The Quotient Image: Class-Based Re-Rendering and Recognition with Varying Illuminations

Amnon Shashua, *Member*, *IEEE*, and Tammy Riklin-Raviv

Abstract—The paper addresses the problem of "class-based" image-based recognition and rendering with varying illumination. The rendering problem is defined as follows: Given a single input image of an object and a sample of images with varying illumination conditions of other objects of the same general class, re-render the input image to simulate new illumination conditions. The class-based recognition problem is similarly defined: Given a single image of an object in a database of images of other objects, some of them are multiply sampled under varying illumination, identify (match) any novel image of that object under varying illumination with the single image of that object in the database. We focus on Lambertian surface classes and, in particular, the class of human faces. The key result in our approach is based on a definition of an illumination invariant signature image which enables an analytic generation of the image space with varying illumination. We show that a small database of objects—in our experiments as few as two objects—is sufficient for generating the image space with varying illumination of any new object of the class from a single input image of that object. In many cases, the recognition results outperform by far conventional methods and the re-rendering is of remarkable quality considering the size of the database of example images and the mild preprocess required for making the algorithm work.

Index Terms—Visual recognition, image-based rendering, photometric alignment.

1 Introduction

Consider the image space generated by applying a source of variability, say changing illumination or changing viewing positions, on a 3D object or scene. Under certain circumstances the images generated by varying the parameters of the source can be represented as a function of a small number of sample images from the image space. For example, the image space of a 3D Lambertian surface is determined by a basis of three images, ignoring cast-shadows [18], [19], [9], [4], [12], [17]. In this case, the low-dimensionality of the image space under lighting variations is useful for synthesizing novel images given a small number of model images or, in other words, provides the means for an "image-based rendering" process in which sampled images replace geometric entities formed by textured micropolygons for rendering new images.

Visual recognition and image re-rendering (synthesis) are intimately related. Recognizing a familiar object from a single picture under some source of variation requires a handle on how to capture the image space created by that source of variation. In other words, the process of visual recognition entails an ability to capture an equivalence class relationship that is either "generative," i.e., create a new image from a number of example images of an object, or "invariant," i.e., create a "signature" of the object that remains invariant under the source of variation under consideration. For example, in a generative process a set of basis images may form a compact representation of the image space. A novel input image is then considered part of the image space if it can be synthesized from the set of basis images. In a process based on invariance, on the other hand, the signature may be a "neutral" image, say the object under a canonical lighting condition or viewing position. A novel image is first transformed into its neutral form and then matched against the database of (neutral) images.

In this paper, we focus on recognition and image re-rendering under lighting condition variability of a *class* of objects, i.e., objects that belong to a general class, such as the class of faces. In other words, for the re-rendering task, given sample images of members of a class of objects and a *single* image of a new object of the class, we wish to render new images of the new object that simulate changing lighting conditions.

Our approach is based on a new result showing that the set of all images generated by varying lighting conditions on a collection of Lambertian objects all having the same shape but differing in their surface texture (albedo) can be characterized analytically using images of a prototype object and a (illumination invariant) "signature" image per object of the class. The Cartesian product between the signature image of an object $\mathbf{y}$ and the linear subspace determined by the images of the prototype object generates the image space of $\mathbf{y}$ (Proposition 1). The second result is on how to obtain the signature image from a data base of example images of several objects while proving that the signature image obtained is invariant to illumination conditions (Theorems 1 and 2).

Our method has two advantages: First and foremost, the method works remarkably well on real images (of faces) using a very small set of example objects—as few as two example objects (see Fig. 7). The re-rendering results are, in

• *The authors are with the Institute of Computer Science, The Hebrew University, Jerusalem 91904, Israel.*
E-mail: shashua@cs.huji.ac.il, tammyr@tiogatech.com.

Manuscript received 15 July 1999; revised 1 Aug. 2000; accepted 2 Aug. 2000. Recommended for acceptance by D.J. Kriegman.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 110241.

many cases, indistinguishable from the "real" thing and the recognition results outperform by far conventional methods. Second, since our approach is based on a simple and clean theoretical foundation, the limitations and breaking points can be clearly distinguished thus further increasing this algorithm's practical use.

1.1 Related Work

The basic result about the low-dimensionality of the image space under varying lighting conditions was originally reported in [18], [19] in the case of Lambertian objects. Applications and related systems were reported in [9], [4], [8]. Re-rendering under more general assumptions, yet exploiting linearity of light transport, was reported in [12], [17].

Work on "class-based" synthesis and recognition of images (mostly with varying viewing positions) was reported in [5], [3], [7], [27], [26], [24], [25], [6], [2], [15]. These methods adopt a "reconstructionist" approach (also see Section 3) in which a necessary condition for the process of synthesis is that the original novel image be generated, reconstructed, from the database of examples. For example, the "linear class" of [27], [13] works under the assumption that 3D shapes of objects in a class are closed under linear combinations (in 3D). Recently, Sali and Ullman [16] have proposed to carry an additive error term, the difference between the novel image and the reconstructed image from the example database. During the synthesis process, the error term is modified as well, thus compensating for the difference between the image space that can be generated from the database of examples and the desired images. Their error term is somewhat analogous to our signature image. However, instead of an error term, we look for an illumination invariant term (signature image) that makes for the difference (in a multiplicative sense) between the image space spanned by a single prototype (or reference) object and the novel image. The database of examples is used for recovering a number of parameters required for generating the signature image.

2 Background and Definitions

We will restrict our consideration to objects with a Lambertian reflectance function, i.e., the image can be described by the product of the albedo (texture) and the cosine angle between a point light source and the surface normal: $\rho(x,y)n(x,y)^\top s$, where $0 \leq \rho(x,y) \leq 1$ is the surface reflectance (gray-level) associated with point x, y in the image, $n(x,y)$ is the surface normal direction associated with point x, y in the image, and s is the (white) light source direction (point light source) and whose magnitude is the light source intensity.

The basic result we will use in this paper is that the image space generated by varying the light source vector s lives in a three-dimensional linear subspace [18], [19]. To see why this is so consider three images I_1, I_2, I_3 of the same object (ρ, n are fixed) taken under linearly independent light source vectors s_1, s_2, s_3, respectively. The linear combination $\sum_j \alpha_j I_j$ is an image $I = \rho n^\top s$, where $s = \sum_j \alpha_j s_j$. Thus, ignoring shadows, three images are sufficient for generating the image space of the object. The basic principle can be extended to deal with shadows, color images, nonwhite light sources, and non-Lambertian surfaces [19], [12], [8], but will not be considered here as our approach can be likewise extended. This principle has been proven robust and successfully integrated in recognition schemes [19], [8], [4]. See Fig. 7 for an example of using this principle for image synthesis.

Next, we define what is meant by a "class" of objects. In order to get a precise definition with which we can base analytic methods on, we define what we call an "ideal" class as follows:

Definition 1 (Ideal Class of Objects). *An ideal class is a collection of 3D objects that have the same shape but differ in the surface albedo function. The image space of such a class is represented by:*

$$\rho_i(x,y)n(x,y)^T s_j,$$

where $\rho_i(x,y)$ is the albedo (surface texture) of object i of the class, $n(x,y)$ is the surface normal (shape) of the object (the same for all objects of the class), and s_j is the point light source direction, which can vary arbitrarily.

In practice, objects of a class do have shape variations, although to some coarse level the shape is similar; otherwise, we would not refer to them as a "class." The ideal class could be satisfied if we perform pixel-wise dense correspondence between images (say frontal images) of the class. The dense correspondence compensates for the shape variation and leaves only the texture variation. For example, Vetter et al. [25] have adopted such an approach in which the flow field and the texture variation were estimated simultaneously during the process of synthesizing novel views from a single image and a (pixel-wise prealigned) database. The question we will address during the experimental section is what is the degree of sensitivity of our approach to deviations from the ideal class assumption. Results demonstrate that one can tolerate significant shape changes without noticeable degradation in performance or, in other words, there is no need to establish any dense alignment among the images beyond the alignment of the center of mass and scale.

From now on when we refer to a class of objects, we mean an "ideal" class of objects as defined above. We will develop our algorithms and correctness proofs under the ideal class assumption. Next, we define the "recognition" and "synthesis" (re-rendering) problems.

Definition 2 (Recognition Problem). *Given $N \times 3$ images of N objects under three lighting conditions and $M \gg N$ other objects of the same class illuminated under some arbitrary light conditions (each), identify the $M + N$ objects from a single image illuminated by some novel lighting conditions.*

Note that we require a small number N of objects, three images per object, in order to "bootstrap" the process. We will refer to the $3N$ images as the "bootstrap set." The synthesis problem is defined similarly.

Definition 3 (Synthesis (Re-Rendering) Problem). *Given $N \times 3$ images of N objects of the same class, illuminated under three distinct lighting conditions and a single image of a novel object of the class illuminated by some arbitrary lighting condition, synthesize new images of the object under new lighting conditions.*

(a) (b) (c)

Fig. 1. Illustration of the "reconstructionist" approach. (a) Original image, (b) image reconstructed from the the bootstrap set of Fig. 2, and (c) image reconstructed from a larger bootstrap set of 20 objects (60 images). The reconstruction is poor in both cases. See text for further details.

To summarize up to this point, given the ideal class and the synthesis/recognition problem definitions above, our goal is: *to extend the linear subspace result of [19] that deals with spanning the image space $\rho n^\top s$, where only s varies, to the case where both ρ and s vary*. We will do so by showing that it is possible to map the image space of one object of the class onto any other object, via the use of an illumination invariant signature image. The recovery of the signature image requires a bootstrap set of example images, albeit a relatively small one (as small as images generated from two objects in our experiments). The remainder of the paper deals with exactly this problem. We first describe a "brute-force" approach for addressing the inherent bilinearity of the problem, detailed next, and then proceed to the main body of this paper.

3 A Reconstructionist Approach and Its Shortcomings

We would like to span the image space $\rho n^\top s$, where both ρ and s vary. Let s_1, s_2, s_3 be a basis of three linearly independent vectors, thus $s = \sum_j x_j s_j$ for some coefficients $x = (x_1, x_2, x_3)$. Let $\rho_1, \ldots, \rho_N$ be a basis for spanning all possible albedo functions of the class of objects, thus $\rho = \sum_i \alpha_i \rho_i$ for some coefficients $\alpha_1, \ldots, \alpha_N$. Let y_s be the image of some new object $\mathbf{y}$ of the class with albedo ρ_y and illuminated by illumination s, i.e.,

$$y_s = \rho_y n^\top s = \left(\sum_{i=1}^{N} \alpha_i \rho_i\right) n^\top \left(\sum_{j=1}^{3} x_j s_j\right).$$

Let $A_1, \ldots, A_N$ be $m \times 3$ matrices whose columns are the images of object i, i.e., the columns of A_i are the images $\rho_i n^\top s_1, \rho_i n^\top s_2, \rho_i n^\top s_3$. We assume that all images are of the same size and contain m pixels. Therefore, we have

$$\min_{x,\alpha_i} \mid y_s - \sum_{i=1}^{N} \alpha_i A_i x \mid^2, \tag{1}$$

which is a bilinear problem in the $N+3$ unknowns x, α_i (which can be determined up to a uniform scale). Clearly, if we solve for these unknowns, we can then generate the image space of object $\mathbf{y}$ from any desired illumination condition simply by keeping α_i fixed and varying x.

One way to solve for the unknowns is first to solve for the pairwise product of x and α_i, i.e., a set of $3N$ variables $z = (\alpha_1 x, \ldots, \alpha_N x)$. Let $A = [A_1, \ldots, A_N]$ be the $m \times 3N$ matrix (we assume $m >> 3N$) obtained by stacking the matrices A_i column-wise. Thus, the vector z can be obtained by the pseudoinverse $A^\# = (A^\top A)^{-1} A^\top$ as the least-squares solution $z = A^\# y_s$. From z, we can decouple x and α_i as follows: Since the system is determined up to scale, let $\sum_i \alpha_i = 1$. Then, group the entries of z into $z = (z_1, \ldots, z_N)$, where z_i is a vector of size 3. We have,

$$x = \sum_{i=1}^{N} z_i$$

and

$$\alpha_i = \frac{1}{3} \sum_{j=1}^{3} \frac{z_{ij}}{x_j}.$$

There are a number of observations that are worth making. First, this approach is a "reconstructionist" one in the sense that one is attempting to reconstruct the image y_s from the dataset of example images, the bootstrap set (for example, [25], [24], [7]). In practice, especially when the size of the bootstrap set is relatively small, $Az \neq y_s$. Moreover, for the same reasons, the decoupling of the variables x_j and α_i from the vector z adds another source of error. Therefore, before we begin creating synthetic images (by varying x_j), we are faced with the problem of having only some approximate rendering of the original image y_s. This problem is acute for small bootstrap sets and, therefore, this approach makes practical sense only for large example sets. The second point to note is that there is some lack of "elegance" (which inevitably contributes to lack of numerical stability and statistical bias due to overfitting[1]) in blowing up the parameter space from $N+3$ to $3N$ in order to obtain a linear least-squares solution.

We illustrate the reconstructionist approach in practice in Fig. 1. We use a bootstrap set of 10 objects (30 images) displayed in Fig. 2, and a bootstrap set of 20 objects (not displayed here). The results of reconstruction are poor for both sets, although one notices some improvement with the larger set of 20 objects. The poor reconstruction is attributed to two main sources. First, is the size of the database. A database of 10 (or 20) objects is apparently not sufficient for capturing the variation among objects in the class. Second, and probably a more dominant source, is the lack of dense pixel-wise alignment among the database and the novel image. Previous work by [26], [24], [25] demonstrate very good results with large databases (around 100 objects) under pixel-wise alignment. The bilinear model proposed by [7] does have implicitly $N+3$ parameters for representing the novel image but requires more parameters for fitting the database images to the model. Hence, the performance of the Freeman and Tenenbaum's bilinear model should be stronger than the simplistic reconstructionist approach demonstrated above, but weaker than the performance of the Qimage approach described in the sequel.

1. Numerical problems due to "blowing" up parameter space for purpose of linearization can be reduced by solving a *heteroscedastic* optimization problem [10], which could be quite unwieldy for large systems.

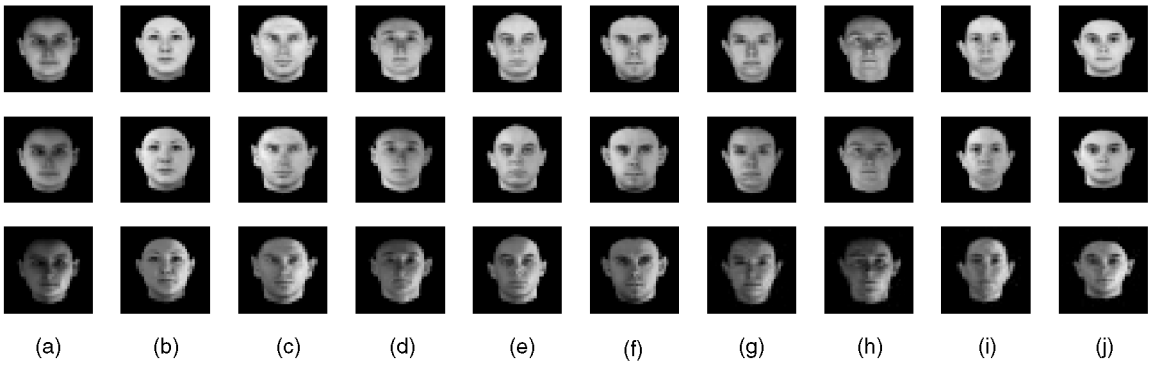


Fig. 2. The bootstrap set of 10 objects from Vetter's database of 200 objects.

In our approach, detailed below, we achieve two major goals: First, we do not make a reconstructionist assumption and thereby tolerate small databases without pixel-wise alignment. Second, we solve (linearly) for a system of $N+3$ parameters (instead of $3N$). As a byproduct of the method of optimization, we obtain an intermediate image, an illumination invariant signature image which can also be used for purposes of visual recognition.

4 The Quotient Image Method

Given two objects $\mathbf{a}, \mathbf{b}$, we define the quotient image Q by the ratio of their albedo functions ρ_a/ρ_b. Clearly, Q is illumination invariant. In the absence of any direct access to the albedo functions, we show that Q can nevertheless be recovered, analytically, given a bootstrap set of images. Once Q is recovered, the entire image space (under varying lighting conditions) of object $\mathbf{a}$ can be generated by Q and three images of object $\mathbf{b}$. The details are below.

We will start with the case $N=1$, i.e., there is a single object (three images) in the bootstrap set. Let the albedo function of that object $\mathbf{a}$ be denoted by ρ_a and let the three images be denoted by a_1, a_2, a_3, therefore, $a_j = \rho_a n^\top s_j$, $j=1,2,3$. Let $\mathbf{y}$ be another object of the class with albedo ρ_y and let y_s be an image of $\mathbf{y}$ illuminated by some lighting condition s, i.e., $y_s = \rho_y n^\top s$. We define below an illumination invariant signature image Q_y of $\mathbf{y}$ against the bootstrap set (in this case, against $\mathbf{a}$).

Definition 4 (Quotient Image). *The quotient image Q_y of object $\mathbf{y}$ against object $\mathbf{a}$ is defined by*

$$Q_y(u,v) = \frac{\rho_y(u,v)}{\rho_a(u,v)},$$

where u, v range over the image.

Thus, the image Q_y depends only on the relative surface texture information and, thus, is independent of illumination. The reason we represent the relative change between objects by the ratio of surface albedos becomes clear from the proposition below:

Proposition 1. *Given three images a_1, a_2, a_3 of object $\mathbf{a}$ illuminated by any three linearly independent lighting conditions and an image y_s of object $\mathbf{y}$ illuminated by some light source s, then there exists coefficients x_1, x_2, x_3 that satisfy*

$$y_s = \left(\sum_j x_j a_j\right) \otimes Q_y,$$

where $\otimes$ denotes the Cartesian product (pixel by pixel multiplication). Moreover, the image space of object $\mathbf{y}$ is spanned by varying the coefficients.

Proof. Let x_j be the coefficients that satisfy $s = \sum_j x_j s_j$. The claim $y_s = (\sum_j x_j a_j) \otimes Q_y$ follows by substitution. Since s is arbitrary, the image space of object $\mathbf{y}$ under changing illumination conditions is generated by varying the coefficients x_j. □

We see that once Q_y is given, we can generate y_s (the novel image) and all other images of the image space of $\mathbf{y}$. The key is obtaining the quotient image Q_y. Given y_s, if somehow we were also given the coefficients x_j that satisfy $s = \sum_j x_j s_j$, then Q_y readily follows: $Q_y = y_s/(\sum_j x_j a_j)$, thus the key is to obtain the correct coefficients x_j. *For that reason, and that reason only, we need the bootstrap set*—otherwise, a single object $\mathbf{a}$ would suffice (as we see above).

Let the bootstrap set of $3N$ pictures be taken from three fixed (linearly independent) light sources s_1, s_2, s_3 (the light sources are not known). Let A_i, $i=1,\ldots,N$, be a matrix whose columns are the three pictures of object $\mathbf{a_i}$ with albedo function ρ_i. Thus, $A_1, \ldots, A_N$ represent the bootstrap set of N matrices, each is a $m \times 3$ matrix, where m is the number of pixels of the image (assuming that all images are of the same size). Let y_s be an image of some novel object $\mathbf{y}$ (not part of the bootstrap set) illuminated by some light source $s = \sum_j x_j s_j$. We wish to recover $x = (x_1, x_2, x_3)$ given the N matrices $A_1, \ldots, A_N$ and the vector y_s.

We define the *normalized albedo* function ρ of the bootstrap set as:

$$\rho(u,v) = \sum_{i=1}^{N} \rho_i^2(u,v)$$

which is the sum of squares of the albedos of the bootstrap set. In cases where there exist coefficients $\alpha_1, \ldots, \alpha_N$ such that

$$\frac{\rho(u,v)}{\rho_y(u,v)} = \alpha_1\rho_1(u,v) + \ldots + \alpha_N\rho_N(u,v),$$

where ρ_y is the albedo of the novel object $\mathbf{y}$, we say that ρ_y is in the *rational span* of the bootstrap set of albedos. With these definitions, we show the major result of this paper: If the albedo of the novel object is in the rational span of the bootstrap set, we describe an energy function $f(\hat{x})$ whose global minimum is at x, i.e., $x = \operatorname{argmin} f(\hat{x})$.

Theorem 1. *The energy function*

$$f(\hat{x}) = \frac{1}{2}\sum_{i=1}^{N} \mid A_i\hat{x} - \alpha_i y_s \mid^2 \tag{2}$$

has a (global) minimum $\hat{x} = x$, if the albedo ρ_y of object $\mathbf{y}$ is rationally spanned by the bootstrap set, i.e., if there exist $\alpha_1, \ldots, \alpha_N$ such that

$$\frac{\rho}{\rho_y} = \alpha_1\rho_1 + \ldots + \alpha_N\rho_N.$$

Proof. Let $\hat{s} = \sum_j \hat{x}_j s_j$, thus, $A_i\hat{x} = \rho_i n^\top \hat{s}$. In vectorized form:

$$A_i\hat{x} = \begin{bmatrix} \rho_{i1}n_1^\top \\ \rho_{i2}n_2^\top \\ \cdot \\ \cdot \\ \cdot \\ \rho_{im}n_m^\top \end{bmatrix} \hat{s} = W_i\hat{s},$$

where $\rho_{i1}, \ldots, \rho_{im}$ are the entries of ρ_i in vector format. The optimization function $f(\hat{x})$ can be rewritten as a function $g(\hat{s})$ of $\hat{s}$:

$$\begin{aligned} g(\hat{s}) &= \frac{1}{2}\sum_{i=1}^{N} \mid W_i\hat{s} - \alpha_i W_y s\mid^2 \\ &= \sum_i \frac{1}{2}\hat{s}^\top W_i^\top W_i\hat{s} + \sum_i \alpha_i \hat{s}^\top W_i^\top W_y s \\ &\quad + \sum_i \frac{1}{2}\alpha_i^2 s^\top W_y^\top W_y s, \end{aligned}$$

where W_y is defined similarly to W_i by replacing the albedo ρ_i by ρ_y. Because the variables of optimization $\hat{x}, \hat{s}$ in $f(\hat{x})$ and in $g(\hat{s})$ are linearly related, it is sufficient to show that the global minimum of $g(\hat{s})$ is achieved when $\hat{s} = s$. We have,

$$0 = \frac{\partial g}{\partial \hat{s}} = \left(\sum_i W_i^\top W_i\right)\hat{s} - \left(\sum_i \alpha_i W_i^\top\right) W_y s.$$

Hence, we need to show that

$$\sum_i W_i^\top W_i = \left(\sum_i \alpha_i W_i^\top\right) W_y.$$

We note that,

$$W_i^\top W_i = \rho_{i1}^2 n_1 n_1^\top + \ldots + \rho_{im}^2 n_m n_m^\top.$$

Thus, we need to show

$$\begin{aligned} &\left(\sum_i \rho_{i1}^2\right) n_1 n_1^\top + \ldots + \left(\sum_i \rho_{im}^2\right) n_m n_m^\top \\ &= \left(\sum_i \alpha_i\rho_{i1}\right)\rho_{y1} n_1 n_1^\top + \ldots + \left(\sum_i \alpha_i\rho_{im}\right)\rho_{ym} n_m n_m^\top. \end{aligned}$$

Note that the coefficients of the left-hand side are the entries of the normalized albedo ρ. Thus, we need to show that

$$\sum_{i=1}^{N} \rho_{ik}^2 = \left(\sum_{i=1}^{N} \alpha_i\rho_{ik}\right)\rho_{yk}$$

for all $k = 1, \ldots, m$. But this holds, by definition, because ρ_y is rationally spanned by $\rho_1, \ldots, \rho_N$. □

The proof above was not constructive, it only provided the existence of the solution as the global minimum of the energy function $f(\hat{x})$. Finding $\min f(\hat{x})$ is a simple technicality (a linear least-squares problem), but note that the system of equations is simplified due to substitution while decoupling the role of $\hat{x}$ and the coefficients α_i. This is shown below.

Theorem 2. *The global minima x_o of the energy function $f(\hat{x})$ is*

$$x_o = \sum_{i=1}^{N} \alpha_i v_i,$$

where

$$v_i = \left(\sum_{r=1}^{N} A_r^\top A_r\right)^{-1} A_i^\top y_s$$

and the coefficients α_i are determined up to a uniform scale as the solution of the symmetric homogeneous linear system of equations

$$\alpha_i y_s^\top y_s - \left(\sum_{r=1}^{N} \alpha_r v_r\right)^\top A_i^\top y_s = 0$$

for $i = 1, \ldots, N$.

Proof.

$$0 = \frac{\partial f}{\partial \hat{x}} = \left(\sum_i A_i^\top A_i\right)\hat{x} - \left(\sum_i \alpha_i A_i^\top\right) y_s$$

from which it follows that:

$$\hat{x} = \left(\sum_i A_i^\top A_i\right)^{-1}\left(\sum_i \alpha_i A_i^\top\right) y_s = \sum_i \alpha_i v_i.$$

We also have

$$0 = \frac{\partial f}{\partial \alpha_i} = \alpha_i y_s^\top y_s - \hat{x}^\top A_i^\top y_s,$$

which following the substitution $\hat{x} = \sum_i \alpha_i v_i$, we obtain a homogeneous linear system for $\alpha_1, \ldots, \alpha_N$:

$$\alpha_i y_s^\top y_s - (\sum_r \alpha_r v_r)^\top A_i^\top y_s = 0$$

for $i = 1, \ldots, N$. Written explicitly,

$$\begin{array}{llll} \alpha_1(v_1^\top A_1^\top y_s - y_s^\top y_s) & + \ldots + & \alpha_N v_N^\top A_1^\top y_s & = 0 \\ \alpha_1 v_1^\top A_2^\top y_s & + \ldots + & \alpha_N v_N^\top A_2^\top y_s & = 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \alpha_1 v_1^\top A_N^\top y_s & + \ldots + & \alpha_N(v_N^\top A_N^\top y_s - y_s^\top y_s) & = 0. \end{array}$$

Let the estimation matrix (above) be denoted by F, we show next that F is symmetric. The entries F_{ij}, $i \neq j$, have the form:

$$F_{ij} = y_s^\top A_j \left(\sum_r A_r^\top A_r \right)^{-T} A_i^\top y_s = y_s^\top A_j B A_i^\top y_s.$$

Note that B is a symmetric matrix (inverse of a sum of symmetric matrices). Let $E_{ij} = A_j B A_i^\top$, then it is easy to notice that $E_{ji} = E_{ij}^\top$ due to the symmetric property of B. Thus, $F_{ij} = F_{ji}$ because

$$F_{ij} = y_s^\top E_{ij} y_s = (E_{ij} y_s)^\top y_s = y_s^\top E_{ij}^\top y_s = F_{ji}.$$

□

The energy function $f(\hat{x})$ in (2) consists of a simultaneous projection of y_s onto the subspaces spanned by the columns of A_1, columns of A_2, and so on. In addition, during the simultaneous projection there is a choice of overall scale per subspace—these choices of scale, the α_i, are directly related to the scaling of the axes represented by $\rho_1, \ldots, \rho_N$ such that the albedos of the bootstrap set span (rationally) the albedo of the novel object. When $N = 1$, the minimum of $f(\hat{x})$ coincides with x iff the albedo of the novel object is equal (up to scale) to the albedo of bootstrap object. The more objects in the bootstrap set, the more freedom we have in representing novel objects. If the albedos of the class of objects are random signals, then at the limit a bootstrap set of m objects ($3m$ images) would be required to represent all novel objects of the class. In practice, the difference in the albedo functions do not cover a large spectrum and instead occupy a relatively small subspace of m, therefore, a relatively small size $N << m$ is required and that is tested empirically in Section 6.

Once the coefficients x have been recovered, the quotient image Q_y can be defined against the average object: Let $\mathcal{A}$ be a $m \times 3$ matrix defined by the average of the bootstrap set,

$$\mathcal{A} = \frac{1}{N} \sum_{i=1}^{N} A_i,$$

and then the quotient image Q_y is defined by:

$$Q_y = \frac{y_s}{\mathcal{A}x}.$$

To summarize, we describe below the algorithm for synthesizing the image space of a novel object $\mathbf{y}$, given the bootstrap set and a single image y_s of $\mathbf{y}$.

1. We are given N matrices, $A_1, \ldots, A_N$, where each matrix contains three images (as its columns). This is the bootstrap set. We are also given a novel image y_s (represented as a vector of size m, where m is the number of pixels in the image). For good results, make sure that the objects in the images are roughly aligned (position of center of mass and geometric scale).
2. Compute N vectors (of size 3) using the equation:
$$v_i = \left(\sum_{r=1}^{N} A_r^\top A_r \right)^{-1} A_i^\top y_s,$$
where $i = 1, \ldots, N$.
3. Solve the homogeneous system of linear equations in $\alpha_1, \ldots, \alpha_N$ described in (3). Scale the solution such that $\sum_i \alpha_i = N$.
4. Compute $x = \sum_i \alpha_i v_i$.
5. Compute the quotient image $Q_y = y_s / \mathcal{A}x$, where $\mathcal{A}$ is the average of $A_1, \ldots, A_N$. See [14] for more details on noise-handling, such as when there is a division by zero.
6. The image space created by the novel object, under varying illumination, is spanned by the product of images Q_y and $\mathcal{A}z$ for all choices of z.

5 A Note about Color

The process described so far holds for black-and-white images, not color images. We describe a simple approach to handle color images, *while still maintaining a gray-value bootstrap set*. In other words, given a bootstrap set of gray-value images and a color image (represented by RGB channels) y_s of a novel object, we wish to create the *color* image space of that object under varying illumination. To that end, we will make the assumption that varying illumination does not affect the saturation and hue composition of the image, only the gray-value distribution (shades of color) of the image.

Given this assumption, we first must decouple the hue, saturation, and gray-value (lightness) components of the image y_s from its RGB representation. This is achieved by adopting the Hue Saturation Value (HSV) color space [21] often used for splitting color into meaningful conceptual categories. The transformation (nonlinear) from RGB to HSV and vice versa can be found, for example, in MATLAB. The HSV representation decouples the color information into three channels (images): Hue (tint or color bias), Saturation (amount of hue present—decreasing saturation corresponds to adding white pigment to a color), and Value (the luminance, or black-and-white information; the diagonal from $(1,1,1)$ to $(0,0,0)$ of the RGB cube). Saturation can vary from a maximum corresponding to vivid color, to a minimum, which is equivalent to a black-and-white image. Once the H, S, and V images are created (from the R, G, B images), the novel image we work with is simply V. The algorithm above is applied and a synthetic image V' is created (a new image of the object under some novel illumination condition). The corresponding color image is the original H, S, and the new V′. Similar approaches for augmenting black-and-white images using a color prototype image can be found in [15].

This approach allows using only gray-level images in the bootstrap set, yet accommodates the synthesis of color images from a novel color input image. Fig. 8 display examples on synthesizing color images from a gray-value bootstrap set.

6 Experiments

We have conducted a wide range of experimentation on the algorithm presented above. We first used a high quality database prepared by Vetter et al. [25] and Vetter and Poggio [24]. We have chosen a bootstrap collection of 10 objects shown in Fig. 2. The images of the bootstrap set and the novel images to be tested are "roughly" aligned, which means that the center of mass was aligned and scale was corrected (manually).

Our first test, shown in Fig. 3, was to empirically verify that the quotient image is indeed invariant to illumination changes. The Q-images where thresholded (above one standard deviation) for display purposes. One can see that with a bootstrap set of 10 objects one obtains a fairly invariant quotient image in spite of the large variation in the illumination of the novel images tested. The Q-images should also be invariant to the choice of the light sources s_1, s_2, s_3 used in the bootstrap set. This is demonstrated in Fig. 5 where the quotient image was generated against different choices of s_1, s_2, s_3 for the bootstrap object set (Vetter's database includes nine images per object thus enabling us to experiment with various bootstrap sets of the same 10 objects). Note that the novel image that was tested was not part of Vetter's database but an image of one of our lab members.

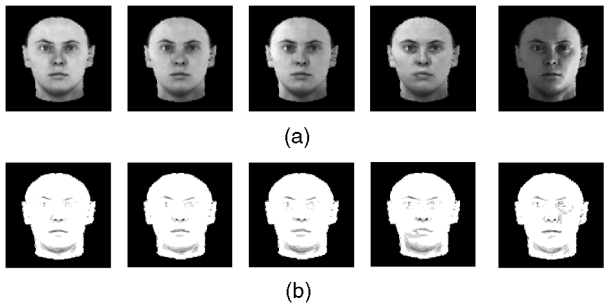


Fig. 3. Testing the invariance of the quotient image to varying illumination. (a) Original images of a novel face taken under five different illuminations. (b) The Q-images corresponding to the novel images above computed with respect to the bootstrap set of Fig. 2.

The next experiment was designed to test the role of the size of the bootstrap set on the accurate determination of the coefficients $x = (x_1, x_2, x_3)$. The accuracy of the coefficient vector x is measured by the invariance of the quotient image against varying illumination, hence Fig. 4 displays Q-images generated by various bootstrap sets, as follows: We have tested the case $N = 1$, i.e., bootstrap set of a single object (row b), compared to a bootstrap set of $N = 10$ but where the reference object is the same object used in case $N = 1$ (instead of the average object), shown in row f.

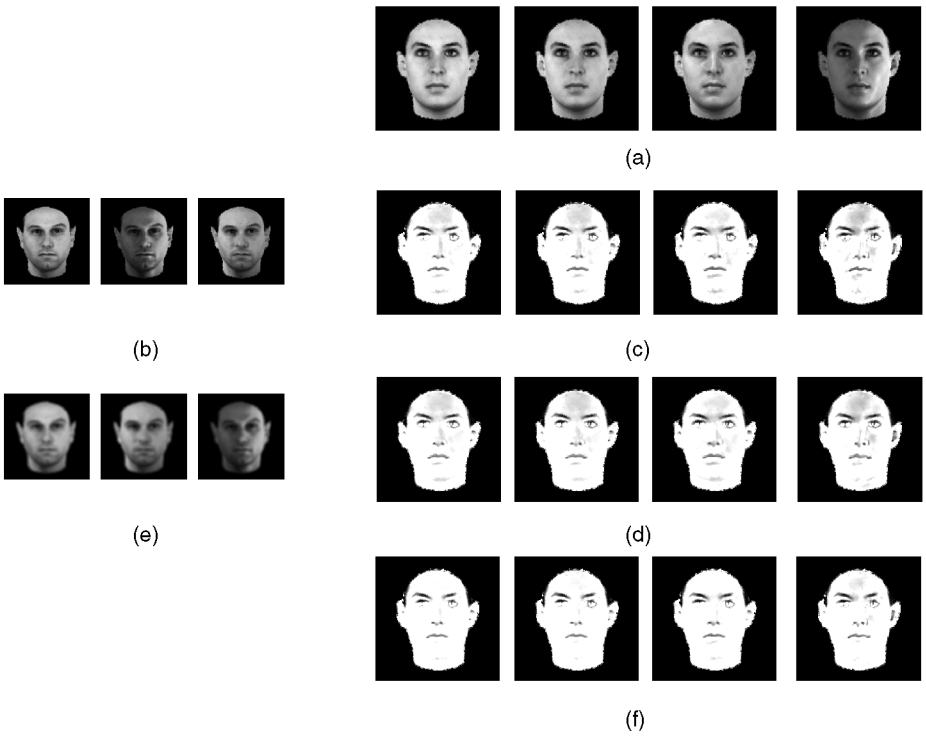


Fig. 4. Testing accuracy of Theorem 1 against the size of the bootstrap set. (a) Original images taken under four distinct light conditions. (b) Bootstrap set of $N = 1$ objects used for generating the Q-images of (a) displayed in row (c). Note that the quotient images are not strictly invariant as they change with the illumination. (d) Q-images of the bootstrap set ($N = 1$) displayed in (e). Note that the bootstrap set is blurred in order to test whether using the "average" object when $N > 1$ makes a difference compared to the machinery described in Theorem 1. We see that blurred images do not improve the invariance of the Q-images. (f) Q-images of (a) against the object (b) but where the coefficient vector x was recovered using the $N = 10$ bootstrap set of Fig. 2. The comparison should be made between rows (c) and (f). Note that in (f), the images are invariant to changing illumination more so than in (c).

Fig. 5. Q-images should be invariant to the three illumination conditions of the database images, as long as they span a three-dimensional subspace. The 3 Q-images were generated against different bootstrap sets of the same 10 objects but of different triplets of light sources. Note that the novel object is not part of the original database of 200 objects, but of a member of our lab.

Therefore, the difference between rows c and f is solely due to the effect of Theorem 1 on computing the coefficient vector x. The result supports the claim of Theorem 1 in the sense that the larger the bootstrap set, the more accurate is the recovery of x. In order to rule out any special influence the average object has on the process (recall that once x has been recovered it was suggested to use the average object ψ as the reference object for the quotient image), we have also tested the case $N = 1$, where the images were deliberately blurred (to simulate an average object), yet the Q-images (row d) have not improved (compared to row c).

In Figs. 6 and 7, we demonstrate the results of image synthesis from a single input image and the bootstrap set. Note the quality and the comparison between results of bootstrap size $N = 10$ and $N = 2$ (there are differences but relatively small).

So far, we have experimented with objects and their images from the same database of 200 objects. Even though the input image is of an object outside the bootstrap set, there is still an advantage by having all the images taken with the same camera, same conditions, and same quality level. Our next experiments were designed to test the algorithm on source images taken from sporadic sources, such as from magazines or from the Web. The bootstrap set in all experiments is the one displayed in Fig. 2.

Fig. 8 shows four novel (color) images of celebrity people (from magazines) and the result of the synthesis procedure. These images are clearly outside the circle of images of the original database of Vetter, for example, the images are not cropped for hair adjustment and the facial details are markedly different from those in the bootstrap set. Finally, we have experimented with other bootstrap sets shown in Fig. 9a. A bootstrap set of three objects varying in hair-style, uncropped, and generally taken under much less attention compared to the bootstrap set of Fig. 2 is sufficient, nevertheless, to generate quite reasonable re-renderings, as shown in Fig. 9d. The degradation is indeed graceful and affects mainly the degree of illumination changes, not as

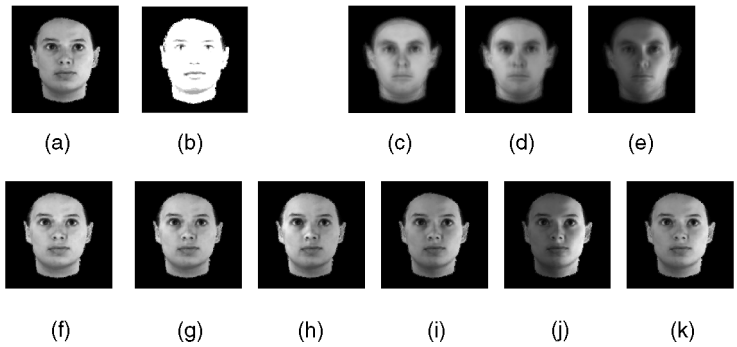


Fig. 6. Image Synthesis Example. (a) Original image and its quotient image (b) from the $N = 10$ bootstrap set. The quotient image is generated relative to the average object of the bootstrap set shown in (c), (d) and (e). Images (f) through (k) are synthetic images created from (b), (c), (d), and (e) using Proposition 1.

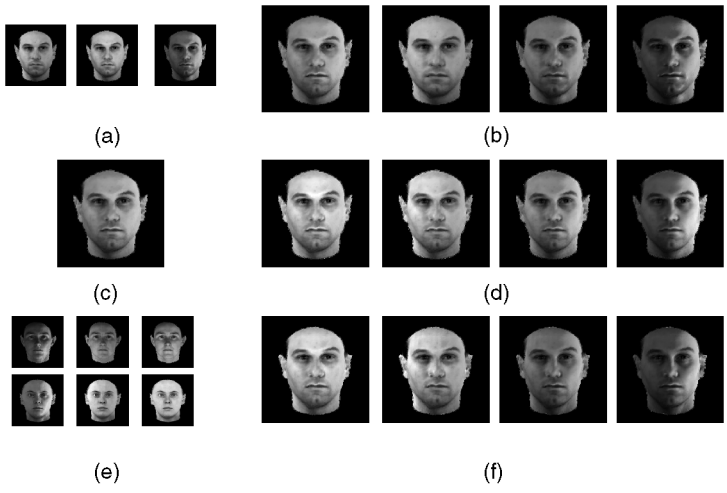


Fig. 7. Image synthesis examples. (a) Original images under three distinct lighting conditions and the synthesized images (b) using linear combinations of those three images. The synthesized images using the original single image (c) and a $N = 10$ bootstrap set are shown in (d). Finally, (e) is an $N = 2$ bootstrap set for generating the synthesized images (f) from the single original image (c).

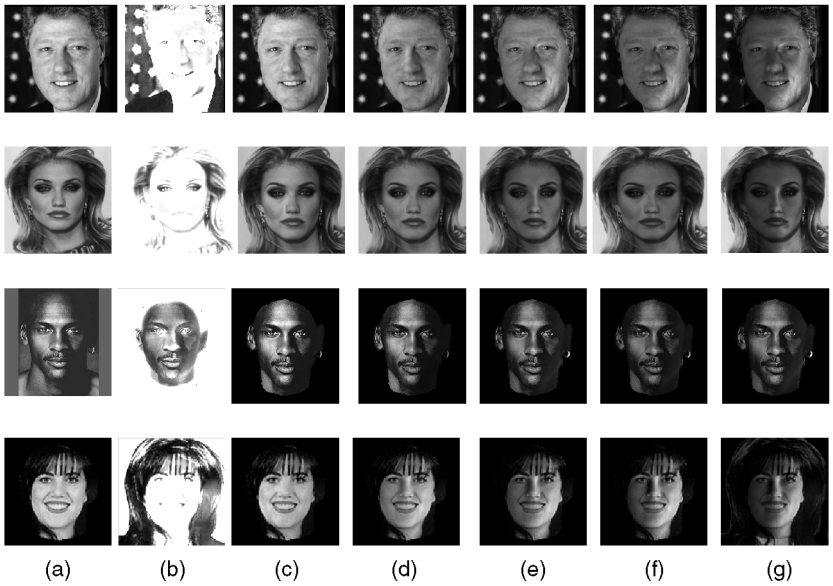


Fig. 8. (a) Original images. (b) Q images and (c) through (g) Synthesized images. (Bill Clinton's image available at http://hopeusa.com/.clinton/believe.html. Cameron Diaz's image available at http://beautiful-women.simplenet.com/cameron. Michael Jordan's image available at http://web4.sportsline.com/u/jordan/. Monica Lewinsky's image available at http:// homepages.newnet.co.uk/epm/photos/html.)

much the quality of the resulting image (compared to the source image).

6.1 When Does the Algorithm Fail?

An inherent assumption throughout the algorithm is that for a given pixel (x, y), $n(x, y)$ is the same for all the images—the bootstrap set as well as the test images. This was referred to in the paper as the *ideal class assumption*. We have seen that the performance for faces is fairly robust despite the fact the ideal class assumption does not strictly hold for roughly aligned images of faces. The performance degrades when dominant features between the bootstrap set and the test set are misaligned. This could arise in a variety of situations such as: 1) the class is of nonsmooth objects like objects with sharp corners (chairs, for instance), 2) objects are seen from varying viewing positions (see [22] for handling such cases with the Qimage approach), and 3) the class of objects is smooth (like human faces) but gross misalignment is caused by facial expressions, mustache, eye-glasses, etc.

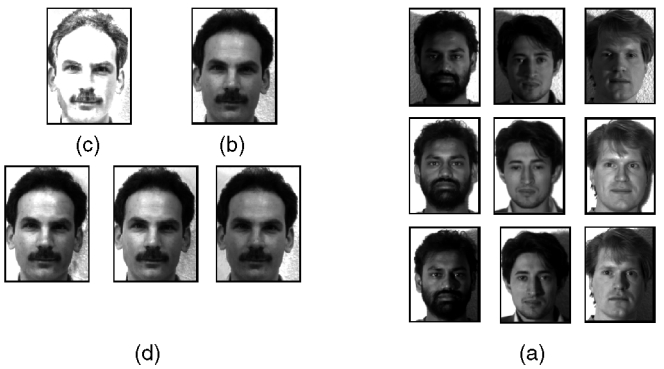


Fig. 9. Image synthesis using other, lower quality, bootstrap sets (Yale data sets). The bootstrap set ($N = 3$) is shown in (a). Note that the objects vary considerably in appearance (hair style and facial hair) and are thus less controlled as in Vetter's data set. The source image (b), its quotient image (c), and synthesized images (d).

7 Other Routes for a Signature Image?

The quotient image approach is based on the idea that an illumination invariant image $Q = \rho_y/\rho_a$ can be used to map the image space of object $\mathbf{a}$ to the image space of object $\mathbf{y}$ using a single image y_s of $\mathbf{y}$. The equation $(\sum_j x_j a_j) \otimes Q$ generates the image space of $\mathbf{y}$ (Proposition 1). There are two points worth making.

First, Q is analogous to an "error correction term." However, it is important to distinguish between error correction and an illumination invariant term. For example, let $\hat{y}$ be the reconstructed image of y_s from the bootstrap set (after solving for x, α_i that minimize (1) in the "reconstructionist" approach), and let $\bar{Q}$ be defined such that $y_s = \hat{y} \otimes \bar{Q}$. *There is no reason to expect that $\bar{Q}$ would be illumination invariant*. This is demonstrated in Fig. 10b showing that the $\bar{Q}$ images are not invariant to changing illumination. In other words, one would not obtain an admissible image space of $\mathbf{y}$, or correct re-rendering, if we simply correct for the reconstruction error by a Cartesian product with $\bar{Q}$.

Second, notice that the optimization criteria described in Theorem 1 involves a somewhat complex definition of what constitutes a "family" of albedo functions (rational span). This is unlike the more intuitive definition, that one would typically adopt under such circumstances, that albedo functions are closed under linear combinations (the definition adopted in the optimization criteria behind (1) for the "reconstructionist" approach). However, the rational span definition has an important role because through it we were able to remove of the intrinsic bilinearity among the illumination parameters $x = (x_1, x_2, x_3)$ and the albedo parameters $\alpha_1, \ldots, \alpha_N$ and obtain a linear system for $N + 3$ variables (instead of $3N$ if the linear span definition

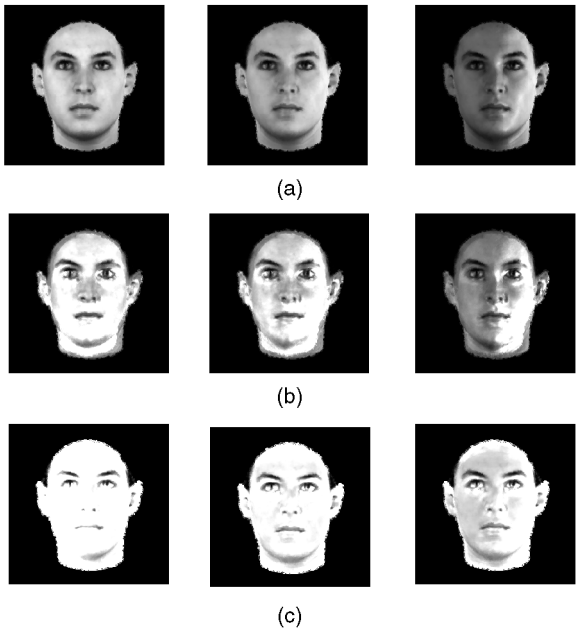


Fig. 10. Alternatives approaches for a quotient image. (a) Original images under varying illumination. (b) Quotient images defined as a multiplicative "error" image, i.e., the ratio of the original image and the least-squares reconstruced image from the bootstrap set. Note that the resulting quotient images are not illumination invariant. (c) Quotient images defined by Proposition 1 where x is the minima of (1) (instead of (2)). Again, the images are not illumination invariant.

were to be adopted). The importance of all this depends on the numerical behavior of the system. In principle, however, one could solve for x from (1) and use it for obtaining the quotient image as defined in Proposition 1. In other words, in the algorithm described in the previous section, simply replace Steps 2 through 4 with the procedure described in Section 3 for obtaining x. We expect a degradation in performance due to numerical considerations (due to the enlargement of parameter space). The results of doing so are illustrated in Fig. 10c. The quotient images clearly show a dependence on illumination change, indicating that the parameters x_1, x_2, x_3 were not recovered well.

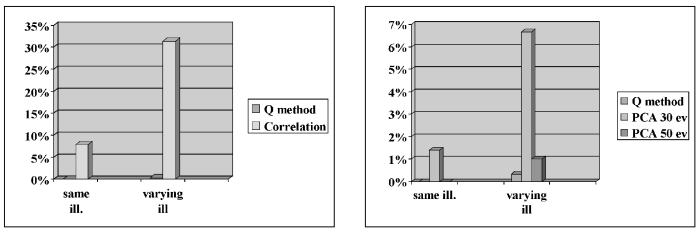
In summary, the combination of an illumination invariant correction term (the quotient image) and a simple optimization criteria (1)—with the price of somewhat complicating the definition of when albedos form a "family"—gives rise to both practical and a provenly correct procedure for class-based re-rendering (under the terms stated for ideal class definition and Lambertian surfaces).

8 Recognition

The Q-images are illumination invariant signatures of the objects in the class. We can therefore make use of the invariance property for purposes of recognition. Vetter's database contains 200 faces each under nine lighting conditions, making a total of 1,800 images. We used a bootstrap set of 20 objects (60 images) and created the Q-images of all the 200 objects—these 200 images serve as the database, we refer to as Q-database, for purposes of recognition. Given any of the 1,800 source images, its Q-image is created from the bootstrap set and matched (by correlation) against the Q-database while searching for the best match.

We made two tests (summarized in Fig. 11). In the first test, the Q-database was generated from images under the same illumination (we have nine images per object in Vetter's database). The results of recognition was compared to correlation where the database for correlation where those images used for creating the Q-database. The match against the Q-database was error free (0 percent). The match against the original images, instead of the Q-images, had 142 mismatches (7.8 percent). In the second test, the images used for creating the Q-database were drawn randomly from the set of nine images (per object). The match against the Q-database produced only six mismatches (.33 percent), whereas the match against the original images produced 565 mismatches (31.39 percent). The sharp increase in the rate of mismatches for the regular correlation approach is due to the dominance of illumination effects on the overall brightness distribution of the image (cf. [19], [1]).

We also made a comparison against the "eigenfaces" approach [20], [11] which involves representing the database by its Principle Components (PCA). In the first test, the PCA was applied to the bootstrap set (60 images) and 180 additional images, one per object. In the first test, the additional images were all under the same illumination and, in the second test, they were drawn randomly from the set of nine images per object. The recognition performance depends on the number of principle components. With 30 principle components (out of 240), the first test had 25 mismatches (1.4 percent), and the second test 120 mismatches (6.6 percent). The performance peaks around 50 principle components in which case the first test was



(a) (b)

Fig. 11. Recognition results on Vetter's database of 1,800 face images. We compare the Q-image method with correlation and Eigenfaces. See text for details.

error free (like in the Q-image method), and the second test had 18 mismatches (1 percent).

To summarize, in all recognition tests, except one test of equal performance with PCA, the Q-image outperforms and, in some cases, in a significant manner, conventional class-based approaches.

9 Summary

We have presented a class-based, image-based, re-rendering and recognition method. The key element of our approach was to show that under fairly general circumstances it is possible to extract from a small set of example images an illumination invariant "signature" image per novel object of the class from a single input image alone. We have proven our results (under the "imaginary" world of ideal class assumption) and demonstrated the applicability of our algorithm on the class of real pictures of human faces. In other words, we have shown that in practice a remarkably small number of sample images of human frontal faces (in some of our experiments, images of two objects were sufficient for making a database) can generate photo-realistic re-rendering of new objects from single images.

The ideas presented in this paper can, without too much difficulty, be turned onto a system for image compositing and relighting of general faces, with very high quality of performance. To that end, further implementation elements may be required, such as using collections of bootstrap sets (while choosing among them manually or automatically using sparse optimization approaches like Support Vector Machines [23]), and automatic or semiautomatic tools for morphing the bootstrap set onto the novel image in order to better compensate for changes of shape (such as [25]).

References

[1] Y. Adini, Y. Moses, and S. Ullman, "Face Recognition: The Problem of Compensating for Changes in Illumination Direction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 721-732, July 1997.

[2] J.J. Atick, P.A. Griffin, and N.A. Redlich, "Statistical Approach to Shape-from-Shading: Deriving 3D Face Surfaces from Single 2D Images," *Neural Computation,* 1997.

[3] R. Basri, "Recognition by Prototypes," *Int'l J. Computer Vision,* vol. 19, no. 2, pp. 147-168, 1996.

[4] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *Proc. European Conf. Computer Vision,* 1996.

[5] D. Beymer and T. Poggio, "Image Representations for Visual Learning," *Science,* vol. 272, pp. 1905-1909, 1995.

[6] S. Edelman, "Class Similarity and Viewpoint Invariance in the Recognition of 3D Objects," *Biological Cybernetics,* vol. 72, pp. 207-220, 1995.

[7] W.T. Freeman and J.B. Tenenbaum, "Learning Bilinear Models for Two-Factor Problems in Vision," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 554-560, 1997.

[8] A.S. Georghiades, D.J. Kriegman, and P.N. Belhumeur, "Illumination Cones for Recognition under Variable Lighting: Faces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 52-59, 1998.

[9] P. Hallinan, "A Low-Dimentional Representation of Human Faces for Arbitrary Lightening Conditions," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 995-999, 1994.

[10] P. Meer and Y. Leedan, "Estimation with Bilinear Constraints in Computer Vision," *Proc. Int'l Conf. Computer Vision,* pp. 733-738, Jan. 1998.

[11] M. Turk and A. Pentland, "Eigen Faces for Recognition," *J. Cognitive Neuroscience,* vol. 3, no. 1, 1991.

[12] J. Nimeroff, E. Simoncelli, and J. Dorsey, "Efficient Re-rendering of Naturally Illuminated Environments," *Proc. Fifth Ann. Eurographics Symp. Rendering,* June 1994.

[13] T. Poggio and T. Vetter, "Recognition and Structure from One 2D Model View: Observations on Prototypes, Object Classes and Symmetries," Technical Report AI Memo 1347, MIT, 1992.

[14] T. Riklin-Raviv, "The Quotient Image: Class-Based Re-rendering and Recognition with Varying Illuminations," master's thesis, School of Computer Science and Eng., 2000.

[15] D.A. Rowland and D. Perrett, "Manipulating Facial Appearance through Shape and Color," *IEEE Computer Graphics and Applications,* pp. 70-76, Sept. 1995.

[16] E. Sali and S. Ullman, "Recognizing Novel 3D Objects under New Illumination and Viewing Position Using a Small Number of Examples," *Proc. Int'l Conf. Computer Vision,* pp. 153-161, 1998.

[17] C. Schoeneman, J. Dorsey, B. Smits, J. Arvo, and D. Greenberg, "Painting with Light," *Computer Graphics Proc., Ann. Conf. Series,* pp. 143-146, 1993.

[18] A. Shashua, "Illumination and View Position in 3D Visual Recognition," *Advances in Neural Information Processing Systems 4,* S. J. Hanson, J.E. Moody, and R.P. Lippmann, eds., pp. 404-411, San Mateo, Calif.: Morgan Kaufmann, 1992.

[19] A. Shashua, "On Photometric Issues in 3D Visual Recognition from a Single 2D Image, *Int'l J. Computer Vision,* vol. 21, pp. 99-122, 1997.

[20] L. Sirovich and M. Kirby, "Low Dimensional Procedure for the Characterization of Human Faces," *J. Optical Soc. Am.,* vol. 4, no. 3, pp. 519-524, 1987.

[21] C. Smith, "Color Gamut Transformation Pairs," *Computer Graphics,* vol. 12, pp. 12-19, 1978.

[22] A. Stoschek, "Image-Based Re-rendering of Faces for Continuous Pose and Illumination Directions," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 582-587, June 2000.

[23] V.N. Vapnik, *The Nature of Statistical Learning.* Springer, 1995.

[24] T. Vetter and V. Blanz, "Estimating Coloured 3D Face Models from Single Images: An Example-Based Approach," *Proc. European Conf. Computer Vision,* pp. 499-513, 1998.

[25] T. Vetter, M.J. Jones, and T. Poggio, "A Bootstrapping Algorithm for Learning Linear Models of Object Classes," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 40-46, 1997.

[26] T. Vetter and T. Poggio, "Image Synthesis from a Single Example View," *Proc. European Conf. Computer Vision,* 1996.

[27] T. Vetter and T. Poggio, "Linear Object Classes and Image Synthesis from a Single Example Image," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 733-742, July 1997.

Amnon Shashua received the BSc degree in mathematics and computer science from Tel-Aviv University, Tel-Aviv, Israel, and the MSc degree in mathematics and computer science from the Weizmann Institute of Science, Rehovot, Israel, in 1986 and 1989, respectively. He received the PhD degree in computational neuroscience, working at the Artificial Intelligence Laboratory, from the Massachusetts Institute of Technology (MIT), in 1993. He is an associate professor at the School of Computer Science and Engineering, Hebrew University of Jerusalem, Israel. His research interests are in computer vision and computational modeling of human vision. His work includes early visual processing of saliency and grouping mechanisms, visual recognition, image synthesis for animation and graphics, and theory of computer vision in the areas of three-dimensional processing from a collection of two-dimensional views. He is a member of the IEEE and an associate editor of *IEEE Transactions on Pattern Analysis Machine Intelligence*.

Tammy Riklin-Raviv received the BSc degree in physics from the Hebrew University of Jerusalem, Israel, in 1993 and the MSc degree in computer science from the Hebrew University of Jerusalem, Israel, in 1999.