The Quotient Image: Class Based Recognition and Synthesis Under Varying Illumination Conditions

Tammy Riklin-Raviv and Amnon Shashua
Institute of Computer Science,
The Hebrew University,
Jerusalem 91904, Israel
http://www.cs.huji.ac.il/$\sim$ $\{\mathrm{tammy}, \mathrm{shashua}\}$/

Abstract

The paper addresses the problem of "class-based" recognition and image-synthesis with varying illumination. The class-based synthesis and recognition tasks are 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, capture the equivalence relationship (by generation of new images or by invariants) among all images of the object corresponding to new illumination conditions.

The key result in our approach is based on a definition of an illumination invariant signature image, we call the "quotient" image, which enables an analytic generation of the image space with varying illumination from a single input image and a very small sample of other objects of the class — in our experiments as few as two objects. In many cases the recognition results outperform by far conventional methods and the image-synthesis is of remarkable quality considering the size of the database of example images and the mild pre-process required for making the algorithm work.

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 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 [10, 11, 6, 2, 8]. 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.

Visual recognition and image 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 data base of (neutral) images.

In this paper we focus on recognition and image synthesis 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 synthesis 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 synthesize 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. 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. 2). The re-rendering results are in 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 [10, 11] in the case of Lambertian objects. Applications and related systems were reported in [6, 2, 5]. Re-rendering under more general assumptions, yet exploiting linearity of light transport was reported in [8].

Work on "class-based" synthesis and recognition of images (mostly with varying viewing positions) was reported in (partial list) [3, 1, 4, 15, 14]. These methods adopt a "reconstructionist" approach 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 [16] works under the assumption that 3D shapes of objects in a class are closed under linear combinations (in 3D). Recently, [9] 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 (grey-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 [10, 11]. 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, non-white light sources, and non-Lambertian surfaces [11, 8, 5], 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 [11, 5, 2]. See Fig. 2 for an example of using this principle for image synthesis.

We define next 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, Poggio and colleagues [14] 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 pre-aligned) data base. 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 alignment of 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. We define next the "recognition" and "synthesis" (re-rendering) problems.

Definition 2 (Recognition Problem) *Given $N \times 3$ images of N objects under 3 lighting conditions and $M \gg N$ images of 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, 3 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 3 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.*

To summarize up to this point, given the ideal class and the synthesis/recognition problem definitions above, our goal is: *we wish to extend the linear subspace result of [11] 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).

3 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 (3 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,*

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

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. $\square$

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, $i = 1, ..., N$, be a matrix whose columns are the three pictures of object $\mathbf{a_i}$ with albedo function ρ_i. Thus, $A_1, ..., A_N$ 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 $A_1, ..., A_N$ 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 case where there exist coefficients $\alpha_1, ..., \alpha_N$ such that

$$\frac{\rho(u,v)}{\rho_y(u,v)} = \alpha_1 \rho_1(u,v) + ... + \alpha_N \rho_N(u,v) \qquad (1)$$

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*

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

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 $\alpha_1, ..., \alpha_N$ that satisfy eqn. 1.

The proof can be found in [12]. Finding $\min f(\hat{x})$ is a simple technicality (a linear least-squares problem):

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

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

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

$$\alpha_i y_s^\top y_s - (\sum_{r=1}^{N} \alpha_r v_r)^\top A_i^\top y_s = 0 \qquad (3)$$

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

The proof can be easily derived and is shown in [12]. 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. In practice, the albedo functions live in a relatively low dimensional subspace of m, therefore a relatively small size $N << m$ is required, and that is tested empirically in Section 4.

Once the coefficients $x = (x_1, x_2, x_3)$ 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, $A_1, ..., A_N$, 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:

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

where $i = 1, ..., N$.

3. Solve the homogeneous system of linear equations in $\alpha_1, ..., \alpha_N$ 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 $A_1, ..., A_N$. Replace divisions by zero by small numbers.

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.

Finally, it is worthwhile noting that it is possible to synthesize color images using a black-and-white bootstrap set by converting the RGB channels of the novel color image into HSV representation. The V channel is the input to the algorithm above and is used to synthesize V'. The corresponding new color image is represented by the original H,S channels and the new V'.

4 Experiments

We have conducted a wide range of experimentation on the algorithm presented above. We will show here two experiments, the rest can be found in [12]. We first used a high quality database prepared by Thomas Vetter and his associates [14, 15]. We have chosen a bootstrap collection of 10 objects shown in Fig. 1. 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).

In Fig 2 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. 1.

Fig. 3 shows 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.

5 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 data base contains 200 faces each under 9 lighting conditions, making a total of 1800 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 1800 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. 4). In the first test the Q-database was generated from images under the same illumination (we have 9 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%). The match against the original images, instead of the Q-images, had 142 mismatches (7.8%). In the second test the images used for creating the Q-database were drawn randomly from the set of 9 images (per object). The match against the Q-database produced only 6 mismatches (0.33%), whereas the match against the original images produced 565 mismatches (31.39%). 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 distri-

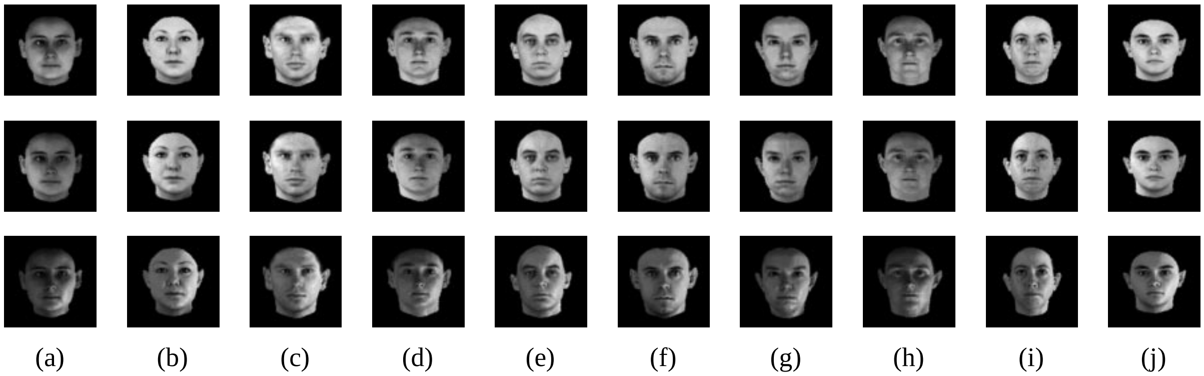

Figure 1. The bootstrap set of 10 objects from Vetter's database of 200 objects.

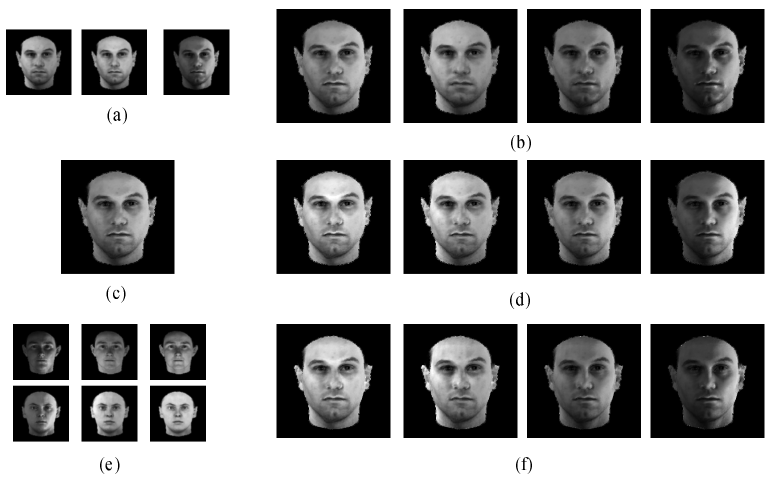

Figure 2. Image synthesis examples. (a) Original images under 3 distinct lighting conditions and the synthesized images (b) using linear combinations of those 3 images. The synthesized images using the original single image (c) and a $N = 10$ bootstrap set are shown in (d). Finally, (e) is an $N = 2$ bootstrap set for generating the synthesized images (f) from the single original image (c).

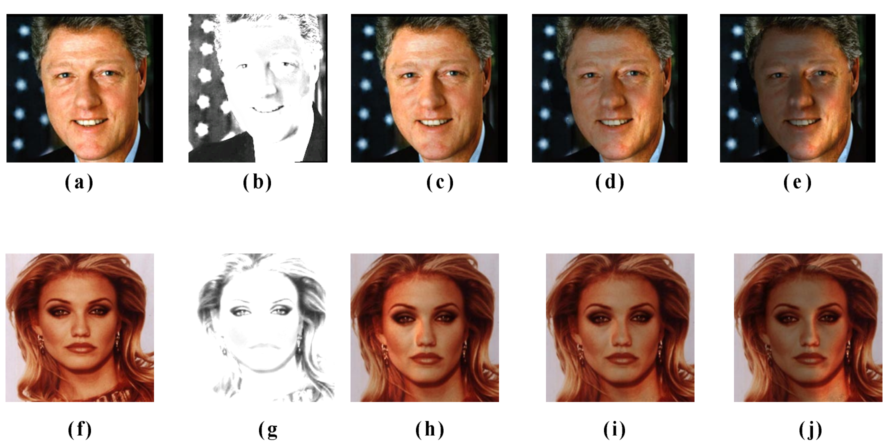

Figure 3. Color examples. Original images (a),(f) and the bootstrap set of Fig. 1 are from completely different sources. The re-rendered images are in (c)–(e) and (h)–(j), respectively.

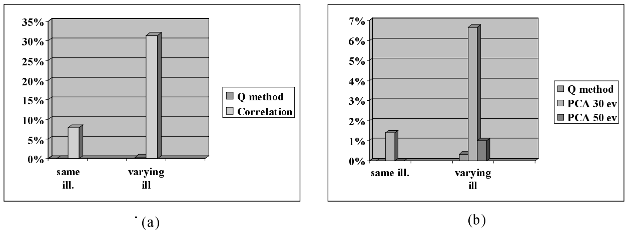

Figure 4. Recognition results on Vetter's database of 1800 face images. We compare the Q-image method with correlation and Eigenfaces. See text for details.

bution of the image.

We also made a comparison against the "eigenfaces" approach [13, 7] 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 9 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%), and the second test 120 mismatches (6.6%). The performance peaks around 50 principle components in which case the first test was error free (like in the Q-image method), and the second test had 18 mismatches (1%).

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.

6 Summary

We have presented a class-based synthesis 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 "toy" 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.

[1] R. Basri. Recognition by prototypes. *International Journal of Computer Vision*, 19(2):147–168, 1996.

[2] P.N Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. In *Proceedings of the European Conference on Computer Vision*, 1996.

[3] D. Beymer and Poggio T. Image representations for visual learning. *Science*, 272:1905–1909, 1995.

[4] W.T. Freeman and J.B. Tenenbaum. Learning bilinear models for two-factor problems in vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 554–560, 1997.

[5] A.S. Georghiades, D.J. Kriegman, and P.N. Belhumeur. Illumination cones for recognition under variable lighting:faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 52–59, 1998.

[6] P. Hallinan. A low-dimentional representation of human faces for arbitrary lightening conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 995–999, 1994.

[7] M.Turk and A.Pentland. Eigen faces for recognition. *J. of Cognitive Neuroscience*, 3(1), 1991.

[8] J. Nimeroff, E. Simoncelli, and J. Dorsey. Efficient rerendering of naturally illuminated environments. In *Proceedings of the Fifth Annual Eurographics Symposium on Rendering*, Darmstadt Germany, June 1994.

[9] E. Sali and S. Ullman. Recognizing novel 3-d objects under new illumination and viewing position using a small number of examples. In *Proceedings of the International Conference on Computer Vision*, pages 153–161, 1998.

[10] A. Shashua. Illumination and view position in 3D visual recognition. In *Proceedings of the conference on Neural Information Processing Systems (NIPS)*, Denver, CO, December 1991.

[11] A. Shashua. On photometric issues in 3D visual recognition from a single 2D image. *International Journal of Computer Vision*, 21:99–122, 1997.

[12] A. Shashua and T. Riklin-Raviv. The quotient image: Class based re-rendering and recognition with varying illuminations. Computer Science TR99-1, Hebrew University, 1999.

[13] L. Sirovich and M. Kirby. Low dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America*, 4(3):519–524, 1987.

[14] T. Vetter, M.J. Jones, and T. Poggio. A bootstrapping algorithm for learning linear models of object classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 40–46, 1997.

[15] T. Vetter and T. Poggio. Image synthesis from a single example view. In *Proceedings of the European Conference on Computer Vision*, 1996.

[16] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):733–742, 1997.