $\bar{z}(x,y)$. With this extended image formation model, the generated images exhibit realistic shading and, unlike the images in Figure 2.a, have strong attached and cast shadows.

2.3 Image Synthesis Under Differing Pose and Lighting

The reconstructed surface and the illumination cones can be combined to synthesize novel images of an object under differing pose and lighting. However, one complication arises because of the generalized bas-relief (GBR) ambiguity. Even though shadows are preserved under GBR transformations [2], without resolution of this ambiguity, images with non-frontal view-point synthesized from a GBR reconstruction will differ from a valid image by an affine warp of image coordinates. (It is affine because GBR is a 3-D affine transformation and the weak perspective imaging model assumed here is linear.) Since the affine warp is an image transformation, one could perform recognition over variation in viewing direction and affine image transformations. Alternatively, one can attempt to resolve the GBR ambiguity to obtain a Euclidean reconstruction using class-specific information. In our experiments with faces, we essentially try to fit the GBR reconstructions to a canonical face. We take advantage of the left-to-right symmetry of faces and the fairly constant ratios

Figure 3: Synthesized images under variable pose and lighting. The representation was constructed from the images in Figure 2.a.

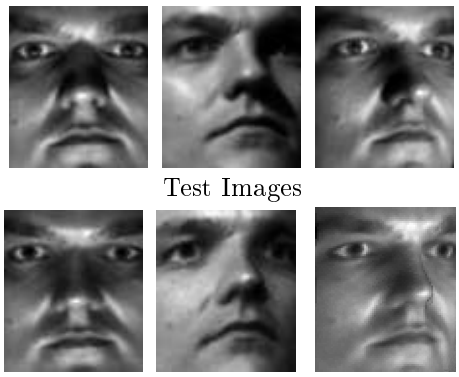
of distances between facial features such as the eyes, the nose, and the forehead to resolve the three parameters of the GBR ambiguity. Once resolved, it is a simple matter to use ray-tracing techniques to render synthetic images under variable pose and lighting.

Figure 3 shows synthetic images of the face under novel pose and lighting. These images were generated from the images in Fig. 2.a where the pose is fixed and there are only small, unknown variations in illumination. In contrast, the synthetic images exhibit not only large variations in pose but also a wide range in shading and shadowing.

3 Representations for Recognition

It is clear that for every pose of the object, the set of images under all lighting conditions is a convex cone. Therefore, the previous section provides a natural way for generating synthetic representations of objects suitable for recognition under variable pose and illumination. For every sample pose of the object, generate its illumination cone and with the union of all the cones form its representation.

However, the number of independent normals in B can be large (more than a thousand) hence the number of extreme rays needed to completely define the illumination cone can run in the millions (see Section 2). Therefore, we must approximate the cone in some fashion; in this work, we choose to use a small number of extreme rays (images). The hope is that a subsampled cone will provide an approximation that negligibly decreases recognition performance; in our experience, around 80 images are sufficient, provided that the corresponding light source directions $\mathbf{s}_{ij}$ are more or less uniform on the illumination sphere. The resulting cone $\mathcal{C}^*$ is a subset of the object's true cone $\mathcal{C}$ for a particular pose.



Test Images

Closest image in cone representation

Figure 4: TOP ROW: Three images from the test set. BOTTOM ROW: The closest reconstructed image from the representation. Note that these images are not explicitly stored, but lie within the closest matching linear subspace.

Another simplifying factor that can reduce the size of the representation is the assumption of a weak perspective imaging model. Under this model, the effect of pose variation can be decoupled into that due to image plane translation, rotation, and scaling (a similarity transformation), and that due to the viewpoint direction. Within a face recognition system, the face detection process generally provides estimates for the image plane transformations. Neglecting the effects of occlusion or appearance of surface points, the variation due to viewpoint can be seen as a non-linear warp of the image coordinates with only two degrees of freedom.

Yet, recognition using this representation consisting of sub-sampled illumination cones will still be costly since computing distance to a cone is $O(n\,e^2)$, where n is the number of pixels and e is the number of extreme rays (images). From an empirical study, it was conjectured in [1] that the cone for typical objects is flat (i.e., all points lie near a low-dimensional linear subspace), and this was confirmed for faces in [5]. Hence, an alternative is to model a face in fixed pose but over all lighting conditions by a low-dimensional linear subspace. Finally, for a set of sample viewing directions, we construct subspaces which approximate the corresponding cones. We chose to use an 11-D linear subspace for each pose since 11 dimensions capture over 99% of the variation in the sample extreme rays. Recognition of a test image $\mathbf{x}$ is then performed by finding the closest linear subspace to $\mathbf{x}$. Figure 4 shows the closest match for images of an individual in three poses. This figure qualitatively demonstrates how well the union of 11-D subspaces approximates the true cones.

For the experimental results reported below, subspaces were constructed by sampling the viewing sphere at 4° intervals over the elevation from $-24°$ to $+24°$ and the azimuth from $-4°$ to $+28°$ about frontal. As a final speed-up, the 117 11-D linear subspaces were projected down to a 100-dimensional subspace of the image space whose basis vectors were computed using

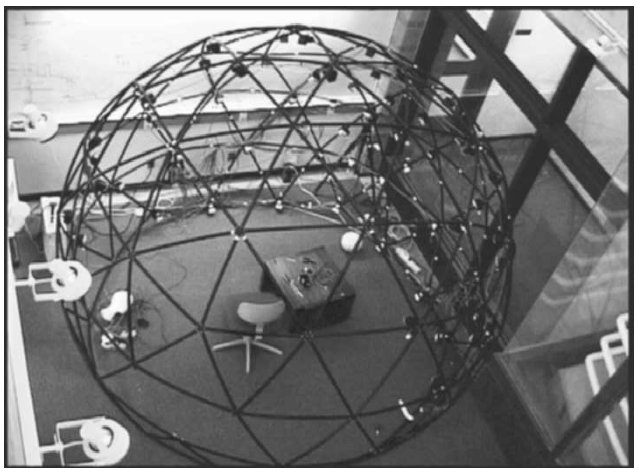


Figure 5: A geodesic dome with 64 strobes used to gather images under variable illumination and pose.

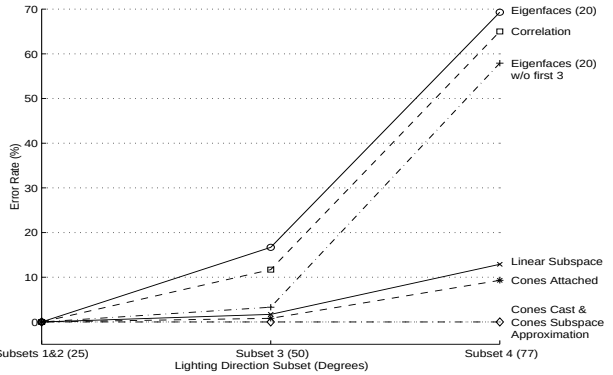
SVD. In summary, each person's face was represented by the union of 117 11-D linear subspaces within a 100-dimensional subspace of the image space. Recognition was then performed by computing the distance of a test image to each 100-D subspace plus the distance to the 11-D subspaces within the 100-D space.

4 Recognition Results

The experimentation reported here was performed on the Yale Face Database B. For capturing this database, we have constructed a geodesic lighting rig with 64 computer controlled xenon strobes shown in Figure 5. With this rig, we can modify the illumination at frame rates and capture images under variable pose and illumination. Images of ten individuals were acquired under 64 different lighting conditions in nine poses (frontal pose, five poses at 12° and three poses at 24° from the camera's axis). Of the 64 images per person in each pose, 45 were used in our experiments, a total of 4050 images. The images from each pose were divided into 4 subsets (12°, 25°, 50° and 77°) according to the angle of the light source with the camera's axis (see Figure 1). Subset 1 (respectively 2, 3, 4) contains 70 (respectively 120, 120, 140) images per pose. Throughout, the 19 images of Subsets 1 and 2 from the frontal pose of each face were used as training images for generating its representation.

4.1 Extrapolation in Illumination

The first set of experiments was performed under fixed pose on the 450 images from the frontal pose (45 per person). This was to compare three other recognition methods to the illumination cones representation. From a set of face images labeled with the person's identity (*the learning set*) and an unlabeled set of face images from the same group of people (*the test set*), each algorithm is used to identify the person in the test images. For more details about the comparison algorithms, see [3] and [7]. We assume that each face has been located and aligned within the image.



Extrapolation in Illumination

Method	Error Rate (%) vs. Illumination		
	Subsets 1 & 2	Subset 3	Subset 4
Correlation	0.0	11.7	65.0
Eigenfaces	0.0	16.7	69.3
Eigenfaces w/o 1st 3	0.0	3.3	57.9
Linear subspace	0.0	1.7	12.9
Cones-attached	0.0	0.8	9.3
Cones-cast (Subspace Approx.)	0.0	0.0	0.0
Cones-cast	0.0	0.0	0.0

Figure 6: **Extrapolation in Illumination:** Each of the methods is trained on images with near frontal illumination (Subsets 1 and 2) from Pose 1 (frontal pose). This graph shows the error rates under more extreme light source conditions in fixed pose.

The simplest recognition scheme is a nearest neighbor classifier in the image space [4]. An image in the test set is recognized (classified) by assigning to it the label of the closest point in the learning set, where distances are measured in the image space. When all of the images are normalized to have zero mean and unit variance, this procedure is also known as Correlation.

A technique now commonly used in computer vision—particularly in face recognition—is principal components analysis (PCA) which is popularly known as *Eigenfaces* [8, 12, 13, 19]. One proposed method for handling illumination variation in PCA is to discard the three most significant principal components; in practice, this yields better recognition performance [3]. For both the Eigenfaces and Correlation tests, the images were normalized to have zero mean and unit variance, as this improved the performance of these methods. This also made their results independent of light source intensity. For the Eigenfaces method, we used 20 principal components; recall that performance approaches correlation as the dimension of the feature space is increased [3, 13]. Error rates are also presented when the principal components four through twenty-three were used.

A third approach is to model the illumination variation of each face with the three-dimensional linear subspace $\mathcal{L}$ described in Section 2.1. To perform recogni-

tion, we simply compute the distance of the test image to each linear subspace and choose the face corresponding to the shortest distance. We call this recognition scheme the *Linear Subspace* method [2]; it is a variant of the photometric alignment method proposed in [16] and is related to [9, 14]. While this models the variation in image intensities when the surface is completely illuminated, it does not model shadowing.

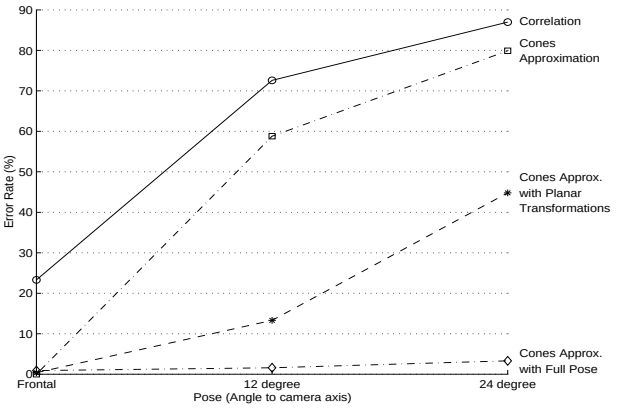
Finally, recognition is performed using the illumination cone representation. In fact, we tested on three variations. In the first (Cones-attached), the representation was constructed without cast shadows, so the extreme rays are generated directly from Equation 3. In the second variation (Cones-cast), the representation was constructed as described in Section 2.2 where we employed ray-tracing that uses the reconstructed surface of a face $\bar{z}(x, y)$ to determine cast shadows. In both variations, recognition was performed by computing the distance of the test image to each cone and choosing the face corresponding to the shortest distance. Since cones are convex, the distance can be found by solving a convex optimization problem (see [7]).

In the last variation, the illumination cone of each face with cast shadows $\mathcal{C}^*$ is approximated by an 11-D dimensional linear subspace (Cones-cast subspace approximation). As mentioned before, it was empirically determined that 11 dimensions capture over 99% of the variance in the sample extreme rays. The basis vectors for this space are determined by performing SVD on the extreme rays in $\mathcal{C}^*$ and then picking the 11 eigenvectors associated with the largest singular values. Recognition was performed by computing the distance of the test image to each linear subspace and choosing the face corresponding to the shortest distance. Using the cone subspace approximation reduces both the storage and the computational time. Since the basis vectors of each subspace are orthogonal the computational complexity is only $O(n\,m)$ where n is the number of pixels and m is the number of the basis vectors.

Similar to the extrapolation experiment described in [3], each method was trained on samples from Subsets 1 and 2 (19 samples per person) and then tested on samples from Subsets 3 and 4. Figure 6 shows the results from this experiment. (This test was also performed on the Harvard Robotics Lab face database and was reported in [7].) Note that the cone subspace approximation performed as well as the raw illumination cones without any mistakes on 450 images. This supports the use of low dimensional subspaces in the full representation of Section 3 that models image variations due to viewing direction and lighting.

4.2 Recognition Under Variable Pose and Illumination

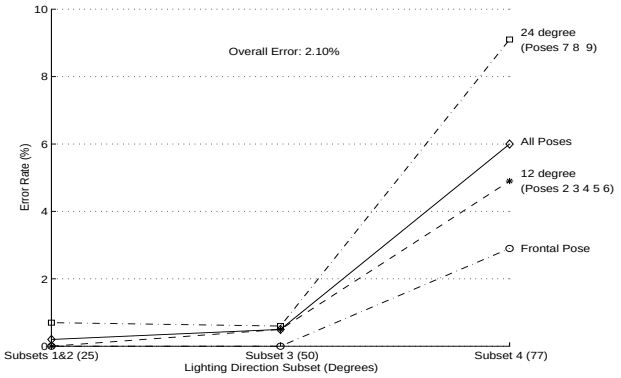
Next, we performed recognition experiments on images in which the pose varies as well as illumination. Images from all nine poses in the database were used in these tests. Four recognition methods were compared on 4050 images. Each method was trained on images with near frontal illumination (Subsets 1 and 2) from the frontal pose, and tested on all images from all nine poses—an extrapolation in both pose and illumination.



Extrapolation in Pose			
Method	Error Rate (%) vs. Pose		
	Frontal (Pose 1)	12° (Poses 2 3 4 5 6)	24° (Poses 7 8 9)
Correlation	23.3	72.6	87.0
Cones Approximation	0.0	58.8	79.9
Cones Approx. with Planar Transformations	0.4	13.3	44.8
Cones Approx. with Full Pose	0.9	1.6	3.3

Figure 7: **Extrapolation in Pose:** Error rates as the viewing direction becomes more extreme. Again, the methods were trained on images with near frontal illumination (Subsets 1 and 2) from Pose 1 (frontal pose). Note that each reported error rate is for *all* illumination subsets (1 through 4).

The first method was Correlation as described in the previous section. The next one (Cones approximation) modeled a face with an 11-D subspace approximation of the cone (with cast shadows) in the frontal pose. No effort was done to accommodate pose during recognition, not even a search in image plane transformations. The next method (Cones approximation with planar transformations) also modeled a face with an 11-D subspace approximation of the cone in the frontal pose, but unlike the previous method, recognition was performed over variations of planar transformations. Finally, a face was modeled with the representation described in Section 3. Each of the 10 individuals was represented by a 100-D subspace which contained 117 11-D linear subspaces each modeling the variation in illumination for each sampled view-point. As with the previous method, recognition was performed over a variation of planar transformations. The results of these experiments are shown in Figure 7. Note that each reported error rate is for *all* illumination subsets (1 through 4). Figure 8, on the other hand, shows the break-down of the results of the last method for different poses against



Error Rates (%)			
	Lighting Variation		
Pose	Subsets 1 & 2	Subset 3	Subset 4
Frontal (Pose 1)	0.0	0.0	2.9
12° (Poses 2 3 4 5 6)	0.0	0.5	4.9
24° (Poses 7 8 9)	0.7	0.6	9.1
All Poses	0.2	0.5	6.0

Figure 8: Error rates for different poses against variable lighting using the representation of Section 3.

variable illumination. As demonstrated in Figure 7, the method of cone subspace approximation with planar transformations performs reasonably well for poses up to 12° from the viewing axis but fails when the viewpoint becomes more extreme.

We note that in the last two methods the search in planar transformations did not include image rotations (only translations and scale) to reduce computational time. We believe that the results would improve if image rotations were included or even if the view-point space and illumination cones were more densely sampled and the 11-D subspaces were not projected down to a 100-D subspace.

5 Discussion

In constructing the representation of an object from a small set of training images, we have assumed that the object's surface exhibited a Lambertian reflectance function. Although our results support this assumption, more complex reflectance functions may yield better recognition results. Other exciting domains for these representations include facial expression recognition and object recognition with occlusions.

References

[1] P. Belhumeur and D. Kriegman. What is the set of images of an object under all possible illumination conditions. *Int. J. Computer Vision*, 28(3):245–260, July 1998.

[2] P. Belhumeur, D. Kriegman, and A. Yuille. The bas-relief ambiguity. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 1040–1046, 1997.

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 19(7):711–720, 1997. Special Issue on Face Recognition.

[4] R. Brunelli and T. Poggio. Face recognition: Features vs templates. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 15(10):1042–1053, 1993.

[5] R. Epstein, P. Hallinan, and A. Yuille. 5+/-2 eigenimages suffice: An empirical investigation of low-dimensional lighting models. In *Physics Based Modeling Workshop in Computer Vision*, Session 4, 1995.

[6] R. T. Frankot and R. Chellapa. A method for enforcing integrabilty in shape from shading algorithms. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 10(4):439–451, 1988.

[7] A. Georghiades, D. Kriegman, and P. Belhumeur. Illumination cones for recognition under variable lighting: Faces. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 52–59, 1998.

[8] P. Hallinan. A low-dimensional representation of human faces for arbitrary lighting conditions. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 995–999, 1994.

[9] P. Hallinan. *A Deformable Model for Face Recognition Under Arbitrary Lighting Conditions.* PhD thesis, Harvard University, 1995.

[10] H. Hayakawa. Photometric stereo under a light-source with arbitrary motion. *J. Opt. Soc. Am. A*, 11(11):3079–3089, Nov. 1994.

[11] D. Jacobs. Linear fitting with missing data: Applications to structure from motion and characterizing intensity images. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, 1997.

[12] L. Sirovitch and M. Kirby. Low-dimensional procedure for the characterization of human faces. *J. Optical Soc. of America A*, 2:519–524, 1987.

[13] H. Murase and S. Nayar. Visual learning and recognition of 3-D objects from appearence. *Int. J. Computer Vision*, 14(5–24), 1995.

[14] S. Nayar and H. Murase. Dimensionality of illumination manifolds in appearance matching. In *Int. Workshop on Object Representations for Computer Vision*, page 165, 1996.

[15] A. Shashua. *Geometry and Photometry in 3D Visual Recognition.* PhD thesis, MIT, 1992.

[16] A. Shashua. On photometric issues to feature-based object recognition. *Int. J. Computer Vision*, 21:99–122, 1997.

[17] H. Shum, K. Ikeuchi, and R. Reddy. Principal component analysis with missing data and its application to polyhedral object modeling. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 17(9):854–867, September 1995.

[18] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *Int. J. Computer Vision*, 9(2):137–154, 1992.

[19] M. Turk and A. Pentland. Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3(1):71–96, 1991.