INVITED PAPER

Face Recognition Using 3-D Models: Pose and Illumination

Novel face recognition algorithms, based on three-dimensional information, promise to improve automated face recognition by dealing with different viewing and lighting conditions.

By Sami Romdhani, Jeffrey Ho, *Member IEEE*,
Thomas Vetter, *Member IEEE*, and David J. Kriegman

ABSTRACT | Unconstrained illumination and pose variation lead to significant variation in the photographs of faces and constitute a major hurdle preventing the widespread use of face recognition systems. The challenge is to generalize from a limited number of images of an individual to a broad range of conditions. Recently, advances in modeling the effects of illumination and pose have been accomplished using three-dimensional (3-D) shape information coupled with reflectance models. Notable developments in understanding the effects of illumination include the nonexistence of illumination invariants, a characterization of the set of images of objects in fixed pose under variable illumination (the illumination cone), and the introduction of spherical harmonics and low-dimensional linear subspaces for modeling illumination. To generalize to novel conditions, either multiple images must be available to reconstruct 3-D shape or, if only a single image is accessible, prior information about the 3-D shape and appearance of faces in general must be used. The 3-D Morphable Model was introduced as a generative model to predict the appearances of an individual while using a statistical prior on shape and texture allowing its parameters to be estimated from single image. Based on these new understandings, face recognition algorithms have been developed to address the joint challenges of pose and lighting. In this paper, we review these developments and provide a brief survey of the resulting face recognition algorithms and their performance.

KEYWORDS | Biometrics; computer vision; face image analysis; face recognition; illumination invariant; inverse rendering; pose invariant

I. INTRODUCTION

The goal of face recognition is to identify individuals in photographs or videos from their facial appearance. Compared to other biometrics, face recognition is passive and does not require cooperative subjects who are near or in contact with a sensor. Images of faces are widespread and archived, and digital cameras are so inexpensive that they are embedded in inexpensive consumer devices (e.g., mobile phones). For video surveillence, a camera may be located in the corner of a room, and the goal would be to identify the occupants without their awareness. Yet, to achieve this capability, face recognition systems must be effective irrespective of the person's gaze or the illumination.

Face recognition has numerous applications including access control, human computer interfaces, security and surveillance, e-commerce, entertainment, annotation of photographic and video databases, etc. Consequently, it has been an attractive research problem, and the most recent comprehensive survey [56] cites 168 papers while a survey of face detection [49]—a preprocessing step of face recognition—cites over 180 sources.

Yet despite this enormous effort, accurate and robust recognition over a broad range of conditions has remained elusive. A recent large scale evaluation of commercial face recognition systems called the Facial Recognition Vendor Test (FRVT) 2002 [31] showed that face recognition and verification accuracy deteriorated significantly when there were differences in pose and lighting between images used for enrollment and recognition, and that errors increased as the elapsed time between enrollment and recognition increased [31]. Why has face recognition proven to be so

Manuscript received February 16, 2006; revised August 11, 2006.

S. Romdhani and **T. Vetter** are with the Computer Science Department, University of Basel, CH-4056 Basel, Switzerland (e-mail: sami.romdhani@unibas.ch; thomas.vetter@unibas.ch).
J. Ho is with the Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: jho@cise.ufl.edu).
D. J. Kriegman is with the Department of Computer Science, University of California at San Diego, La Jolla, CA 92093 USA (e-mail: kriegman@cs.ucsd.edu).

Digital Object Identifier: 10.1109/JPROC.2006.886019

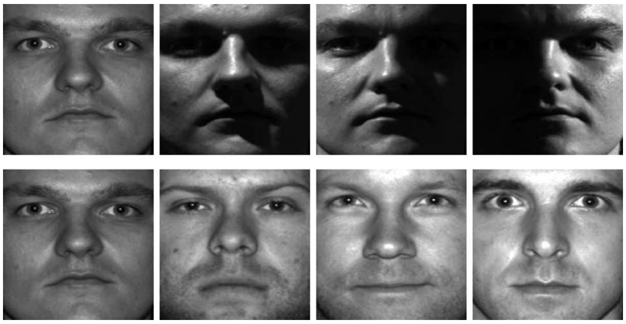


Fig. 1. ***Striking effect of illumination on the appearances of a human face. Top row: images of the same person taken from the same viewpoint but under different illumination. Bottom row: images of four different individuals taken under the same viewpoint and lighting.***

challenging? In part, it is because an individual's face can appear differently in different images yet many people may look somewhat similar to each other under some conditions (identical twins are an extreme case). Facial expression, makeup, eyeglasses, facial hair, weight change, and aging are intrinsic factors that lead to differences in appearance. On the other hand, extrinsic factors such as illumination (source brightness, direction, color), camera viewpoint, and the camera's radiometric response can lead to significant image variability. It is our belief that, to reach unconstrained and accurate face recognition, the fewest limiting assumptions about the problem should be made. It is on these foundations that the systems presented in this article were developed.

Consider for the moment the challenge of illumination variation; Fig. 1 shows that the effect of lighting on the appearance of a human face can be striking. The top row shows four images of an individual taken with the same viewpoint but under differing illumination. The bottom four images, on the other hand, are of four individuals taken under the same lighting and from the same viewpoint. Using the most common measure of similarity between pairs of images, the L^2-difference (i.e., the sum over all pixel locations of the squared difference between pixel values), the similarity between any pair of images in the bottom is always greater than the similarity between any pair of images from the upper row. In other words, face recognition based purely on L^2-similarity (i.e., template matching) will fail for these images. The same remark can also be made for viewpoint changes.

While there are exceptions, the lion's share of face recognition methods are purely two-dimensional (2-D) in the sense that subjects are enrolled into the system using one or more 2-D images, and at recognition time the input is again one or more 2-D images. At no point in the recognition process is a three-dimensional (3-D) model of a face constructed, nor do the algorithms make strong and explicit use of the fact that the images are the result of observing a 3-D object. Instead, they rely on the power of pattern recognition and machine learning techniques to essentially infer from training examples how an individual's appearance may vary and how features of the images can discriminate between individuals. Yet it is our contention that accurate face recognition over the kinds of variation described above can be achieved with a 3-D generative model that can be used to synthesize or render an individual's appearance. In turn, classifiers can be constructed based on estimated parameters of the model, using images or representations derived from a collection of such models, or by estimating extrinsic and intrinsic parameters such that a rendered model is consistent with an imaged face.

We would like to stress that this is not a comprehensive survey article about face recognition algorithms. See [56] for an excellent survey paper reviewing the major problems and solutions. This paper is limited to a few of the most promising methods that address the illumination and combined pose/illumination problems. We will provide some detail and give possible reasons for the success of these methods. A growing area of research in face recognition is the identification of people from a 3-D range data [9], [11]. These methods are not reviewed in this paper, which focuses on recognition from 2-D photographs.

The reader should also be aware that it is not straightforward to compare the performance of face recognition algorithms based on the published results. The standard protocol [31] in evaluating face recognition algorithms requires three separate sets of images: training, gallery, and probe sets. The training set is for learning whereas the gallery and probe sets are used for testing the recognition algorithm. The gallery set contains images with known identities while the identities in the probe set are unknown. In principle, there should be no overlap between training and testing images not only in terms of identity but also in terms of illumination condition, pose, and acquisition device. Ideally, to ensure that the algorithm is not tuned to any specific condition, training and testing image sets should originate from different and independent research institutes. Furthermore, as the complexity of the identification depends on the number of individuals in the gallery and probe sets, this number should be large. Unfortunately, due to various difficulties, these practices are seldom observed, and this makes the comparison of published algorithms and results often difficult and complicated.

A. Organization of the Paper

In this paper, we discuss the state of the art in face recognition over pose and lighting variation. We introduce recent advances in modeling illumination effects [2], [4], [34] and the most promising face recognition algorithms based on these foundational results [2], [12], [17], [28], [42], [52]. We describe how some of these techniques can be generalized to handle both pose and lighting. We describe the 3-D Morphable Model (3DMM), a parametric model of 3-D face shape and RGB texture that can be used to predict the appearances of an individual from an arbitrary viewpoint, illuminated by one arbitrary directed light source along with ambient light [7], [8], [36]–[38]. Using a statistical prior on shape and texture, the 3DMM parameters for an individual can be estimated from single image, and in turn these parameters are the basis for recognition decisions. We present and compare empirical results showing the effectiveness of these methods on face recognition over lighting and pose variation.

II. MODELS OF ILLUMINATION AND REFLECTANCE

The effect of lighting on face images is dramatically illustrated in Fig. 1, and to understand this effect we must model how the environmental illumination strikes the face, how light is reflected from the face, and how shadowing occurs. In general, these are modeled separately based on physically reasonable assumptions.

First, it should be noted that lighting variation is more than simply differences in overall brightness as the strength of the illuminant may vary with direction. It can be modeled as a positive function over the 4-D space of light rays that are incident to the face. For recognition, faces are generally far from the light sources. This allows us to treat light source strength as a function of directions (i.e., a positive function on the sphere), and a single distant light source would then be considered as a delta function.

In general, surface reflectance can be described by a 4-D function $f_r(\theta_i, \phi_i, \theta_o, \phi_o)$ called the bidirectional reflectance distribution function (BRDF), and it gives the reflectance of each point on a surface as a function of the incident illumination direction $\omega_i = (\theta_i, \phi_i)$ and the emitted direction $\omega_o = (\theta_o, \phi_o)$ (see Fig. 2). Specific reflectance models for human skin [23] and hair [30] have been developed, and have lead to very compelling renderings for motion pictures. To more fully model the interaction of light with a face, one might also want to include the effects of subsurface scattering [24], translucency, and interreflections between, say, the nose and the cheek; while these effects lead to greater realism in rendering, they have not yet been considered significant enough for face recognition. Consequently, much simpler reflectance models, such as the Lambertian and Phong models, have been shown to be effective for face recognition, and we will only consider these reflectance models in this paper [4], [15], [17].

Under the Lambertian model, the pixel intensity I of each surface point is given by the inner product between the unit surface normal vector $\vec{n}$ scaled by the *albedo* value ρ and the light vector $\vec{l}$, which encodes the direction and strength of incident light from a single, distant source

$$I(\vec{l}) = \rho \max(\vec{l} \cdot \vec{n}, 0). \tag{1}$$

The Lambertian model effectively collapses the 4-D BRDF f_r into a constant function with value ρ. In particular, the brightness of a Lambertian surface does not depend upon the viewing direction, and so it appears equally bright from all viewing directions.

The Phong model extends the Lambertian model and accounts for specular highlights and ambient illumination, as well as diffuse reflection. Specular highlights arise only for certain viewing directions that depend on the normal and light direction. The specular color does not depend on the albedo of the surface, but only on the color of the light. Under the Phong model, the radiance of a point illuminated with ambient intensity a, and viewed from the direction $\vec{v}$ is given by

$$I(\vec{l}) = a\rho + \rho \max(\vec{l} \cdot \vec{n}, 0) + s(\vec{l}, \vec{n}, \vec{v}). \tag{2}$$

In this equation, $s(\vec{l}, \vec{n}, \vec{v})$ is a function modeling specular highlights, and it is defined in Section V-B.

For a single light source, two types of shadows can appear: attached shadows and cast shadows (Fig. 2). A surface point is within an attached shadow when the inward facing surface normal $\vec{n}$ points away from the light source. This condition can be summarized concisely as

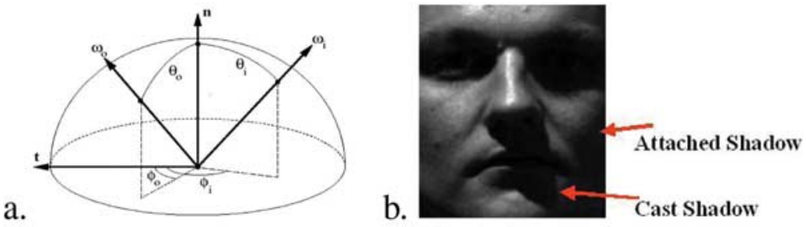


Fig. 2. ***(a) Coordinate system used in defining the BRDF.*** $\omega_i = (\theta_i, \phi_i)$ ***parameterizes the incident lighting direction.*** $\omega_o = (\theta_o, \phi_o)$ ***represents the viewing direction (emitted direction).*** n ***is the normal vector. (b) The formation of shadows on a human face. Attached shadows are in the upper region of the eye socket. Cast shadows appear in the lower region of the eye socket and the lower part of the face.***

$\vec{n} \cdot \vec{l} < 0$. Cast shadows occur when another part of the object occludes the light source (e.g., the shadow that the nose casts onto the cheek). A point p is in cast shadow if the line segment (or ray) from p to the light source intersects the surface. While attached shadows are related to the local geometry, cast shadows are related to the object's global geometry.

The above discussion has focussed on a single distant illuminant $\vec{l}$, and when multiple sources are present, the resulting image is the sum (superposition) of the images produced by the sources individually.

III. ILLUMINATION: THEORY AND FOUNDATIONAL RESULTS

Before delving into the details of various face recognition algorithms, we discuss briefly an interesting result [12] on the nonexistence of illumination invariants.

A. Nonexistence of Illumination Invariants

Consider the following simple question. You are given an image of an object (a face). Now given a second image, can you determine whether this is an image of the same object (face) in the same pose, but under different lighting, or a completely different object? Counterintuitive to our daily experience, [12] demonstrated that for any two images, whether they are of the same object or not, there is always a family of Lambertian surfaces, albedo patterns, and light sources that could have produced them. As a consequence, given two images, it is not possible with absolute certainty to determine whether they were created by the same or different objects.

For face recognition, this negative result, however, is not as devastating as one may have thought, and there are at least two ways of avoiding this apparent quandary. First, while determining whether two images are of the same object is impossible in principle, nothing prevents us from using three or more images, perhaps by increasing the number of training images for each person. Second, this result can be attributed to the unrestricted access to the space of Lambertian objects, and the Lambertian surface accounting for two specific images may be rather bizarre. For example, given an image of Katharine Hepburn and one of Humphrey Bogart, along with any pair of light source directions, there exists a Lambertian surface that could have produced these images. However, it is unlikely to be face-like. This observation is operationalized by using a deterministic prior on the surface shape so that only face-like surfaces are permitted or a probabilistic prior favoring face-like surfaces.

The success of these two approaches depends critically on the ability to acquire an efficient and effective appearance model that can capture good amount of variability of images under all possible conditions. In particular, for an object $\mathcal{O}$ such as a face, we would like to know the set $\mathcal{C}$ of images under all possible conditions. In principle, the problem of recognition becomes rather easy if $\mathcal{C}$ is known: given an input image, simply determine which person's set contains the image. To realize this, one would need a model of a face from which to construct such a set and the means to determine containment. In this section, we will focus particularly on illumination variation and assume that all images were taken from the same viewpoint, e.g., frontal pose. Considering all images to have n pixels, we can regard $\mathcal{C}$ as a subset of the image space $\mathbb{R}^n$. In the rest of this section, we discuss the theoretical results of [2], [4], [34], which give various characterizations of the set $\mathcal{C}$ when the object $\mathcal{O}$ is Lambertian and convex. There are two main themes, the effective low dimensionality of $\mathcal{C}$ and its linearity. While faces are neither convex nor Lambertian, the theory is tractable in this case, and algorithms based on these results have proven to be effective.

We first establish a few conventions and notations. Interreflections will be ignored, and all illumination will be assumed to be generated by distant sources. In particular, the source is represented as a 3-vector $\vec{l}$ such that $|\vec{l}\,|$ encodes the strength of the source, and the unit vector $\vec{l}/|\vec{l}\,|$ represents its direction.

B. Illumination Cones

Under the above assumptions, it was shown in [4] that the set of n-pixel images $\mathcal{C}$ of an object in fixed pose under all lighting conditions is a convex cone in $\mathbb{R}^n$. Recall that a cone in $\mathbb{R}^n$ is simply a convex subset of $\mathbb{R}^n$ that is invariant under nonnegative scalings: if x is in the cone, then λx is also in the cone for any nonnegative λ. A polyhedral cone is simply a cone with a *finite* number of generators $\{e_1, \ldots, e_g\}$: points of the cone are vectors $x \in \mathbb{R}^n$ that can be expressed as some nonnegative linear combination of the generators, $x = a_1 e_1 + \cdots + a_g e_g$ with $a_1, \ldots, a_g \geq 0$.

Convexity is a simple consequence of the superposition principle for illumination. If I_1 and I_2 are two images taken under two different illumination conditions l_1 and l_2, any convex combination of these two images

$$J = aI_1 + bI_2, \qquad a, b \geq 0, \quad a + b = 1$$

is also an image of the same object under a new illumination condition specified by $al_1 \cup bl_2$, i.e., l_1 and l_2 are "turned on" simultaneously with attenuation factors a, b, respectively.

This is the simplest and also the only characterization of $\mathcal{C}$ without any limiting assumptions on reflectance or geometry. When the object is convex, Lambertian and considered to have a single normal projecting to each pixel, the set $\mathcal{C}$ of images of an object under all possible illumination conditions is a convex polyhedral cone, and an upper bound on the number of generators (extreme rays) is $m(m-1)$ where m is the number of distinct surface normals [4].

We now sketch some of the ideas behind these results and refer the reader to [4], [22] for more details. While (1) applies to a single surface point, we now consider the image as a whole. Let $B \in \mathbb{R}^{3\times n}$ be a matrix where each row of B is the product of the albedo with the inward pointing unit normal for a point on the surface projecting to a particular pixel. For a light source $\vec{l}$ that does not produce shadowing, the resulting image is given by $I = B\vec{l}$, and the set of images without shadowing is a subset of a 3-D linear subspace called the illumination subspace $\mathcal{L}$ [39], where

$$\mathcal{L} = \{I | I = Bl, \forall l \in \mathbb{R}^3\}. \tag{3}$$

When attached shadows are considered, the set of images under a single distant source is given by $\mathcal{U} = \{I | I = \max(Bl, 0), \forall l \in \mathbb{R}^3\}$. When multiple light sources illuminate the object, the resulting image is some convex combination of the images produced by the individual lights, and in turn the set of images $\mathcal{C}$ is the convex hull of $\mathcal{U}$. It is important to note that single light source images are on the boundary of $\mathcal{C}$, and in general images produced with multiple or diffuse light sources will lie in the interior of $\mathcal{C}$; hence, recognition in single source conditions is generally more difficult. This partially justifies why research efforts focussing on lighting have used single source conditions whereas diffuse lighting is usually used in controlled operational conditions. However, [1] has initiated an important study on face recognition under multiple illumination sources, and significant results have been obtained therein.

Since the illumination cone $\mathcal{C}$ is completely determined by the illumination subspace $\mathcal{L}$, $\mathcal{C}$ can be determined uniquely if the matrix B (surface normals scaled by albedo) were known. The method of uncalibrated photometric stereo [51], [58] takes three or more images of the same object viewed from the same pose and under unknown and different lighting as input and allows us to recover B up to an invertible 3×3 linear transformation $A \in GL(3)$ since $B\vec{l} = (BA)(A^{-1}\vec{l}) = B^*\vec{l}^*$ where $B^* = BA$ and $\vec{l}^* = A^{-1}\vec{l}^*$. Although B is not uniquely determined, it is easy to see that B and B^* span the same illumination subspace, and hence, the same illumination cone.

It can also be shown that the dimension of $\mathcal{C}$ is equal to the number of distinct surface normals [4]. For images with n pixels, this indicates that the dimension of the illumination cone is one for a planar object, is roughly $\sqrt{n}$ for a cylindrical object, and is n for a spherical object. It is to be noted, however, that having a cone span n dimensions does not mean that it covers $\mathbb{R}^n$, and as we shall see there is sound empirical and theoretical evidence to support the idea that the illumination cone is flat; that is, it can be well-approximated by a low-dimensional linear subspace.

C. Empirical Observations

The fact that for objects with diffuse, Lambertian-like reflectance, the effective dimension of $\mathcal{C}$ is small was noticed quite early [14], [21]. This can be demonstrated by collecting images of an object taken under a number of different illumination conditions. If $\{I_1, \ldots, I_m\}$ are m such vectorized images, we can form an intensity matrix as $I = [I_1 \cdots I_m]$. Singular value decomposition (SVD) of I [18] $I = U\Sigma V^t$, gives the singular vectors as the columns of the matrix U, and the diagonal elements of Σ as the singular values. Let $\{\sigma_1, \ldots, \sigma_m\}$ denote the singular values in descending order. The first k singular values indicate the least squares approximation accuracy of the best fitting k-dimensional subspace spanned by the corresponding k singular vectors. Principal component analysis (PCA) [43], [45] is another commonly used technique for linear dimensional analysis. Instead of the intensity matrix, PCA computes the eigenvalues and eigenvectors of the scatter matrix of the images. In computer vision literature, the singular vectors in SVD and the eigenvectors in PCA are usually called Eigenimages, and in the case of face images they are called, appropriately, *Eigenfaces*.

Fig. 3 displays the magnitude of the first ten singular values obtained by applying SVD to a collection of 45 image of one human face (in frontal pose and under 45 different point light sources) shown in Fig. 9. The magnitude of the singular values decreases rapidly after the first three singular values. In [14], PCA was applied to images of non-Lambertian objects. The conclusion from this empirical study is surprising in that 5 ± 2 Eigenimages are sufficient to model the effect of lighting for objects with a wide range of reflectance properties.

We remark that linear algebraic techniques, such as matrix factorization, SVD and PCA, have been the dominating and preferred mathematical tools in modelling

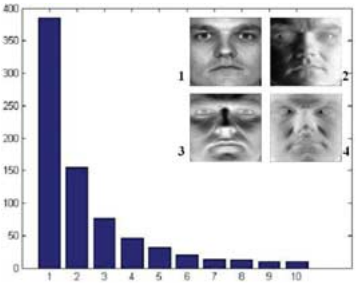


Fig. 3. ***The magnitudes of first ten singular values for 45 images in fixed pose under differing lighting of the person shown Fig. 9. In this example, the first three eigenvalues account for more than 97% of the energy. The four Eigenfaces corresponding to the largest four eigenvalues are also displayed.***

and analyzing illumination effects for a fixed viewpoint and object. This can be partially attributed to the fact that many illumination models are linear. The sources of this apparent pervasive linearity are the superposition principle of illumination and the quasi-linear nature of the Lambertian reflectance model. To analyze the illumination effects across a range of different objects and viewpoints, multilinear techniques have become popular recently [27], [58], [59]. In analyzing the illumination effects for a fixed viewpoint and object, we have stacked the vectorized images horizontally $\{I_1, \ldots, I_m\}$ to form the intensity matrix I. In other words, the data (intensity values) are now indexed by two integers I_{ij}, where i indexes the pixels and j indexes the illumination conditions. To incorporate other variations into the model such as pose variation, we need to stack the vectorized images in different directions. The resulting mathematical structure that encodes the intensity values is a tensor of rank greater than two. In particular, for a collection of images that contains both pose and illumination variations, the intensity values I_{ijk} form a tensor of rank 3, where i indexes the pixels as before, and j and k index the illumination and pose variations, respectively. Using tensorial representations, higher order singular value decomposition and other tensor decomposition results can be applied to compute linear subspaces. We refer the reader to [27], [58], [59] for more details.

D. Modeling Reflectance and Illumination Using Spherical Harmonics

The effective low dimensionality of $\mathcal{C}$ that we have just discussed clearly begs for explanations. This empirical observation can be elegantly explained via a signal processing framework using spherical harmonics [2], [33], [34]. The key conceptual advance is to treat a Lambertian object as a "low-pass filter" that turns complex (high frequency) external illumination into a smoothly shaded image. In the context of illumination, the signals (the illumination and BRDF) are functions defined on the sphere, and spherical harmonics are the analogue of the Fourier basis functions.

Spherical harmonics, Y_{lm}, are a set of functions that form an orthonormal basis for the set of all square-integrable (L^2) functions defined on the unit sphere. Y_{lm}, indexed by two integers l (degree) and m (order) obeying $l \geq 0$ and $-l \leq m \leq l$, has the following form:

$$Y_{lm}(\theta, \phi) = N_{lm} P_l^{|m|}(\cos\theta) e^{im\phi} \tag{4}$$

where N_{lm} is a normalization constant, and $P_l^{|m|}$ are the Legendre polynomials. Illumination can then be expressed in terms of this basis as $L(\theta, \phi) = \sum_{l=0}^{k} \sum_{m=-l}^{l} a_{lm} Y_{lm}(\theta, \phi)$. Since the spherical harmonics are complex-valued functions while the illumination function L is real-valued, in practice, the expansion above is computed using real-valued basis functions Y'_{lm}, which are the real and imaginary components of the spherical harmonics. Conveniently, each Y'_{lm} can be written as a polynomial of degree l in the usual Cartesian coordinates x, y, z.

As shown in [2], [34], Lambertian reflectance can be cast as a linear filter whose truncated cosine kernel can also be expressed in a spherical harmonic basis. Because the kernel does not have a ϕ-dependency, odd order terms for $l > 1$ vanish. In addition, more that 99% of the L^2-energy of the kernel is captured by the terms where $l \leq 2$. Because any high-frequency $(l > 2)$ component of the lighting function $L(\theta, \phi)$ will be severely attenuated, the Lambertian kernel is said to act as a low-pass filter.

For $l \leq 2$, there are nine spherical harmonic basis functions (one of order 0, three of order 1, and five of order 2). We can define the nine *harmonic images* I_i taken under the *virtual* lighting conditions specified by the nine spherical harmonics. The far-reaching consequence is that although lighting conditions are infinite-dimensional (the function space for $L(\theta, \phi)$), the illumination effects on a Lambertian object can be approximated by a 9-D linear subspace $\mathcal{H}$ (the harmonic subspace) spanned by the harmonic images I_i, i.e., $\mathcal{C}$ can be approximated well by $\mathcal{H}$.

While the illumination cone provides a satisfying characterization of $\mathcal{C}$, its exact computation is, in principle, not feasible for most objects. This is because the number of generators is quadratic in the number of distinct surface normals, and for many objects, this number is on the same order as the number of pixels. As an example, for a typical 200 × 200 image, there are roughly 1.6 billion generators. Each generator is stored as a 200 × 200 image, and hence it requires at least 64 terabytes to store all generators. The analysis based on spherical harmonics, and the empirical evidence suggests that an illumination cone can be approximated by a 9-D subspace, and this only requires storing $9n$ numbers. In [28], [29], an alternative approximation is presented.

IV. RECOGNITION FOR FRONTAL POSE, VARIABLE ILLUMINATION

In this section, we discuss six recently published algorithms [2], [12], [17], [28], [42], [52] for face recognition under varying illumination. All of these use an image-based metric, and except for [12], the common feature among them is the ability to produce an approximation to the illumination cone (a low-dimensional linear subspace) that models the illumination effect using only a handful of training images.

These algorithms can roughly be categorized into two types, algorithms that explicitly estimate 3-D face geometry (depth and/or normals) and albedo [2], [17], [42], [52] and algorithms that do not [12], [28]. Using 3-D shape allows the recognition system to consider cast shadows as well as pose variation. The dimensionality of the datum (images) can be reduced, for example, using PCA [42]. On the reduced space, different classifiers (such as support vector machines, Bayesian classifiers and nearest neighbor classifiers) can be brought to bear. The approach of harmonic subspaces [2], [52] provides one way of utilizing surface normals without explicitly computing the 3-D structure. Since the basis images are simply polynomials of the surface normals and albedos, they can be easily computed if the normals and albedos are known. Obviously, a part of these algorithms is the recovery of surface normals and albedos using a few training images. This can be accomplished either using photometric stereo techniques [17] or employing probabilistic methods using some learned prior distributions of normals and albedos [42], [52]. Though not discussed here, other shape reconstruction techniques such as stereo and laser range finders could be used to acquire geometry and registered reflectance.

The algorithms in [12] and [28] do not require 3-D shape information. In [12], the joint probability density function (pdf) of image gradients at a given location is learned empirically from training data, and recognition is then performed by calculating the maximum likelihood using this pdf. In [29] a set of training images (as few as five) are taken under prescribed lighting conditions, and these images form a basis of a linear subspace that effectively approximates the illumination cone, and in turn this leads to good recognition rates.

Georghiades *et al.* **[17]**: In this algorithm, surface depth, normals and albedos of each face are recovered using an uncalibrated photometric stereo technique [5], [51] to first estimate surface normals and albedo which is then integrated to provide 3-D face shape. The input is three or more images of the face from the same pose but under differing lighting (see Fig. 4). As constructing a full illumination cone with $O(n^2)$ extreme rays is unreasonable, the sphere of light source directions is sampled, and a collection of s images is rendered and used as generators $\{e_1, \ldots, e_s\}$ of a cone. Given a probe image x, the face is recognized by finding the closest cone to x. The distance from x to a cone is found as the minimum of $\|x - \sum_{i=1}^{s} a_i e_i\|^2$ subject to the constraint that $a_i \geq 0$. This is a convex programming problem, which can be solved efficiently. A variation is to approximate the sampled cone as a low dimensional linear subspace computed using singular value decomposition (11-D in the reported experiments); finding the nearest subspace is easily performed.

Basri & Jacobs [2]: This face recognition algorithm is a direct application of the illumination model based on spherical harmonics. Similar to the variation in the preceding algorithm, it is also a subspace-based algorithm, but here the appearance model for each individual is a 9-D linear subspace spanned by the nine harmonic images. To compute the harmonic subspace, surface normals and albedo are needed. The effect of cast shadows is not modeled. Let $B = [b_1, \ldots, b_9]$ be the matrix whose columns are harmonic images of an individual. The face recognition decision is based on finding the nearest neighbor using the L^2 reconstruction error; for a query image x, it is given by

$$\min_a \|Ba - x\|^2 \tag{5}$$

where a can be any 9-by-1 vector.

Sim & Kanade [42]: Surface normals and albedos are necessary components in the previous two algorithms. While photometric stereo requires at least three images, [42] presents a maximum likelihood estimation method using just one image along with a statistical prior. In this method, the Lambertian model is augmented with an additional term

$$i(x) = b(x)^t l + e(x, l) \tag{6}$$

The extra term $e(x, l)$ models the effective ambient illumination, and it depends both on x and l. With aligned images, it is assumed that the normals of human faces at

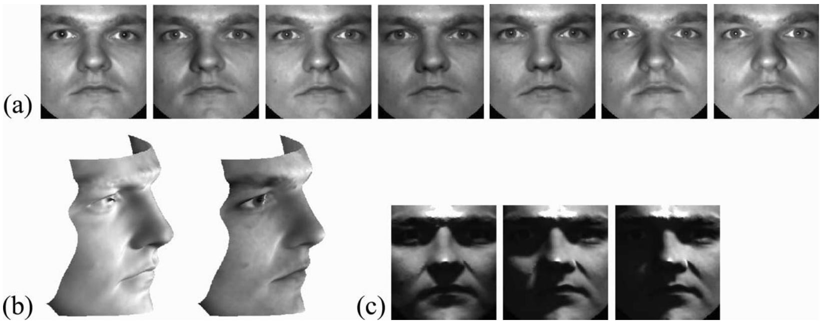


Fig. 4. *3-D reconstruction of a human face. (a) The seven gallery images. (b) Reconstruction results rendered with constant albedo (left) and using the estimated albedos (right). (c) Three synthesized images with new lighting conditions. Note the large variations in shading and shadowing as compared to the seven training images above.*

pixel x form a Gaussian distribution with some mean $\mu_b(x)$ and covariance matrix $C_b(x)$. Similarly, $e(x,l)$ is also assumed to form a Gaussian distribution with mean $\mu_e(x,l)$ and variance $\sigma_e^2(x,l)$. All the parameters forming the prior can be estimated from a collection of training images along with known normals and lighting directions.

With the prior in hand and a given enrollment image, the face shape is estimated by first estimating the unknown illumination l [55], [57]. This allows $\mu_e(x,l)$ and $\sigma_e^2(x,l)$ to be computed. The normal and albedo $b(x)$ are then recovered as a *maximum a posteriori* (MAP) estimate $b_{\mathrm{MAP}}(x) = \arg\max_{b(x)} \Pr(b(x)|i(x))$. From the estimated normals and albedo, a collection of images is rendered under differing lighting, and a subspace is computed. Recognition is again based on finding the nearest subspace.

Zhang & Samaras [52]: Instead of estimating normals and albedos, rendering images, and computing a linear subspace as in [42], Zhang & Samaras directly estimate the nine harmonic images from a single training image. The starting point is an equation similar to (6), namely $i(x) = b(x)^t\alpha + e(x,\alpha)$ where this $b(x)$ is a 9×1 vector that encodes the values of the nine harmonic images at x, and e is the error term exactly as before. In place of l, there is a 9×1 vector α representing the nine coefficients in the truncated spherical harmonics expansion of s. Assuming that $b(x)$ forms a Gaussian distribution at each pixel x, with some mean $\mu_b(x)$ and covariance matrix $\mathcal{C}_b(x)$. As in [42], these parameters can be estimated from a bootstrap collection of images. And like [42], α is estimated first and then a MAP estimate of $b(x)$ is determined. Recognition is performed as in [2].

Chen *et al.* **[12]**: Unlike the previous algorithms, this one does not estimate surface normals and albedos, and it requires only a single training image. It is essentially probabilistic in the sense that the algorithm depends critically on a prior distribution. In this case, the distribution is on the angles between image gradients, and it is obtained empirically from a set of training images of human faces. Here, the joint probability density function ρ for two image gradients can be used as an illumination insensitive measure. If we treat each pixel independently, the joint probability of observing the images gradients ∇I and ∇J of two images I and J of the same object is

$$\begin{aligned} P(\nabla I, \nabla J) &= \prod_{\mathrm{i}\in\mathrm{Pixels}} \rho(\nabla I_i, \nabla J_i) \\ &= \prod_{\mathrm{i}\in\mathrm{Pixels}} \rho(r_1(i), \phi(i), r_2(i)) \end{aligned} \tag{7}$$

where $r_1(i) = |\nabla I(i)|$, $r_2(i) = |\nabla J(i)|$, and ϕ is the angle between the two gradient vectors. With this prior probability $\rho(r_1(i), \phi(i), r_2(i))$ on image gradients and given a query image I, $P(\nabla I, \nabla J)$ is computed for every training image J. The training image having the largest value of $P(\nabla I, \nabla J)$ is considered to be the likeliest to have come from the same face as the query image I. No subspace is involved, and the computation is exceptionally fast and efficient.

Lee *et al.* **[29]**: Implementation-wise, this is perhaps the simplest algorithm. The main insight is to use a specific configuration of point light sources such that real images taken under these lights can directly serve as basis vectors spanning a subspace that accounts for shading variation, and to some extent cast shadows and non-Lambertian effects. Since spherical harmonic functions have negative values, such lighting is not directly physically realizable, and in general it is extremely difficult to precisely realize a source distribution that broadly varies as a function of direction. Instead, it is shown in [28], [29] that there exists a configuration of nine (or fewer) lighting directions such that for any individual, the subspace spanned by images taken under these lighting conditions lies near the individual's harmonic subspace $\mathcal{H}$ and effectively approximates the individual's illumination cone $\mathcal{C}$. Recognition again proceeds by finding the nearest subspace. Fig. 6 shows the nine light source positions and nine images of a face taken under the corresponding lighting.

V. RECOGNITION UNDER VARIABLE POSE AND LIGHTING

The previous sections considered the theory and algorithms for recognizing faces in fixed pose, but under varying illumination. In this section, we consider the situation where both sources of variability are present. As most systems for frontal face recognition are appearance-based, it is natural to try to extend these methods to more general settings. As demonstrated for example by the work on Eigenlight fields [20], [59] and Bayesian Face Subregions [25], appearance-based methods can be extended to handle pose variation. Yet, the recognition rates of such appearance-based techniques when extended to handle both lighting and pose are not yet satisfactory [19]. What has proven to be effective is the use of a 3-D generative model. First, we briefly describe an extention of the methods for variable lighting from the previous section, and then consider 3-D Morphable Models in more detail.

A. Illumination Cone Models for Varying Pose and Lighting

While the set of images of an object in fixed pose but over all illumination conditions is a convex cone (which can be approximated by a linear subspace), the set of images over variable pose and lighting can be characterized as a family of cones (subspaces) swept out as the pose is continuously varied with one cone per pose. The method proposed in [17] for fixed pose and described in Section IV can be generalized immediately to multiple poses by sampling the pose space, and for each pose a subsapce approximation to the cone can be constructed. Hence, the representation for an individual is the union of lowdimenional subspaces (In the experiments reported below, 117 subspaces of dimension 11 were used). For a probe image, recognition is performed by computing the minimum distance to all of these linear subspaces. To speed this computation, principal component analysis can be performed on the $117 * 11 = 1287$ basis images, and the basis vectors of each subspace can be projected to a lower dimensional subspace of the image space.

B. 3-D Morphable Model

In this section, we consider a generative model for predicting an individual's appearance under arbitrary pose and lighting, namely the 3-D Morphable Model [7], [8] which can be used to address one of the most challenging problem of face recognition: estimating the 3-D shape, albedo, pose, and illumination from a *single image viewed from unconstrained pose and illumination conditions*. It is based on strong prior knowledge of the statistical variation of the shape and texture of human faces (the texture is defined as the albedo values for a complete object) and a model simulating the physical interaction of light and a face's surface. The Morphable Model is generic and effectively characterizes a broad population of people. It can be applied to face recognition on images exhibiting simultaneous variations of pose and illumination, using just one photograph per individual in both the gallery and probe sets (see Fig. 7).

One of the reasons for the accuracy of the Morphable Model is that many of the illumination phenomena are modeled, thereby maximizing the image information used in the estimation: non-Lambertian effects such as specular highlights and cast shadows provide cues about the 3-D shape that are used to increase the quality of its estimation. In particular, the 3-D shape is explicitly modeled by a dense set of 3-D points, and so global illumination effects, such as cast shadows, can be used in the image analysis algorithm (see Section V-C). Furthermore, the relative 2-D image location of facial features depends on both the individual's 3-D shape and on her *pose*. The 3-D shape model of the Morphable Model represents these geometric variations explicitly. Hence, 1) the pose can be naturally separated from the shape and 2) facial features can be registered together on a reference frame, thereby producing photorealistic novel views (see Fig. 7) free of artifacts such as double contours and blur visible, for example, on Fig. 5.

The essence of the 3-D Morphable Model is that accurate and general face recognition is possible by the separation of the different sources of variation present in facial images, and their representation using independent sets of parameters: physical effects, such as pose and illumination variations, are modeled using physical principles borrowed from the computer graphics field, whereas variations of identity present in a large population sample are modeled statistically.

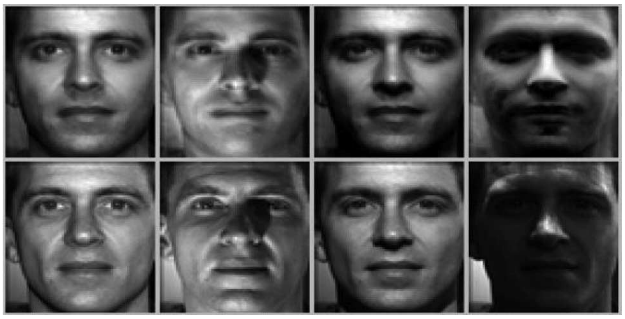


Fig. 5. ***Hallucinated images (courtesy of [42]). Four synthetic images using estimated normals $n(x)$ and $e(x, s)$ (top row) and actual images under the same illumination (bottom row).***

The Morphable Model originated more than a decade ago [7], [8], [36]–[38], [46]–[48]. Initially, ideas as Linear Object Class, separation of shape and texture and view generalization from a single image were developed, which led to the model part of the 3-D Morphable Model. Then, emphasis was put on developing a robust, fast and accurate analysis algorithm. The image analysis algorithm has now matured and is outlined in Section V-C.

There is no illumination invariant (see Section III-A). This means that estimating the 3-D shape of a face from a single photograph is an ill-posed problem: the only clue available in a single image about depth is contained in the shading and the shadows. However, using this information requires the illumination environment, the reflectance, and the texture of the face to be known. As these are generally not known, the problem is ill-posed. One way out of this dilemma is by using prior knowledge. The Morphable Model is a representation of this prior knowledge in that it parameterizes a large population of possible faces, and is independent of extrinsic factors such as pose and illumination. We will see, in the rest of this paper, that this prior constrains the estimation problem sufficiently.

In computer graphics, objects such as human faces can be represented by a mesh that includes a dense set of vertices and a triangle list that specifies the connectivity between the vertices. Associated with each of the N_v vertices is its 3-D position and RGB albedo. These can be arranged in a $3 \times N_v$ shape matrix, $\mathbf{S}$ and a $3 \times N_v$ texture matrix, $\mathbf{T}$

$$\mathbf{S} = \begin{pmatrix} x_1 & x_2 & \cdots & x_{N_v} \\ y_1 & y_2 & \cdots & y_{N_v} \\ z_1 & z_2 & \cdots & z_{N_v} \end{pmatrix}$$

$$\mathbf{T} = \begin{pmatrix} r_1 & r_2 & \cdots & r_{N_v} \\ g_1 & g_2 & \cdots & g_{N_v} \\ b_1 & b_2 & \cdots & b_{N_v} \end{pmatrix}. \qquad (8)$$

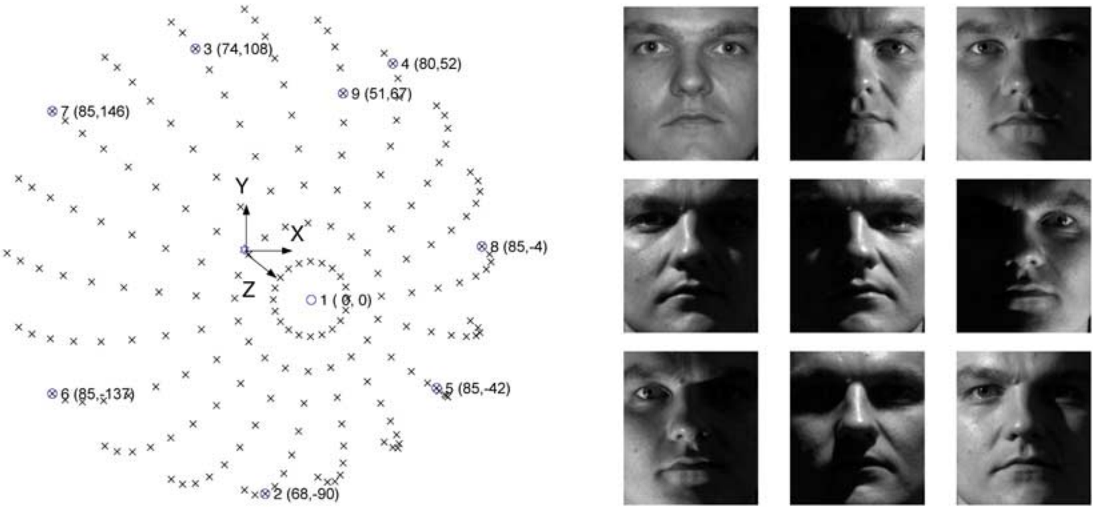


Fig. 6. *Left: The configuration of nine light source directions under which the images on the right are acquired spherical coordinates (ϕ,θ) (on S^2) of the nine directions are {(0, 0), (68, −90), (74, 108), (80, 52), (85, −42), (85, −137), (85, 146), (85, −4), (51, 67). Right: nine images of a person illuminated by these lights.*

An image of this object can then be produced by rotating the object in 3-D, then projecting it onto the image frame, using a perspective projection that represents the way a human observer (with one eye) perceives a 3-D scene

$$\mathbf{W} = \mathbf{R}_\gamma \mathbf{R}_\theta \mathbf{R}_\phi \mathbf{S} + \tau_w \mathbf{1}_{1\times N_v} \qquad (9)$$

$$x_i = f\frac{\mathbf{W}_{1,i}}{\mathbf{W}_{3,i}} \quad y_i = f\frac{\mathbf{W}_{2,i}}{\mathbf{W}_{3,i}}. \qquad (10)$$

In this equation, $\mathbf{R}_\gamma$, $\mathbf{R}_\theta$, and $\mathbf{R}_\phi$ are rotation matrices along the three Euclidean axes, τ_w is a 3×1 vector representing a 3-D translation, f is the focal length of the camera model, and x_i and y_i, $i = 1, \ldots, N_v$, are the 2-D projections of each modeled vertex. These vertices are generally projected at noninteger locations in the image frame. In order to obtain the values at pixel positions, the vertices are interpolated using the triangle list.

This geometrical transformation details where a vertex is to be drawn in the image. The color that must be drawn at that location depends not only on the texture matrix $\mathbf{T}$, but also on the way light is reflected from the surface. As mentioned in Section II, a model of the surface reflection is known as a BRDF. In computer graphics, many elaborate BRDF models have emerged that account to various degrees for complex reflection phenomena. For the Morphable Model, we chose to use the Phong reflectance model that considers the diffuse and specular reflections. More elaborate BRDF models could be used instead, such as the Cook–Torrance [13], Torrance–Sparrow [44], or Lafortune [26] models, which better approximate more complex effects. The Phong model specifies that, when illuminated from a distant light source with unit-length direction $\vec{l}$ and colored intensity (l_r^d, l_g^d, l_b^d), and with an ambient light of intensity (l_r^a, l_g^a, l_b^a), a vertex albedo represented in a column, $\vec{\mathbf{t}}_i$, of the texture matrix, $\mathbf{T}$, is transformed as follows:

$$\mathbf{t}_i^I = \begin{pmatrix} l_r^a & 0 & 0 \\ 0 & l_g^a & 0 \\ 0 & 0 & l_b^a \end{pmatrix} \cdot \mathbf{t}_i + \begin{pmatrix} l_r^d & 0 & 0 \\ 0 & l_g^d & 0 \\ 0 & 0 & l_b^d \end{pmatrix} \cdot \left((\vec{n}_i \cdot \vec{l}) \mathbf{t}_i + k_s (\vec{v}_i \cdot \vec{r}_i)^\nu \mathbf{1}_{3\times1} \right). \qquad (11)$$

In these equations, k_s represents the specular reflectance of human skin (the higher k_s, the more shiny), and ν, the angular distribution of the specular reflections of human skin (the lower ν, the larger the highlight). $\vec{v}_i$ is the unit-length viewing direction between vertex i and the camera center. The unit-length vector $\vec{r}_i$ is the reflection direction of the light, computed from $\vec{n}_i$ and $\vec{l}$.

The first term of (11) is the contribution of the ambient light. The first term of the last parenthesis is the diffuse component of the directed light, and the second term is its specular component. To account for attached shadows, these two scalar products are lower bounded to zero. To account for cast shadows, a shadow map is computed from the global 3-D shape matrix, $\mathbf{S}$, using standard computer graphics techniques [15]. The vertices in shadows are illuminated by the ambient light only. As can be seen in (11), only one directional light source is modeled. In theory, any finite number of light sources can be used.

However, using multiple light sources could increase the difficulty of the analysis algorithm (as the optimization algorithm would have more local minima). The techniques based on modelling the effects of illumination using linear subspaces described in the previous section can handle multiple, distributed sources and recently there has been an effort to wed the shape representation of morphable models with spherical harmonic lighting [54].

Using a mesh model as described above, the image of a particular human face, at any pose when illuminated from a particular direction, can be synthesized given its shape $\mathbf{S}$, and its albedo $\mathbf{T}$. To extend the mesh model so that it can be generic and represent many faces, $\mathbf{S}$ and $\mathbf{T}$ in the 3-D morphable are parameterized as a linear combination of exemplar 3-D faces, so that a particular set of parameters gives the $\mathbf{S}$ and $\mathbf{T}$ for a particular individual. To this end, 200 human faces were acquired using a *Cyberware* laser scanner. If linear combinations were computed on raw scans, nonrealistic faces would be generated: blurry faces with artifacts such as double contours. To eliminate this undesirable effect, a preprocess is applied to the scanned faces: the ensemble of scans is registered and put into correspondence with a reference face. A consistent labeling of all vertices is introduced. A vertex, say, the tip of the nose, is represented by the same vertex index for all faces. This ensures that when a linear combination is performed, the same facial features are added together.

For some values of the shape and texture linear coefficients α_i and β_i, the resulting mesh may not actually be very face-like, and so a probability distribution is learned from examples and placed over the model parameters. We chose to model this as Gaussian distribution. PCA is applied to the registered shape and texture exemplars, yielding N_S shape and N_T texture principal components, $\mathbf{S}^i$ and $\mathbf{T}^i$, respectively, and the standard deviation for each principal component, $\sigma_{S,i}$ and $\sigma_{T,i}$. The

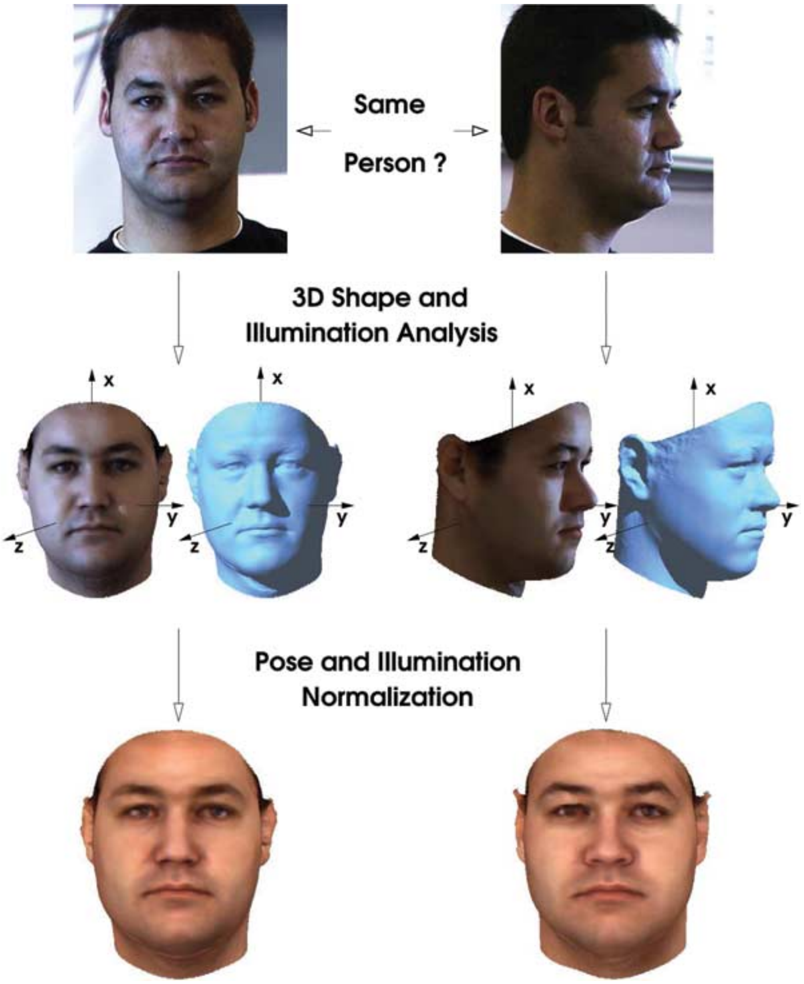


Fig. 7. *To compare two images, the 3-D Morphable Model analyzes each image individually by estimating the 3-D shape, albedo, pose, and illumination parameters. Then, a pose and illumination normalization is performed. Face recognition can be applied by comparing the normalized shape and albedo model parameters.*

face of any individual can then be obtained by a linear combination of principal components

$$\mathbf{S} = \overline{\mathbf{S}} + \sum_{i=1}^{N_S} \alpha_i \cdot \mathbf{S}^i, \quad \mathbf{T} = \overline{\mathbf{T}} + \sum_{i=1}^{N_T} \beta_i \cdot \mathbf{T}^i \tag{12}$$

where $\overline{\mathbf{S}}$ and $\overline{\mathbf{T}}$ are the average of the shape and texture datasets, respectively. The probability of a given shape and texture are then directly obtained from their coefficients, α_i and β_i

$$p(\mathbf{S}) \sim e^{-\frac{1}{2}\sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2}}, \quad p(\mathbf{T}) \sim e^{-\frac{1}{2}\sum_i \frac{\beta_i^2}{\sigma_{T,i}^2}}. \tag{13}$$

Using the Morphable Model framework, the image of the face of any individual seen from any angle and illuminated from any direction can be obtained from the shape parameters α_i, the texture parameters β_i, the shape projection parameters, and the illumination parameters as follows:

$$I^m(x_i(\theta), y_i(\theta)) = \mathbf{t}_i^I(\theta) \tag{14}$$

where x_i and y_i are computed by (10), and $\vec{\mathbf{t}}_i^I$, by (11), and θ denotes the ensemble of model parameters.

In a nutshell, the prior models accounting for the variations of the face image are devised as follows: Gaussian probability models for the registered 3-D shape and albedo, a Phong reflectance model, a single directed light source for the illumination model, and, finally, pose variations are modeled by the physical law of rigid body.

C. Face Image Analysis

The previous section described a generative model able to synthesize a photorealistic image of a human face from model parameters. In vision, the problem is the inverse: how to infer the model parameters explaining a given input face image? We address this problem in an *analysis by synthesis* framework. The task is to find the model parameters such that the face image rendered from these parameters $I^m(x,y)$ is as close as possible to the input image $I(x,y)$. Mathematically, this can be formulated by maximizing the posterior probability of the model parameters given the input image. Using Bayes theorem and assuming that the image pixels are independent and identically distributed with a Gaussian noise σ_I gives rise to the following equation:

$$\min_\theta \sum_i \frac{1}{\sigma_I^2} \left\| I(x_i(\theta), y_i(\theta)) - \mathbf{t}_i^I(\theta) \right\|^2 + \sum_{i=1}^{N_S} \frac{\alpha_i^2}{\sigma_{S,i}^2} + \sum_{i=1}^{N_T} \frac{\beta_i^2}{\sigma_{T,i}^2}. \tag{15}$$

In this cost function, the first sum is the likelihood of the model image, which ensures that the resulting face agrees with the given input image, and the next two sums represent the prior probabilities, that ensure that the face obtained is a likely face [see (13)].

As with most nonlinear optimization algorithms, it must be started using an initial estimate. Typically, the initial shape and texture parameters are their means, and the pose and light direction are taken to be frontal. Additionally, seven landmark points (around the eyes, mouth, and nose) must be manually provided and put into correspondence with the model.

In (15), the texture and shape PCA models are employed to constrain the set of possible solutions. A lighting model is also used, the Phong reflectance model, which has a few parameters when only one light source is handled. However, even these prior models are not strong enough to obtain an accurate estimate of the 3-D shape when only a few manually set anchor points are used as input. This is because the cost function to be minimized is highly nonconvex and exhibits many local minima. In fact, the shape model requires the correspondence between the input image and the reference frame to be found for every visible vertices. Using only facial color information to recover the correspondence is not optimal and may be trapped in regions that present similar intensity variations (eyes/eyebrows, for instance). This is why, we use not only the pixel intensities but also other features of the input image to obtain a more accurate estimate of the correspondence and, as a result, of the 3-D shape. One example of such a feature is the *edges*. Other features that improve the shape and texture estimate are the *specular highlights* and the *texture constraints*. The specular highlight feature uses the specular highlight location, detected on the input image, to refine the normals and, thereby, the 3-D shape of the vertices affected. The texture constraint enforces that the estimated texture lies within a specific range (typically [0, 255]), which improves the illumination estimate. The overall resulting cost function is smoother and easier to minimize, making the system more robust and reliable. A question raised by this problem is how to fuse the different image cues to form the optimal parameter estimate. We chose a Bayesian framework and maximize the posterior probability of the parameters given the image and its features.

This analysis algorithm, called the Multiple Feature Fitting algorithm, is briefly outlined here; a more detailed explanation is provided in [35], [38]. It is demonstrated in [35], [38] that, if the features (pixel intensities, edges and specular highlights) are independent and extracted from the input image by a deterministic algorithm, then the overall cost function is a linear combination of the cost function of each feature taken separately

$$\min_\theta \tau^c C^c + \tau^e C^e + \tau^s C^s + \tau^p C^p + \tau^t C^t \tag{16}$$

where $C^c = \sum_i (1/\sigma_I^2) \| I(x_i(\theta), y_i(\theta)) - \mathbf{t}_i^I(\theta) \|^2$ denotes the pixel intensity feature, $C^P = \sum_{i=1}^{N_S} (\alpha_i^2/\sigma_{S,i}^2) + \sum_{i=1}^{N_T} (\beta_i^2/\sigma_{T,i}^2)$ denotes the prior feature, and C^e, C^s, and C^t denote, respectively, the edge, specular highlights, and texture constraints cost functions. The τ's are weighting parameters. A detailed explanation of these cost functions is provided in [38]. The overall cost function is minimized using a Levenberg–Marquardt optimization algorithm [32]. Graphically, the Multiple Feature Algorithm is described in Fig. 8.

The image edges provide information about the 2-D shape independent of the texture and of the illumination. Hence, the cost function used to fit the edge features provides a more direct constraint on the correspondences and on the shape and pose parameters. This is why it can be seen in Fig. 8 that the plot of the edge cost function across azimuth direction is much smoother than the one of the pixel intensity feature. The edge feature is useful to recover the correspondences of specific facial characteristics (eyes, eyebrows, mouth, nose). On the other hand, it does not carry much depth information. So it is beneficial to use the edge and intensity features in combination.

The specular highlights are easy to detect: the pixels with a specular highlight saturate. Additionally, they give a direct relationship between the 3-D geometry of the surface at these points, the camera direction, and the light direction: a point on a specular highlight has a normal that has the direction of the bisector of the angle formed by the light source direction and the camera direction. Hence, the specular highlight cost function is used to refine the shape estimate for the vertices that are projected onto specular highlights of the input image.

In order to accurately estimate the 3-D shape, it is necessary to recover the texture, the light direction, and its intensity. To separate the contribution of the texture from light in a pixel intensity value, a Gaussian texture prior model is used [see (13)]. However, it appears that this prior model is not restrictive enough and is able to instantiate invalid textures (negative and overflowing color values). To constrain the texture model and to improve the separation of light source strength from albedo, we introduce a feature that constrains the range of valid albedo values.

VI. EXPERIMENTS AND ALGORITHMS COMPARISON

In this section, we discuss the performance of the face recognition algorithms summarized in the previous sections. In the identification task reported here, an image

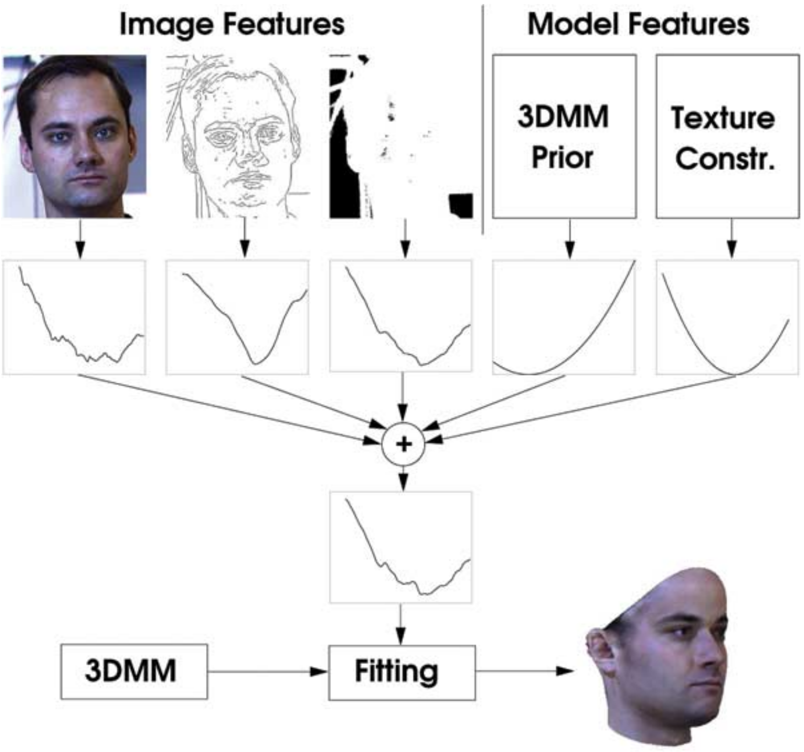


Fig. 8. ***3-D shape, texture, and imaging parameters are estimated using the pixel intensity, the edges, and the specular highlights detected in the input image, shown on the top row. Additionally, two model-based features are used: the 3-D Morphable Model prior and texture constraints. Second row: plots of each feature cost function along the azimuth angle. These cost functions are combined yielding a smoother function that the one based on pixel intensity alone, which is then minimized.***

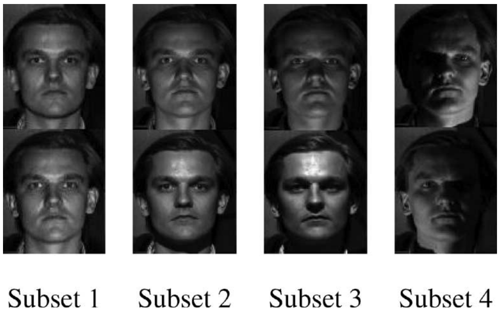


Fig. 9. ***Example images of a single individual in frontal pose from the Yale Face Database B showing the variability due to illumination (Set 1). The images have been divided into four subsets according to the angle the light source direction makes with the camera axis—Subset 1 (up to 12°), Subset 2 (up to 25°), Subset 3 (up to 50°), Subset 4 (up to 77°).***

of an unknown person is provided to the system under evaluation. The unknown face image is then compared to a database of known people, called the gallery set, which may include one or several images of each individual. The ensemble of unknown images is called the probe set. It is assumed that the individual in the unknown image is present in the gallery (i.e., this is not a verification experiment). The algorithms mentioned in Section VI-A need to adjust their internal parameters for each individual in the gallery set. Hence, the gallery set is called training set, as the gallery images are used to train the system.

The algorithms are tested below using subsets of two face databases that have become the **de facto** standards in the past few years for studies of variable lighting and pose. **Set 1**: The Yale Face Database B [17] contains ten individuals acquired under 64 different illumination conditions and nine poses (a sample of the Yale database is shown in Fig. 9). For the Experiments reported in Section VI-A, only frontal pose and 45 illumination conditions were used, and this is called Set 1. The images are grouped into four subsets according to the angle of the lighting with respect to the camera axis. The first two subsets cover the angular range 0° to 25°, the third subset covers 25° to 50°, and the fourth subset covers 50° to 77°. The heavily shadowed images in subset four are the most challenging for face recognition.

Datasets 2 and 3 include two different portions of the Pose, Illumination and Expression (PIE) database from CMU [41]. The full PIE database contains images of 68 individuals, 43 illumination conditions (21 source direction with or without additional ambient light), 13 poses, and with four different expressions. **Set 2**: This set includes a portion of the CMU-PIE database, which, similar to Set 1, is restricted to the frontal pose photographs taken in a neutral expression. Apart from being extracted from two different databases, the main differences between Sets 1 and 2 is that Set 2 has more images of more individuals than in Set 1, while Set 1 has more images under more difficult lighting conditions. **Set 3**: The purpose of this set is to experiment with the performance of algorithm when confronted with both pose and illumination variation. It includes another portion of the CMU-PIE database. It contains photographs of 68 individuals, illuminated from 22 directions plus ambient illumination and viewed from three poses (frontal, side, and profile). Each individual is photographed 3 × 22 times. Example images of this dataset are shown in Fig. 10.

A. Frontal Pose, Varying Illumination

Here we report identification performance of the algorithms outlined in previous sections on the frontal pose, varying illumination problem. While these recognition algorithms are quite robust against illumination variation, they differ from each other in two fundamental ways: by the number of training images and by the way subspaces are computed from the training images.

Table 1 summarizes the experimental results using Set 1 (Yale face database B). Methods are trained from images in Subset 1, and tested on all subsets. The first four rows

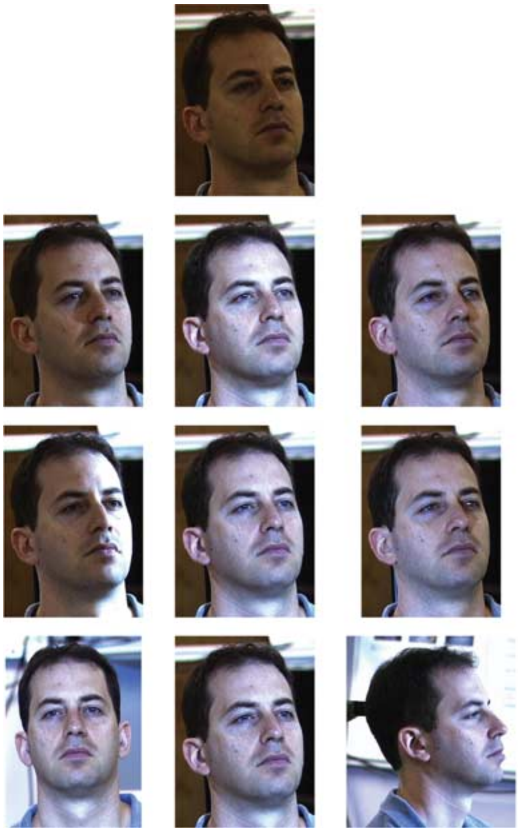
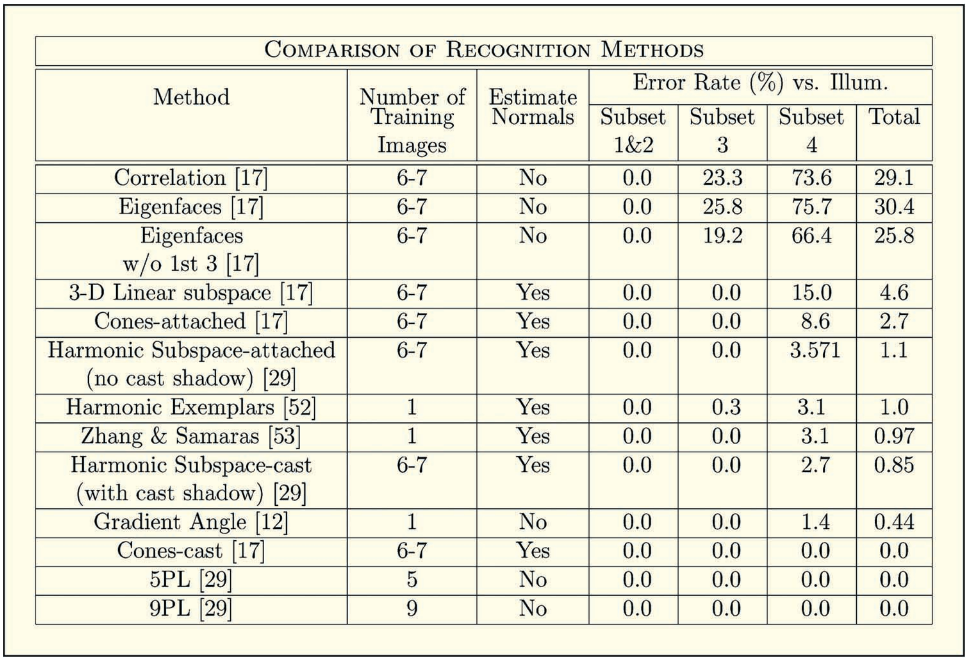


Fig. 10. ***Example of CMU-PIE dataset (Set 3) photographs. Top rows: ambient light only. Two middle rows: illuminated by ambient light and some of the 21 light directions. Bottom row: photographs of the three different poses.***

Table 1 The Error Rates for Various Recognition Methods on Subsets of the Yale Face Database B. Each Entry is Taken Directly From a Published Source Indicated by Citation



Comparison of Recognition Methods

Method	Number of Training Images	Estimate Normals	Error Rate (%) vs. Illum.			
			Subset 1&2	Subset 3	Subset 4	Total
Correlation [17]	6-7	No	0.0	23.3	73.6	29.1
Eigenfaces [17]	6-7	No	0.0	25.8	75.7	30.4
Eigenfaces w/o 1st 3 [17]	6-7	No	0.0	19.2	66.4	25.8
3-D Linear subspace [17]	6-7	Yes	0.0	0.0	15.0	4.6
Cones-attached [17]	6-7	Yes	0.0	0.0	8.6	2.7
Harmonic Subspace-attached (no cast shadow) [29]	6-7	Yes	0.0	0.0	3.571	1.1
Harmonic Exemplars [52]	1	Yes	0.0	0.3	3.1	1.0
Zhang & Samaras [53]	1	Yes	0.0	0.0	3.1	0.97
Harmonic Subspace-cast (with cast shadow) [29]	6-7	Yes	0.0	0.0	2.7	0.85
Gradient Angle [12]	1	No	0.0	0.0	1.4	0.44
Cones-cast [17]	6-7	Yes	0.0	0.0	0.0	0.0
5PL [29]	5	No	0.0	0.0	0.0	0.0
9PL [29]	9	No	0.0	0.0	0.0	0.0

contain the result of using well-established and "baseline" algorithms that do not provide significant illumination modeling. The next eight rows display the results of using more sophisticated illumination modeling. The difference in performance between these two categories of algorithms is apparent: while the total error rates for the former category hover above 20%, algorithms in the later category can achieve less than 1% error rates. Note that different algorithms require different numbers of training images, and in evaluating algorithm performance, we have tried to use the same number of training images whenever possible.

Before going further, we briefly describe the five baseline algorithms [17]. Correlation is a nearest-neighbor classifier in the image space [10] in which all of the images are normalized to have zero mean and unit variance. Eigenfaces uses PCA to obtain a subspace from the training images [45]. One proposed method for handling illumination variation using PCA is to discard the first three most significant principal components, which, in practice, yields better recognition results [3]. The 3-D linear subspace method uses the 3-D illumination linear subspace $\mathcal{L}$ in (3) as a representation. While this method models the variation in shading when the surface is completely illuminated, it does not model shadowing. Note also that two variations of the illumination cones method were employed. In the "Cone-attached" and "Cone-cast" methods, images without and with cast shadows were used to compute the illumination cones, respectively.

There are several ways to understand the results in Table 1. First, recognition is generally easier in images taken under frontal illumination. As expected, the images with more shadowing (those from Subsets 3 and 4) are the main challenges. As the first four "baseline" algorithms clearly demonstrate, it is difficult to robustly perform recognition for these images without any significant illumination modeling. Second, linear subspace models are indeed an effective tool for modeling illumination. This is validated by the following experiment: Instead of computing the minimum distance to the 9-D subspace spanned by the 9 gallery images per subject as in the 9PL method, recognition can be performed based on the minimum distance to these nine images. In this later case, the error rate becomes 22.6% versus the perfect rate reported in Table 1. While the same gallery images are being used, the effectiveness of the 9PL method is due to the ability of the subspace to correctly extrapolate images under novel illumination conditions.

While the on-line recognition processes for the more accurate algorithms in Table 1 are similar, they differ significantly, however, in their off-line training processes. For algorithms that required surface normals and/or 3-D shape, at least three training images are needed for each subject to be identified. In this experiment, typically six frontally illuminated images were used to estimate the shape and albedos using photometric stereo techniques. Although "Harmonic Exemplar" can use just one training image, it requires the priors on harmonic images that can only be obtained using an off-line training process that typically requires a large number of training images. "Gradient Angle" is similar in that priors on the angles between image gradients have to be estimated empirically. Perhaps, the simplest algorithm conceptually and in implementation is "9PL." Since there is practically no training involved here, one simply needs to obtain images of a person taken under nine specified lighting conditions. Further experiments have also shown that a five-dimensional subspace ("5PL") may be sufficient for robust face recognition.

Experiments have also been reported in the literature using two variants of **Set 2** (CMU-PIE database, frontal pose, 22 or 23 illumination conditions). In [29], it has been demonstrated that using only a 7-D subspace for each individual (i.e., seven training images per person), an overall recognition error rate of 2.8% can be achieved for a variant of Set 2 where only the direct source is present (No ambient lighting). On the same dataset, Sim and Kanade [42] report an error rate of 5%, though only a single image is used for training. On a variant of Set 2 containing images with both ambient lighting and a directional light source, the error rate is 0.1% using the 3-D Morphable Model and with a single image per individual in the gallery set (more details on this experiment are provided in the next section). In general, the results are not directly comparable since as shown in [29], recognition becomes easier when the lighting contains a larger diffuse component.

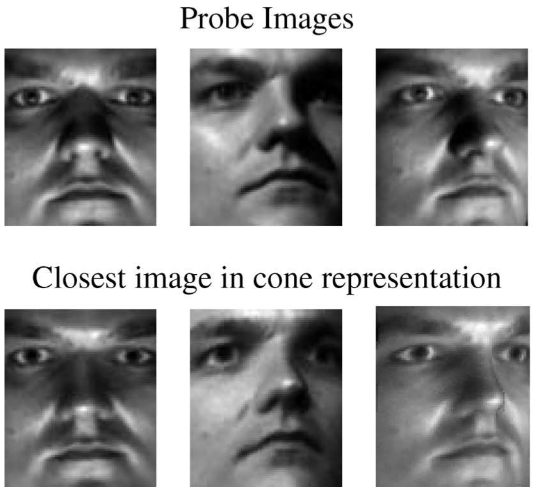
B. Variable Pose and Illumination

We now consider the performance of the two techniques for using a generative model to handle both pose and lighting variation.

Experimental results for the technique based on illumination cones is evaluated on the Yale Face Database B. For ten subjects, the representations are constructed from seven gallery images in frontal pose with near frontal lighting (images from Subset 1). Fig. 11 shows the ability of the method to extrapolate from frontal to nonfrontal poses with complex lighting. On the left, probe images are shown along with the closest (synthetic) image in the representation. In the table on the right, the error rate over 4050 images is shown as a function of increasing pose angle. As a baseline, the illumination cone method is compared to the simplest appearance-based method, nearest neighbor recognition to the gallery images. The illumination cone representation was constructed using photometric stereo in two ways, assuming Lambertian reflectance and the Torrence–Sparrow reflectance model. For more details, see [16], [17].

We now investigate the performance of the 3-D Morphable Model and its Multiple Feature Fitting algorithm in an identification experiment. For these tests, we chose to use **Set 3**, (the subset of the CMU-PIE face image database [40] that presents variation in both pose and illumination for 68 individuals). Observing the best practices for face recognition systems evaluation mentioned in the introduction, the individuals in the PIE database are not included in the training set used to construct the Morphable Model. In this experiment, the gallery set includes a single image per individual photographed in an unknown pose and under unknown illumination conditions.

For each probe image, the 3-D Morphable Model is first fit to the image. Then, identification is performed by comparing the fitting result of the probe image to those of gallery images in two ways: either the identity-specific model parameters (the shape α_i and texture β_i parameters) are compared, or the pose and illumination normalized

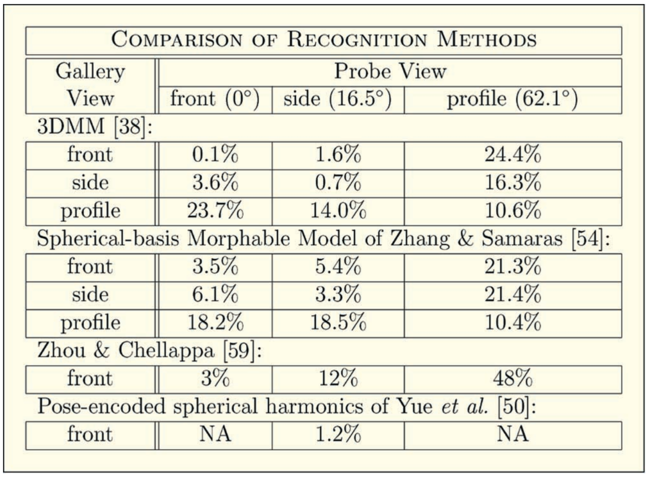


Extrapolation in Pose

Method	Error Rate (%) vs. Pose			Total
	Frontal (Pose 1)	12° (Poses 2 3 4 5 6)	24° (Poses 7 8 9)	Error Rate (%)
Correlation[14]	29.1	75.8	87.2	74.4
Lambertian Cones[13]	0.9	2.5	4.4	2.96
Torrance Sparrow[13] Cones	0.7	1.8	4.2	2.47

Fig. 11. ***Extrapolation in pose: left: upper row shows three images of a face from the probe set, while the lower row shows the closest reconstructed images from illumination cone representations. Note that these images are not explicitly stored or directly synthesized by the generative model, but instead lie within the closest matching 11-D linear subspace. Right: error rates (%) as the viewing direction increases. The three methods have been trained on seven images per person from Subset 1 (near-frontal illumination), Pose 1 (frontal pose). Note that each reported error rate is for all illumination subsets (1 through 4). The "frontal pose" includes 450 test images, the "12 degree" (poses 2, 3, 4, 5, 6) includes test 2250 images, and the "24 degree" (poses 7, 8, 9) includes 1350 test images.***

Table 2 Mean Identification Error Percentage for Different Methods Obtained for the PIE Data Set, Averaged Over All Lighting Conditions for Front, Side, and Profile View Galleries. All the Experiments Included a Single Image per Individual in the Gallery Set With a Front Illumination Condition



COMPARISON OF RECOGNITION METHODS			
Gallery View	Probe View		
	front (0°)	side (16.5°)	profile (62.1°)
3DMM [38]:			
front	0.1%	1.6%	24.4%
side	3.6%	0.7%	16.3%
profile	23.7%	14.0%	10.6%
Spherical-basis Morphable Model of Zhang & Samaras [54]:			
front	3.5%	5.4%	21.3%
side	6.1%	3.3%	21.4%
profile	18.2%	18.5%	10.4%
Zhou & Chellappa [59]:			
front	3%	12%	48%
Pose-encoded spherical harmonics of Yue *et al.* [50]:			
front	NA	1.2%	NA

images (see Fig. 7) can be compared. First, an experiment using the model parameter is described, and then an experiment carried out in the Face Recognition Vendor Test 2002 [31] on normalized images is reported.

1) Identification From Model Parameters: In the 3DMM framework, facial images are compared based on the combined shape and texture model parameters (denoted by $\mathbf{c_g}$ and $\mathbf{c_p}$ for gallery and probe, respectively), using the following metric (denoted by D):

$$\mathbf{c} = [\alpha_1, \ldots, \alpha_{N_S}, \beta_1, \ldots, \beta_{N_T}]$$

$$D = \frac{\mathbf{c_g^T c_p}}{\sqrt{\left(\mathbf{c_g^T c_g}\right) \cdot \left(\mathbf{c_p^T c_p}\right)}}. \tag{17}$$

Three set of experiments were performed with the 3DMM in which the gallery images contained different poses. In each case, the probe set included all the other images of the CMU-PIE dataset. The probe set was divided into three sets also according to the pose. The results reported on the first three rows of Table 2 are identification error percentage averaged over each probe set obtained with the 3DMM. For each cell, the average is taken over all 22 illumination conditions of the probe set (as the gallery set includes only images with a front illumination).

As a comparison, results obtained on the same (or similar) dataset with three recent methods are presented.

Zhang *et al.* **[54]**: combines Spherical Harmonics to model illumination and a 3DMM to model albedo and shape. A drawback of this method is that the normals used for illumination are independent from the normals of the 3-D shape. They are statistically modelled in conjunction with albedo: modifying the albedo cannot be done without modifying the normals and the reflectance, which contradicts the desirable behavior of coding different physical phenomena using distinct parameters. In fact the 3-D shape model is only used for geometrical alignment. A consequence of this is that the 3-D shape is not estimated from the shading as in the case of the 3DMM, but from a set of 60 *manually* marked image feature points located on the silhouette and on the inner part of a face (eyes, eyebrows, mouth, and nose). Thus, this method requires much more human operator input than the 3DMM fitting algorithm to analyse an image (seven feature points are required for the 3DMM). Additionally, it was demonstrated in [35], using a synthetic example, that correspondence information alone cannot be used to estimate accurate 3-D shape. In fact, we think that, in single image analysis, the main reason for using an illumination model is to enable 3-D shape estimation. These impediments may explain the lower identification performance compared to the 3DMM (Table 2). An advantage of the Zhang and Samaras method over the 3DMM is that multiple light sources are explicitly modeled and accounted for in the analysis algorithm.

Zhou & Chellappa [59]: Results of the pose and illumination invariant face recognition of Zhou & Chellappa [59] are provided for a frontal gallery pose. This algorithm is an image-based multiview appearance method that uses an approximation of the Lambertian model for illumination (attached and cast shadows are ignored and the pixels estimated in the shadow treated as outlier).

Yue *et al.* **[50]**: The Pose-encoded spherical harmonics method is a direct extension of the Zhang & Samaras [52] method to the pose invariant case, thereby achieving combined pose and illumination invariance. Similar to [52] variations due to multiple light sources are modeled using a spherical harmonic based model. View invariance is obtained by generating a frontal view from a face image at any view, by applying a 2-D linear warping. This warping is defined by a set of 63 *manually* marked feature points.

Results from this method were only reported on the frontal gallery, side view pose.

We mentioned in the Introduction the difficulty of empirical comparison due to the different testing procedures: indeed, the comparison between the first two systems and the last two is misleading. Although they are evaluated on the CMU-PIE image set, Zhou & Chellappa and Yue *et al.* used the set without ambient illumination whereas images with ambient illumination were used for the 3DMM and the Zhang & Samaras. Zhou & Chellappa trained their system on half of the images of the CMU-PIE database, the other half being used to form the gallery and testing sets. Hence, the gallery sets included only 34 individuals, whereas all 68 individuals of the dataset were used to evaluate the other systems (as they were not trained on the same CMU-PIE dataset, but on a completely different dataset).

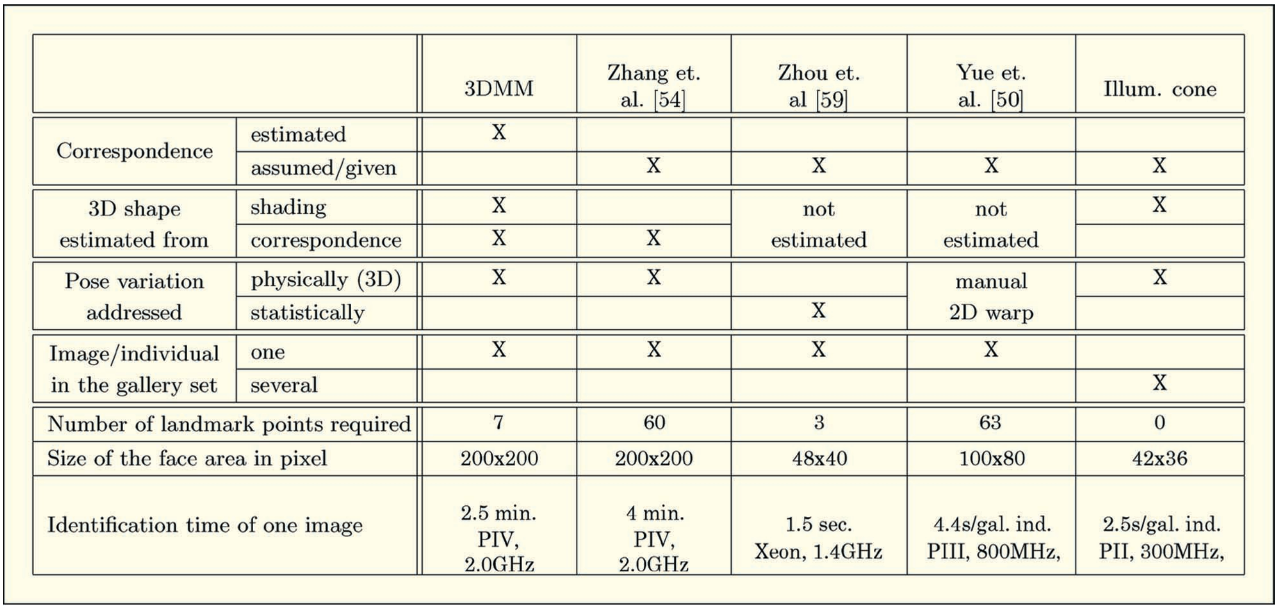
Table 2 indicates that, although the problem addressed by the 3DMM is more difficult (it uses far less human operator input than the Zhang & Samaras and Yue *et al.* systems and its gallery is twice as large as the one of Zhou & Chellappa), the Morphable Model provides significantly better generalization performance across illumination and pose. We conjecture that the reason for its improved performance is that the 3DMM makes fewer assumption on the image formation process and uses more image information than the other systems: Compared with [54], the 3DMM estimates the shape using the shading of the pixels in the face area and not from the 2-D coordinates of a sparse set of manually marked feature points. Compared with [59], the 3DMM treats explicitly a face image as a projection of a 3-D object, whereas [59] attempts to learn pose variation using statistical techniques. Moreover, when this technique compares two images, the corresponding pixels are not adjusted during image analysis, making the comparison less precise (the images are aligned by registering three image points). The recent system of Yue *et al.* [50] is promising, it is regretful, though, that it requires extensive manual interaction. It is an example of the fact that addressing the full face recognition problem is much more involved than addressing part of it (correspondences are not estimated). Similar to the Illumination Cone, the run time of this method scales linearly with the number of individuals in the gallery set, which may set high constraints on a system with a large gallery set.

It should be noted that the results on profile view are significantly less accurate than on front and side views. The major drawback of the 3DMM analysis algorithm is that, to set the initialization parameters, it requires seven landmark points. Additionally, a Matlab implementation of the analysis algorithm takes 2.5 min on a modern computer.

Comparison of the results of the 3DMM and the illumination cone methods of Fig. 11 is difficult as these systems were not tested on the same images. What can be noted is that the error rate for the 3DMM is slightly better than the illumination cone, despite the fact that the gallery set of the CMU-PIE database used to evaluate the 3DMM includes almost seven times more persons that the Yale dataset used with the illumination cone, and this method requires seven gallery images for each subject as opposed to the 3DMM that necessitates only one.

The difference of the aforementioned pose and illumination invariant face recognition algorithms are summarized in Table 3. In this table the methods are

Table 3 Characteristics of the Face Recognition Algorithms Described Earlier



		3DMM	Zhang et. al. [54]	Zhou et. al [59]	Yue et. al. [50]	Illum. cone
Correspondence	estimated	X				
	assumed/given		X	X	X	X
3D shape estimated from	shading	X		not estimated	not estimated	X
	correspondence	X	X			
Pose variation addressed	physically (3D)	X	X		manual 2D warp	X
	statistically			X		
Image/individual in the gallery set	one	X	X	X	X	
	several					X
Number of landmark points required		7	60	3	63	0
Size of the face area in pixel		200x200	200x200	48x40	100x80	42x36
Identification time of one image		2.5 min. PIV, 2.0GHz	4 min. PIV, 2.0GHz	1.5 sec. Xeon, 1.4GHz	4.4s/gal. ind. PIII, 800MHz,	2.5s/gal. ind. PII, 300MHz,

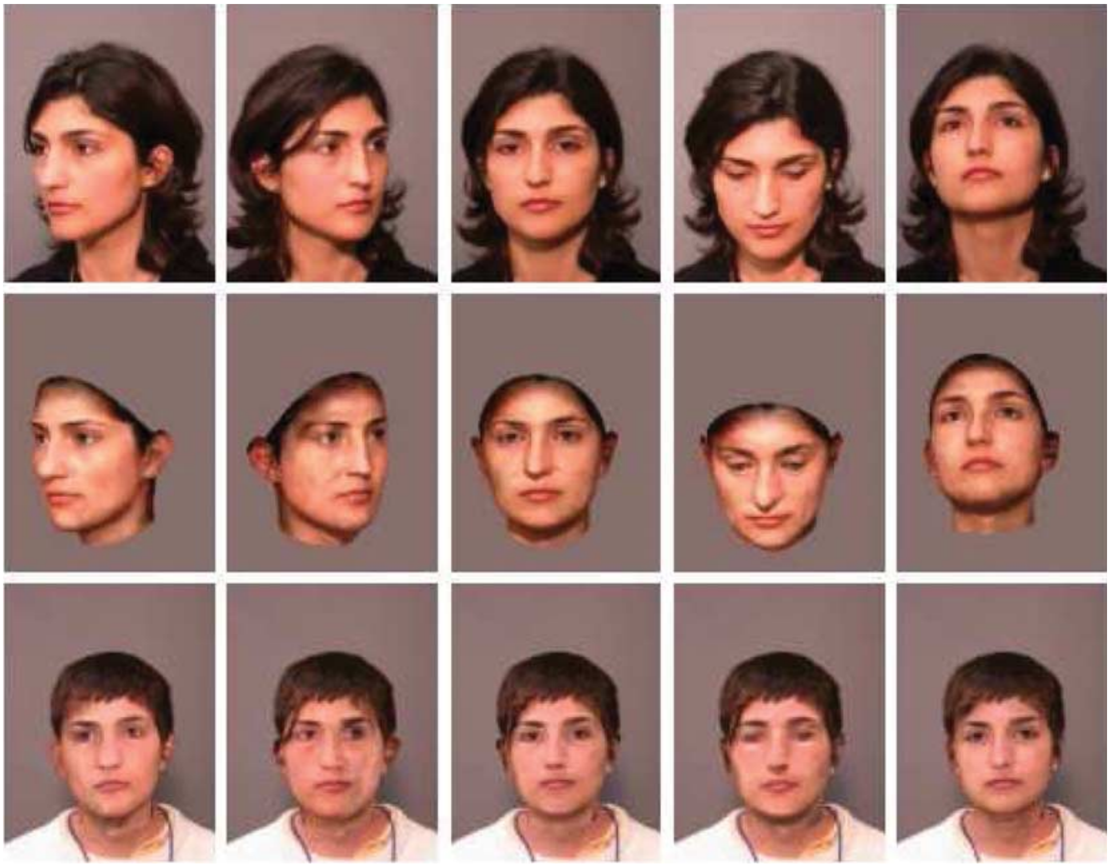


Fig. 12. ***3DMM is fit to the original images (top row). Renderings of the fitting result are shown in the middle row. Mapping the texture of visible face regions on the surface and rendering it with a standard background produce virtual front views (bottom row). This figure was first published in the face Recognition Vendor Test 2002 [31].***

classified according to some major image analysis algorithm features: 1) The corresponding pixel in a gallery and probe image are either estimated from the pixel intensities, or are given by a sparse set of landmark points. We conjecture that if a large set of corresponding pixels is estimated, the image comparison is more precise. 2) Given one or several images of the same object, the 3-D shape can be either estimated from the shading, from the corresponding pixels (using prior knowledge), or from both. 3) Image variations induced by pose variations can be modeled using an explicit 3-D model, taking advantage of the intrinsic 3-D nature of heads, or can be learned statistically. 4) Several images per individual may be required, which limits the applicability of the system. 5) Finally, the last rows provide information on the operating conditions of the algorithms. The time required to identify one input image, listed on the last row, is only indicative as the experiments were performed on different machines (as indicated) and moreover, the methods were all implemented in Matlab and the code was not optimized. The timings listed do not include the manual labeling time. For the last two algorithms, the identification time is proportional to the number of individual in the gallery set, thus, for these methods, we chose to indicate the identification time per individual in the gallery set.

2) Identification From Normalized Images: The Face Recognition Vendor Test (FRVT) 2002 [31] was an independently administered assessment, conducted by the U.S. Government, of the performance of commercially available automatic face recognition systems. It was realized that identification of face images significantly drops if the face image is nonfrontal. Hence, one of the questions addressed by FRVT02 is: Does identification performance of nonfrontal face images improve if the pose is normalized by our 3-D Morphable Model? To answer this question, we normalized the pose of a series of images [6]. Normalizing the pose means to fit an input image where the face is nonfrontal, thereby estimating its 3-D structure, and to render a frontal view of the estimated face on top of a *constant* frontal view photograph of another person. Examples of pose normalized images are shown in Fig. 12. As

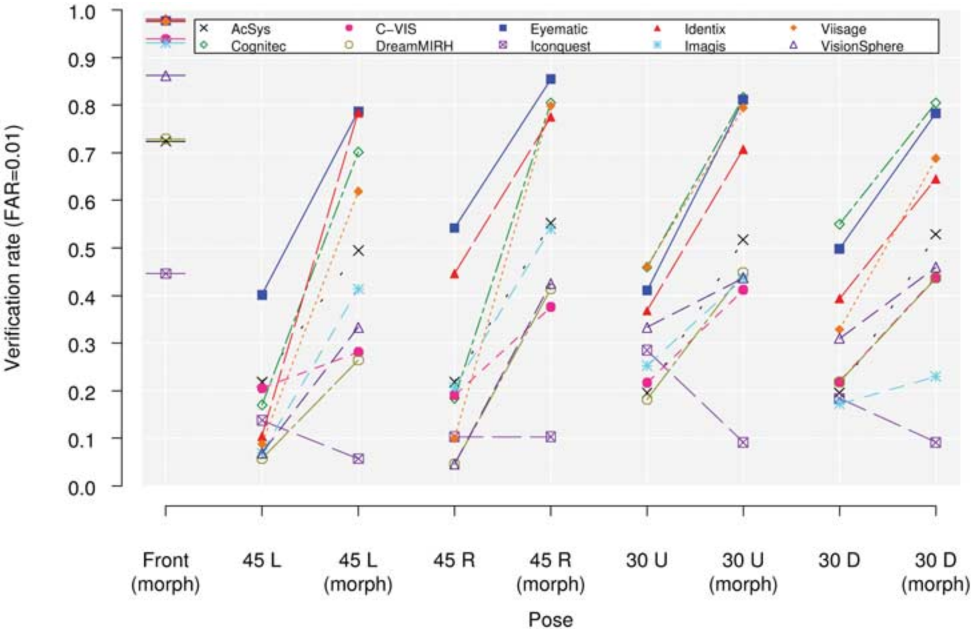


Fig. 13. ***The effect on the verification performance of the original images versus normalized images using the 3-D Morphable Model. The verification rate at a false alarm rate of 1% is plotted. This figure was first published in the face Recognition Vendor Test 2002 [31] and in [6].***

neither the hair nor the shoulders are modeled, the synthetic images are rendered into a standard frontal face image of one person.

The normalization was applied to images of 87 individuals at five poses (frontal, two side views, one up, and a down view). Identification was performed by the ten participant systems evaluated in FRVT02 (see [31], pp. 31–32) using the frontal view images as gallery and nine probe sets: four probe sets with images at nonfrontal views, four probe sets with the normalized images of the nonfrontal views and one probe set with 3DMM preprocessing normalization applied to frontal images. The comparison of performance between the normalized images (a.k.a. morph images) and the raw images is presented on Fig. 13 for a verification experiment (the verification-rate is plotted for a false alarm rate of 1%).

The frontal morph probe set provides a baseline for how normalization affects an identification system. In the frontal morph probe set, normalization is applied to the gallery images. The results on this probe set is shown on the first column of Fig. 13. The verification rates would be 1.0, if a system were insensitive to the artifacts introduced by the Morphable Model and did not rely on the person's hairstyle, collar, or other details which are exchanged by the normalization (which are, anyway, not reliable features for identification). The sensitivity to the Morphable Model of the ten participants ranges from 0.98 down to 0.45. The overall results show that with the exception of Iconquest, Morphable Models significantly improved (and usually doubled) performance.

VII. CONCLUSION

Re-examining the images in Fig. 1 in the Introduction, we now have, at our disposal, a number of face recognition algorithms that can comfortably handle these formidable-looking images. Barely a decade ago, these images would have been problematic for face recognition algorithms of the time. The new concepts and insights introduced in studying illumination modeling in the past decade have bore many fruits in the form of face recognition algorithms that are robust against illumination variation. In many ways, we are very fortunate because human faces do not have more complex geometry and reflectance. Coupled with the superposition nature of illumination, this allows us to utilize low-dimensional *linear* appearance models to capture a large portion of image variation due to illumination. Linearity makes the algorithms efficient and easy to implement, and the appearance models make the algorithms robust. Yet the generalization to pose variation presented for the illumination cone method does not scale well as the number of subjects increases since the probe image must be compared to the representation of each enrolled subject. One approach would be to use the generative models to create a discriminative classifier. Additionally, there remains a challenge of how these techniques can be extended to handle nonrigid shape variation such as facial expression.

At the cost of using nonlinear optimization techniques, the 3-D Morphable Models has been shown to handle combined pose and illumination variations, and state of the art identification performance is obtained. Morphable Models can be extended, in a relatively straightforward manner, to cope with other sources of variation such as expression. Yet, the current implementation requires manual selection of seven feature points on a face image, and this is equivalent to providing a good estimate of 3-D pose; clearly, there is a need either to detect automatically such features or to directly estimate head pose over the full range of lighting conditions.

In all of the presented work, the local reflectance models (Lambertian or Phong) are overly simplistic for skin and facial hair, and interreflections and subsurface scaterring are completely ignored. An open question is whether incorporation of these more sophisticated image formation models would impact recognition performance given other confounding factors. In order to model fine and identity-related details such as freckles, birthmarks, and wrinkles, it might be helpful to extend the Morphable Model framework for representing texture. Indeed, a linear combination of textures is a rather simplifying choice, hence improving the texture model is subject to future research.

A component of a full recognition system is robust face detection and alignment over a wide range of illumination and pose variations. Present face detection techniques [49] are not as effective over the same range of conditions as the presented recognition techniques. Because face tracking is an integral and indispensable part of video face recognition, it is also a challenging problem to develop a tracker that is robust against pose and illumination variations. Expression, partial occlusion, makeup, aging, and other factors must also be considered in concert with work on illumination and pose. ■

REFERENCES

[1] G. Aggarwal and R. Chellappa, "Face recognition in the presence of multiple illumination sources," in *Proc. Int. Conf. Computer Vision*, 2005, pp. 1169–1176.

[2] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 383–390, Feb. 2003.

[3] P. Belhumeur, J. Hespanha, and D. J. Kriegman, "Eigenfaces versus fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[4] P. Belhumeur and D. J. Kriegman, "What is the set of images of an object under all possible lighting conditions," *Int. J. Comput. Vis.*, vol. 28, pp. 245–260, 1998.

[5] P. Belhumeur, D. J. Kriegman, and A. Yuille, "The bas-relief ambiguity," *Int. J. Comput. Vis.*, vol. 35, no. 1, pp. 33–44, 1999.

[6] V. Blanz, P. Grother, J. Phillips, and T. Vetter, "Face recognition based on frontal views generated from non-frontal images," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 454–461.

[7] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH 99)*, 1999, pp. 187–194.

[8] ——, "Face recognition based on fitting a 3-D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003.

[9] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Three-dimensional face recognition," *Int. J. Comput. Vis.*, vol. 64, no. 1, pp. 5–30, 2005.

[10] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 10, pp. 1042–1053, Oct. 1993.

[11] K. I. Chang, K. W. Bowyer, and P. J. Flynn, "An evaluation of multimodal 2D + 3D face biometrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 4, pp. 619–624, Apr. 2005.

[12] H. Chen, P. Belhumeur, and D. Jacobs, "In search of illumination invariants," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2000, pp. 1–8.

[13] R. L. Cook and K. E. Torrance, "A reflectance model for computer graphics," *ACM Trans. Graph.*, vol. 1, no. 1, pp. 7–24, 1982.

[14] R. Epstein, P. Hallinan, and A. Yuille, "$5 + / - 2$ eigenimages suffice: An empirical investigation of low-dimensional lighting models," in *Proc. Workshop Physics-Based Modeling in Computer Vision*, 1995, p. 108.

[15] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics: Principles and Practice*. Reading, MA: Addison-Wesley, 1996.

[16] A. Georghiades, "Incorporating the Torrance and Sparrow model of reflectance in uncalibrated photometric stereo," in *Proc. Int. Conf. Computer Vision*, 2003, pp. 816–825.

[17] A. Georghiades, D. J. Kriegman, and P. Belhumeur, "From few to many: Generative models for recognition under variable pose and illumination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[18] G. Golub and C. van Loan, *Matrix Computation*. Baltimore, MD: John Hopkins Univ. Press, 1989.

[19] R. Gross, S. Baker, I. Matthews, and T. Kanade, *Handbook of Face Recognition, Chapter Face Recognition Across Pose and Illumination*. New York: Springer-Verlag, 2004.

[20] R. Gross, I. Matthews, and S. Baker, "Eigen light-fields and face recognition across pose," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, 2002, pp. 1–7.

[21] P. Hallinan, "A low-dimensional representation of human faces for arbitrary lighting conditions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1994, pp. 995–999.

[22] J. Ho and D. J. Kriegman, "On the effect of illumination and face recognition," in *Face Processing: Advanced Modeling and Methods*, R. Chellappa and W. Zhao, Eds. Burlington, MA: Elsevier, 2005.

[23] T. Igarashi, K. Nishino, and S. K. Nayar, "The appearance of human skin," Columbia Univ. Comput. Sci., Tech. Rep., 2005.

[24] H. Jensen, S. Marschner, M. Levoy, and P. Hanrahan, "A practical model for subsurface light transport," in *Proc. SIGGRAPH*, 2001, pp. 511–518.

[25] T. Kanade and A. Yamada, "Multi-subregion based probabilistic approach toward pose-invariant face recognition," in *Proc. IEEE Int. Symp. Computational Intelligence in Robotics and Automation*, 2003, pp. 954–959.

[26] E. P. F. Lafortune, S.-C. Foo, K. E. Torrance, and D. P. Greenberg, "Non-linear approximation of reflectance functions," in *Proc. 24th Annu. Conf. Computer Graphics and Interactive Techniques*, 1997, pp. 117–126.

[27] J. Lee, B. Moghaddam, H. Pfister, and R. Machiraju, "A bilinear illumination model for robust face recognition," in *Proc. Int. Conf. Computer Vision*, 2005, pp. 1177–1184.

[28] K. Lee, J. Ho, and D. J. Kriegman, "Nine points of lights: Acquiring subspaces for face recognition under variable lighting," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, pp. 519–526.

[29] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[30] S. R. Marschner, H. W. Jensen, M. Cammarano, S. Worley, and P. Hanrahan, "Light scattering from human hair fibers," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 780–791, 2003.

[31] P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, "FRVT 2002: Evaluation report," NIST, Tech. Rep. NISTIR 6965, 2003.

[32] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge, MA: Cambridge Univ. Press, 1992.

[33] R. Ramamoorthi and P. Hanrahan, "An efficient representation for irradiance environment," in *Proc. SIGGRAPH*, 2001, pp. 497–500.

[34] ——, "A signal-processing framework for inverse rendering," in *Proc. SIGGRAPH*, 2001, pp. 117–228.

[35] S. Romdhani, "Face image analysis using a multiple feature fitting strategy," Ph.D. dissertation, Univ. Basel, Basel, Switzerland, Jan. 2005.

[36] S. Romdhani, V. Blanz, and T. Vetter, "Face identification by fitting a 3D morphable model using linear shape and texture error functions," in *Proc. Eur. Conf. Computer Vision*, 2002, pp. 3–19.

[37] S. Romdhani and T. Vetter, "Efficient, robust and accurate fitting of a 3-D morphable model," in *Proc. Int. Conf. Computer Vision*, 2003, pp. 59–66.

[38] ——, "Estimating 3-D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 986–993.

[39] A. Shashua, "On photometric issues in 3-D visual recognition form a single image," *Int. J. Comput. Vis.*, vol. 21, pp. 99–122, 1997.

[40] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination and expression (PIE) database of human faces," CMU, Tech. Rep., 2000.

[41] ——, "The CMU pose, illumination and expression (PIE) database," in *Proc. IEEE Conf. Automatic Facial and Gesture Recognition*, 2002, pp. 53–58.

[42] T. Sim and T. Kanade, "Combining models and exemplars for face recognition: An illuminating example," presented at the Workshop Models Versus Exemplars in Computer Vision, Kauai, HI, 2001.

[43] L. Sirovitch and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Amer. A*, vol. 2, pp. 519–524, 1987.

[44] K. Torrance and E. Sparrow, "Theory for off-specular reflection from roughened surfaces," *J. Opt. Soc. Amer.*, vol. 57, no. 9, pp. 1105–1114, 1967.

[45] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–96, 1991.

[46] T. Vetter, "Learning novel views to a single face image," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, 1996, pp. 22–27.

[47] T. Vetter and V. Blanz, "Estimating coloured 3-D face models from single images: An example based approach," in *Proc. Eur. Conf. Computer Vision (ECCV'98)*, pp. 499–513.

[48] T. Vetter and T. Poggio, "Linear object classes and image synthesis from a single example image," *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 19, no. 7, pp. 733–742, Jul. 1997.

[49] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, Jan. 2002.

[50] Z. Yue, W. Zhao, and R. Chellappa, "Pose-encoded spherical harmonics for robust face recognition using a single image," in *Proc. IEEE Int. Workshop Analysis and Modeling of Faces and Gestures*, 2005, pp. 229–243.

[51] A. Yuille and D. Snow, "Shape and albedo from multiple images using integrability," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 158–164.

[52] L. Zhang and D. Samaras, "Face recognition under variable lighting using harmonic image examplars," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, pp. 19–25.

[53] ——, "Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 351–363, Mar. 2006.

[54] L. Zhang, S. Wang, and D. Samaras, "Face synthesis and recognition from a single image under arbitrary unknown lighting using a spherical harmonic basis morphable model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 209–216.

[55] R. Zhang, P. Tsai, J. Cryer, and M. Shah, "Shape from shading: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 8, pp. 690–706, Aug. 1999.

[56] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

[57] Q. Zheng and R. Chellappa, "Estimation of illuminant direction, albedo and shape from shading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 7, pp. 680–702, Jul. 1991.

[58] K. Zhou, R. Chellappa, and D. Jacobs, "Characterization of human faces under illumination variations using rank integrability, and symmetry constraints," in *Proc. Eur. Conf. Computer Vision*, 2004, pp. 588–601.

[59] S. K. Zhou and R. Chellappa, "Image-based face recognition under illumination and pose variations," *J. Opt. Soc. Amer. A*, vol. 22, pp. 217–229, 2005.

ABOUT THE AUTHORS

Sami Romdhani studied electronics engineering at the Université Libre de Bruxelles, Belgium and received the M.Sc. degree in electronics engineering at the University of Glasgow, Glasgow, U.K. and the Ph.D. degree from the University of Basel, Switzerland, in 2005.

In 1998, he started his research on face image analysis at the University of Westminster, U.K., and joined the University of Freiburg, Germany, in 2000, where he started working on 3-D Morphable Models. He is currently a Postdoctoral Researcher at the University of Basel, and works mainly on extending computer vision methods using accurate probabilistic prior models. His research interests include image modeling and understanding, inverse rendering, and 3-D modeling and reconstruction.

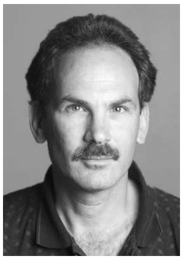
Jeffrey Ho (Member, IEEE) received the M.S. degree in computer science in 2000 and the Ph.D. degree in mathematics in 1999 from the University of Illinois at Urbana-Champaign, Urbana.

He spent the next four years working as a Postdoctoral Researcher first at Beckman Institute and then at the University California at San Diego. He is now an Assistant Professor in the Department of Computer Information Science and Engineering at the University of Florida, Gainesville. He works mainly in computer vision, and his areas of interest include face recognition, visual tracking, and 3-D reconstruction.

Thomas Vetter (Member, IEEE) received the Ph.D. degree in biophysics from the University of Ulm, Germany.

As a Postdoctoral Researcher at the Center for Biological and Computational Learning at the Massachusetts Institute of Technology, he started his research on computer vision. In 1993, he moved to the Max- Planck-Institut in Tubingen, Germany, and, in 1999, he became a professor of computer graphics at the University of Freiburg, Germany. Since 2002, he has been a professor of applied computer science at the University of Basel, Basel, Switzerland. His current research is in image understanding, graphics, and automated model building.

Prof. Vetter is a member of the IEEE Computer Society.



David J. Kriegman received the B.S.E. degree in electrical engineering and computer science from Princeton University, Princeton, NJ, in 1983 and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1984 and 1989, respectively.

He was an Assistant and Associate Professor of Electrical Engineering and Computer Science at Yale University (1990–1998) and an Associate Professor with the Computer Science Department and Beckman Institute at the University of Illinois at Urbana-Champaign (1998–2002). Since 2002, he has been a Professor of Computer Science and Engineering in the Jacobs School of Engineering at the University of California at San Diego, La Jolla (UCSD). He has published over 140 papers on object recognition, reconstruction, illumination, structure from motion, face recognition, microscopy, computer graphics, and robotics.

Prof. Kriegman was chosen for a National Science Foundation Young Investigator Award in 1992, and has received best paper awards at the 1996 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) and the 1998 European Conference on Computer Vision as well as the 2003 Paper of the Year Award from the Journal of Structural Biology. Kriegman has served as Program Co-chair of CVPR 2000 and General Co-chair of CVPR 2005. Currently, he is the Editor-in-Chief of the IEEE Transactions on Pattern Analysis and Machine Intelligence.