Learning Object Representations from Lighting Variations

R. Epstein[1], A. L. Yuille[2], and P. N. Belhumeur[3]

[1] Division of Applied Sciences, Harvard University, Cambridge MA, 02138
[2] Smith-Kettlewell Eye Research Institute, San Francisco, CA 94115.
[3] Department of Electrical Engineering, Yale University, New Haven, CT 06520-8267

Abstract. Realistic representation of objects requires models which can synthesize the image of an object under all possible viewing conditions. We propose to learn these models from examples. Methods for learning surface geometry and albedo from one or more images under fixed posed and varying lighting conditions are described. Singular value decomposition (SVD) is used to determine shape, albedo, and lighting conditions up to an unknown 3×3 matrix, which is sufficient for recognition. The use of class-specific knowledge and the integrability constraint to determine this matrix is explored. We show that when the integrability constraint is applied to objects with varying albedo it leads to an ambiguity in depth estimation similar to the bas relief ambiguity. The integrability constraint, however, is useful for resolving ambiguities which arise in current photometric theories.

Object Recognition Workshop. ECCV. 1996.

1 Introduction

The image of an object depends on many imaging factors such as lighting conditions, viewpoint, articulation and geometric deformations of the object, albedo of the object, and whether it is partially occluded by other objects. It is therefore necessary to design object represention which capture all the image variations caused by these factors. Such representations can then be used for object detection and recognition.

We believe that realistic represention of objects will require models which can *synthesize* the image of an object for all possible values of the imaging factors. Such an approach has long been advocated by people influenced by Bayesian probability theory [8]. This approach has some similarities to "appearance based models" [23] but, as we will argue in the next section, there are some important differences.[1]

We propose to learn these representations from examples. Learning from examples allows us the possibility of representing objects which are too complicated

[1] Alternative approaches, such as extracting invariant features, [22] may only be applicable to limited classes of objects such as industrial parts.

for current modelling systems. If statistical techniques are used, this allows us to concentrate on the important characteristics of the data and ignore unimportant details. For example, aspect graphs [18] give an elegant way of characterizing the different views of objects. But for many objects, they are difficult to calculate and hard to use. By contrast, the statistical methods used in [23] are able to recognize certain objects from different viewpoints using simpler techniques. It seems therefore, that some of the complexities of the aspect graph representation are unnecessary, at least for some classes of objects.

We argue that it is important to model the variations of all the factors affecting the image *independently* and *explicitly*. This will allow the object representations to be more general, suitable for more complicated objects, and more easy to generalize to new instances. For example, the appearance based matching algorithm of Murase and Nayar [23] is highly successful within its choosen domain of simple rigid objects but its avoidance of geometric and reflectance models means that it could be fooled by simply repainting one of the learned objects. The new (repainted) object would then have to be learnt again, requiring a costly training procedure. Similar problems would arise if the object is allowed to deform geometrically.

Modelling variations explicitly also makes it easy to incorporate prior knowledge about the object class into the learning procedure. If the object class is known, and explicit models are used, then far less training data will be needed. It appears that humans can make use of this type of class specific knowledge in order to generalize rapidly from one instance of an object [21]. In related work, we are exploring whether our models can account for these and other psychophysical experiments.

In this paper, therefore, we will describe methods for learning the geometry and reflectance functions of objects from one of more images of the object. We assume fixed pose but vary the lighting conditions[2]. For this paper we assume Lambertian reflectance functions with non-constant albedo, but we are currently generalizing our work to other types of relectance models.

We describe mechanisms for learning the shape, reflectance, and albedo of an object with or without the use of class specific knowledge. In particular, we make use of the surface integrability constraint and discover a close relation between the bas relief ambiguity and integrability. We illustrate the usefulness of our represention by synthesizing images. In related work, Belhumeur and Kriegman [2] characterize the set of images that can be generated by using Lambertian models, of the type we learn here, and give further examples of image synthesis.

Our approach makes use of singular value decomposition (SVD) which has previous been applied to the related problem of photometric stereo by Hayakawa [13]. For Lambertian sources with a single illuminant, SVD allows one to estimate shape, albedo, and lighting conditions up to an unknown 3×3 constant matrix, which we call the $\mathbf{A}$ matrix. We observe that the stated assumptions in [13] only determines $\mathbf{A}$ up to an unknown rotation matrix. It can be shown [30] this

[2] An extension to variable pose is described in Epstein and Yuille (in preparation).

assumption is valid for certain types of surfaces but will be incorrect for others. However, we demonstrate that a variety of general purpose and/or class specific assumptions, including surface integrability, can be used to determine the **A** matrix uniquely. Moreover, it can be shown [2] that the set of allowable images of the object (from fixed viewpoint) can be determined *without* knowing **A**.

2 Appearance Based Models and Image Synthesis

To set our work in context, it is important to describe how it relates to other work on image synthesis and the influential work on appearance based models [23].

Appearance based models (ABM's) of objects are learned by applying principal component analysis (PCA) to a representative dataset of images of an object. For certain classes of objects, this produces a low-dimensional subspace which captures most of the variance of the dataset. The object can then be represented by a manifold defined in this low-dimensional space. The position of the image on this manifold will depend on the lighting and viewpoint conditions. An input image, or subpart of an image, can be matched to the appearence manifold and hence recognized. This approach is extremely successful within specific domains.

It is interesting to contrast ABM's with image synthesis models of the type that we use in this paper. Our approach requires specifying a representation for the object and an imaging model. The representation model should be flexible enough to deal with all the variations described previously – due to lighting, articulation, geometric deformations, etc. The imaging model enables us to synthesize an image of the object. The representation and imaging models are learnt by statistical techniques from samples of the data.

Synthesis models and ABM's are similar in two important respects. Firstly, unlike many (most) current object recognition systems, they do not first extract sparse features, such as edges, from the image (see [9]). However, the word "appearance" in ABM's is slightly misleading because the ABM's only model the appearance of the object within the low dimensional subspace. They ignore all image variations that project outside this subspace. The synthesis models, by contrast, generate all possible image variations. Secondly, both synthesis models and ABM's are statistical with their models being generated by the data. This makes them more robust with respect to noise which can destroy more deterministic modelling approaches such as geometric invariants [22].

From our viewpoint, however, the ABM's are limited because they do not represent variables like shape and lighting explicitly. It is straightforward to adapt synthesis models to take into account geometrical deformations or to add paint onto the surface of an object. But an ABM would have to learn all such changes from scratch. Similar problems would also apply in the related eigenface approach [27] where the eigenfaces combine albedo, lighting, and geometrical changes, but represent none of them explicity. Like ABM's this approach involves projecting the image onto a low-dimensional space and ignoring anything that lies outside this space.

The ABM's have gone a long way in demonstrating the advantages of using much richer descriptions than simply sparse features like edges and corners for recognition. Still, a drawback of these approaches is that in order to recognize an object seen from a particular pose and under a particular illumination, they must have previously seen the object under the same conditions. Yet, if one tries to enumerate all possible poses and permutes these with all possible illumination conditions, things get out of hand quite quickly. Fortunately, this brute-force approach to modeling, which requires observing objects under the full range of parametric variation, is unnecessary since appearance can usually be predicted from a modest number of images.

Indeed, both eigenfaces and ABM's can be considered to be feature based methods where the features are extracted by applying linear filters determined by PCA. It can be argued [3] that if the goal is discrimination between objects, rather than representation, then better linear filters can be used based on Fisher's linear discriminant. PCA projects into the subspace which captures most of the variance between objects. By contrast, Fisher's linear discriminant [7] projects into the subspace which maximizes the variation between different objects. This can be illustrated by considering applying both techniques to a set of faces in which a small subclass of people have glasses. The PCA approach would tend to project onto a subspace which ignores the glasses (because they appear in two few samples to significantly affect the variance). By contrast, Fisher's linear discriminant would project into a subspace which included the glasses because they would be powerful cues for distinguishing between people.

A more explicit way of modelling faces occurs in [4] where the eigenfaces are considered to be principal components of the albedoes of faces. Two-dimensional geometrical distortions are applied to allow for changes in viewpoint and expression. These deformations occur by warping a set of feature points, corresponding to the facial features, and interpolating the warp over the rest of the face.

Lighting variations are also handled explicity by a related model by Hallinan [12] which is able to recognize faces under highly variable lighting conditions and to distinguish reliably between faces and non-faces. Lighting variations are represented by a linear combination of lighting basis images obtained from PCA. To model geometric changes, Hallinan [12] uses two-dimensional image warps. Though this not an explicit model of surface geometry, it can be shown that the spatial warps correspond to warps of the surface normal vectors of the underlying three dimensional shape [29]. It is therefore straightforward to recompute the surfaces from the warps. Hallinan's lighting models were the starting point for this current work and we will return to them later in the paper.

Another model, that uses image synthesis and explicit representations is the face recognition system reported in [1]. This face model uses three dimensional geometry and a Lambertian imaging model. By using a dataface of face geometry, obtained by laser scanning, a strong prior distribution for the shape of faces is obtained. Using this prior the three dimensional geometry of the face can be estimated from a single image. However, the types of geometric models used in this system are somewhat limited and only apply to objects made of single parts,

such as faces. For objects with several articulating parts more sophisticated geometrical models should be used, perhaps of the type described in [31].

3 The Lambertian Model and Lighting Basis Functions

Suppose we pick an object and fix its pose and articulation. Then the principle of superposition ensures that the set of images of the object, as the lighting varies, lies within a linear space[3]. How does this obervation relate to reflectance function models of image formation?

The most used reflectance model is the Lambertian model [14] which is often written as:

$$I(x,y) = a(x,y)\mathbf{n}(x,y)\cdot\mathbf{s} \equiv \mathbf{b}(x,y)\cdot\mathbf{s}, \tag{1}$$

where $a(x,y)$ is the albedo of the object, $\mathbf{n}(x,y)$ is its surface normal, $\mathbf{b}(x,y) \equiv a(x,y)\mathbf{n}(x,y)$ and $\mathbf{s}$ is the light source direction (the light is assumed to be at infinity). If this equation applies then it is clear [26],[28],[25] [20], that the space of images of the object, as the light source direction changes, spans a three dimensional subspace. In other words, any image of the object can be expressed as:

$$I(x,y) = \sum_{i=1}^{3} \alpha_i b_i(x,y), \tag{2}$$

for some coefficients $\{\alpha_i\}$, where i labels the vector components. This is a linear subspace model of image formation.

Equation (1), however, has several limitations. It ignores attached shadows (where $\mathbf{b}(x,y)\cdot\mathbf{s} \leq 0$), cast shadows, and partial or hidden shadows (where there are several light sources and the light from some of them are shadowed). It also ignores intereflections. When these effects are taken into account, the dimensionality of the image space rises enormously [2]. Moreover, the model ignores specularities and will break down if the light source is close to the object. These limitations mean that caution is necessary when using this model.

Alternatively, motivated by the principle of superposition, one can try to analyze the empirical structure of the set of possible images . In a series of empirical studies [11], [5] principal component analysis (PCA) was used to analyze the space of images generated by one object at fixed pose with varying lighting conditions. The lighting conditions were sampled evenly on the view hemisphere, so the dataset included extreme lighting configurations. The experimental results showed that 5 ± 2 eigenvalues were typically enough to account for most of the variance. For faces, the percentage of variance covered by the first five eignevalues was approximately 90 %. For objects which were highly specular (such as a helmet) or with many shadows (such as an artificial parrot) the percentage decreased. Nevertheless, the specularities and shadows, though perceptually very saliant, contributed little to the variance. In addition, Hallinan [11] showed that

[3] In fact it can shown to lie within a convex cone inside this linear space [2]

if *different* faces were aligned geometrically, using affine transformations, then the first five eigenvalues still captured approximately 90 % of the variance.

These results meant that for each object we could approximate the image space by a linear combination of the first five eignevectors or *lighting basis functions*. In other words an image of the object, under fixed viewpoint, could be expressed as:

$$I_M(\boldsymbol{x}; \{\alpha_i\}) = \sum_{i=1}^{5} \alpha_i B_i(\boldsymbol{x}), \tag{3}$$

where the $\{B_i(.)\}$ are the lighting basis functions (i.e. the first five principal components), and the $\{\alpha_i\}$ are the coefficients (which depend on the specific lighting conditions).

If this number of coefficients is set equal to three then this would be similar to the Lambertian linear model, see equation (2). Indeed it was observed that the first three lighting basis functions usually corresponded to the image lit from in front, from the side, and from above. This is explained in [30].

The empirical linear subspace model, see equation (3), was used by Hallinan [12] to successfully model lighting variation. Such models are attractive but they do have several limitations: albedo and shading information is combined indiscriminantly and there is no explicit 3-D model. (Although, under certain circumstances [29] it does allow recovery of the three-dimensional shape.)

For reasons described above, we would prefer a more explicit representation based on three-dimensional shape and albedo. We argue, therefore, that the success of the linear subspace results suggest that Lambertian models are a good approximation to a number of real objects. Indeed, it was *conjectured* [5] that the first three principal components of this space correspond to Lambertian illumination of the object and higher order principal components dealt with specularities and sharp shadows.

4 Learning the Models

Our approach consists of learning models of the objects – their surface geometry and albedo – using variants of the Lambertian model which make it robust to shadows and specularities. This is done with four different schemes.

Suppose we have a set of images of an object illuminated by M different point light sources. We denote these light sources by $\{\mathbf{s}(\mu) : \mu = 1, ..., M\}$. The resulting images are represented by $\{I(p, \mu) : \mu = 1, ..., M \quad p = 1, ..., P\}$ where the index p labels the pixels of the image (these pixels lie on a two dimensional grid but it is convenient to represent them as a vector).

Our first scheme assumes that we have multiple images of the object[4] and the light sources are known. This is of least interest since it is a strong assumption and corresponds to standard photometric stereo [26, 28, 14, 17], though with nonconstant albedo. We investigated this scheme mainly to test the Lambertian

[4] Fixed pose and varying illumination.

asumptions about our data. We concluded that the model is a good approximation though robust techniques are needed to reduce the influence of shadows and specularities.

If, however, there are multiple unknown light sources then we show that SVD can be applied (see also [13]) to simultaneously estimate the surface geometry and albedo up to a 3×3 linear transformation, the $\mathbf{A}$ matrix. This transformation arises due to an ambiguity in the Lambertian equation (1). This is because for any arbitrary invertible linear transformation $\mathbf{A}$:

$$\mathbf{b} \cdot \mathbf{s} = \mathbf{b}^T\mathbf{s} = \mathbf{b}^T\mathbf{A}\mathbf{A}^{-1}\mathbf{s}. \tag{4}$$

Our second learning scheme, follows from this result and the proof in [2] that the set of images of the object are *independent* of the precise value of $\mathbf{A}$ provided the objects are viewed from front on. This means that it unecessary to estimate $\mathbf{A}$. Our second scheme, therefore consists merely of applying SVD to the input data and thereby generating the light cone representation described in [2].

For our third learning scheme, we demonstrate that the $\mathbf{A}$ matrix can be recovered by using the surface integrability constraint and the assumption that we either have an image of the object under ambient lighting, or that the sampling set of lighting conditions allows us to generate one. We compare our assumptions to those of [13] and prove that his method relies on an, unstated, assumption about the dataset which will often not be valid. In addition, we describe a new perceptual ambiguity related to the integrability constraint. This scheme results in the full albedo and three-dimensional shape of the object.

In our fourth learning scheme, we consider the use of prior knowledge about the class of the viewed object. We demonstrate that $\mathbf{A}$ can be learnt by merely assuming that we know a prototype object of that class. Not suprising, if the object class is known then fewer images are needed to learn the object model. This seems to agree with current psychophysical results [21].

4.1 Learning the Models with known light source direction

Suppose we assume that the light source vectors $\{\mathbf{s}(\mu) : \mu = 1, ..., M\}$ are known. This is true for our dataset because the images have been gathered under controlled conditions.

We can formulate estimating shape and albedo as a least squares optimization problem:

$$E[b; V] = \sum_{\mu,p} V(p,\mu)\{I(p,\mu) - \sum_i b_i(p)s_i(\mu)\}^2 \tag{5}$$

where $V(p,\mu)$ is a binary indicator function whose value is 1 if point p is not in shadow, or have a specularity, under lighting condition μ, and is zero otherwise.

The arguments of the energy function – b, s, V – represent the sets $\{\mathbf{b}(p) : p = 1, ..., P\}$; $\{\mathbf{s}(\mu) : \mu = 1, ..., M\}$, and $\{V(p,\mu) : p = 1, ..., P \quad \mu = 1, ..., M\}$ respectively.

We observe that the energy can be written as the sum of P independent energies $E_p[\mathbf{b}(p), \{V(p,\mu) : \mu = 1, ..., M\}] = \sum_\mu V(p,\mu)\{I(p,\mu) - \sum_i b_i(p)s_i(\mu)\}^2$.

These energy functions E_p $(p = 1, ..., P)$ are all quadratic in b and so they can be minimized by linear algebra provided the V are specified. This allows us to estimate the surface normal and albedo at all points p independently.

We first assume that there are no specularities or shadows, in other words we set $V(p, \mu) = 1, \ \forall \ p, \ \mu$. This gives the results shown in figures (1, 2). This is equivalent to the photometric stereo techniques described in [26], [28], [14].

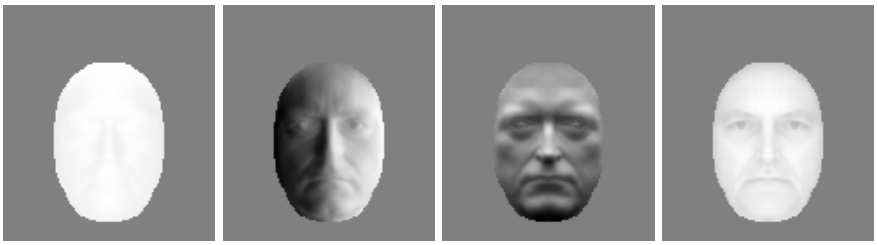

Fig. 1. The albedo and normals estimated directly assuming known light source directions and *without* using robust techniques to remove specularities and shadows. The first three images are the z, x, and y components of the surface normal respectively. The rightmost image is the albedo. Observe that the estimated albedo appears to get darker near the boundaries of the face causing the albedo image to appear to be non-flat. This is due to failure to treat the shadows correctly.

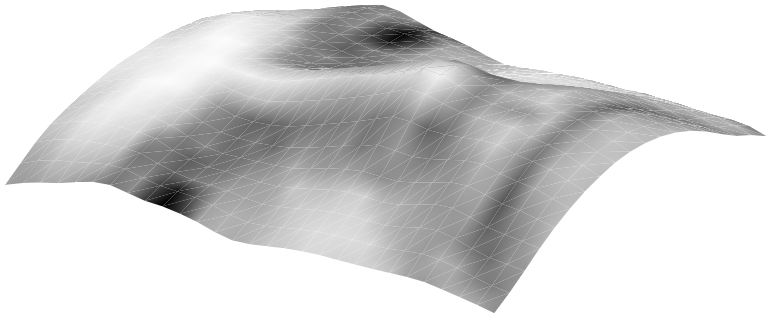

Fig. 2. The surface computed from the normals in the previous figure. The face appears flattened. This is because the algorithm's failure to remove shadows means that it underestimates the albedo in shaded regions and correspondingly makes the surface flatter.

These results are reasonable but close inspection shows that the estimated albedo becomes darker towards the boundaries of the face, see figure (1), and the shape of the face is flattened, see figure (2). This is because the algorithm knows nothing about shadows and tries to model them as regions of dark albedo. This in

turn causes the shape to appear too frontoparallel. We conclude that the object is approximately Lambertian but that it also has shadows and specularities.

We observe, however, that specularities are bright, shadows are dark, and a point will tend to be in shadow or specular only for a limited set of lighting directions. Thus if we histogram the intensity values at a single image point, as it is illuminated from many directions, the brightest and darkest points will tend to be specularities and shadows[5].

Thus we can remove most of the effects of shadows and specularities by plotting the histogram, see figure (3), and set $V = 0$ for the bottom $\alpha_1\%$ and top α_2 %. If α_1 and α_2 are sufficiently large (say 30%) then we set $V = 0$ for the remaining data (which we now assume is purely Lambertian).

We now minimize the E_p again using linear algebra. The results are significantly improved, see figures (4, 5). Observe that the albedo image in figure (4) appears to be much flatter, suggesting that we have removed much of the effects of the shadows. This is further supported by the surface plot, see figure (5), which is no longer foreshortened – compare with figure (2). Thus eliminating the shadows by pruning the histogram gives us significantly more uniform albedoes on the skin and a more accurately estimated shape.

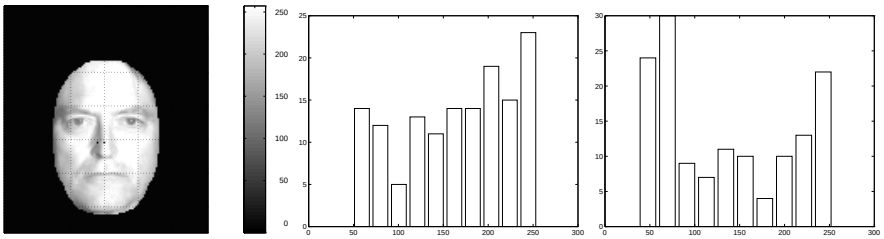

Fig. 3. Histograms for two pixels on the bridge and the side of the nose, locations shown in the left image by the two black dots. The middle image shows the histogram for the pixel on the bridge of the nose. This pixel was never in shadow so there is no peak in the histogram for small intensity values (corresponding to shadows). The right image shows the histogram for the pixel on the side of the nose. This pixel was often in shadow and so its histogram has a peak at low intensity values. Note that background ambient illumination prevents the shadows from being perfectly dark.

Alternatively, instead of eliminating the top $\alpha_2\%$ and the bottom $\alpha_1\%$ of $\{I(p,\mu) : \mu = 1, ..., M\}$ we could instead eliminate all intensities below a *shadow threshold* and above a *specularity threshold*. Or, we could do residual analysis to check whether the intensities thrown away correspond to true shadows or specularities. We can use our estimate $\mathbf{b}^*(p)$ to *predict* what the intensities would be for those cases. Those light source configurations for which the predictions

[5] Ideally perfect shadows would have zero intensity, but our light sources are not true point sources and there was some ambient light present when our database was collected.

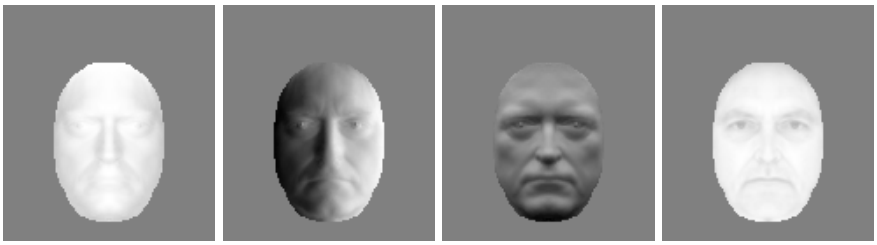

Fig. 4. The normals and albedo calculated directly. Residuals at low intensity values < 70 have been removed.

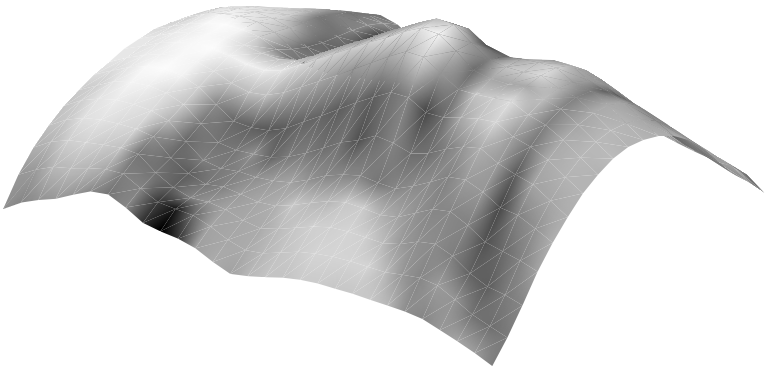

Fig. 5. Surface computed from normals above.

Fig. 6. This figure shows the extent to which each pixel was thresholded. Pixel brightness corresponds to the number of images in which the pixel was over threshold. Hence, dark pixels were thresholded out (considered in shadow) more. Observe that points will low albedos, such as the irises of the eyes, are overrepresented.

agree with the observed intensities are no longer assumed to be due to shadows, or specularities, and so are used to make a second estimate of $\mathbf{b}(p)$. This process can be repeated.

4.2 Light Source Direction Unknown: Using SVD to estimate surface properties and light source directions up to a linear transformation.

It is unrealistic to assume that the light source directions will be given. Thus we need a method which can estimate them and the surface properties simultaneously. In other words, we need to minimize the energy function $E[b,s] = \sum_{\mu,p}\{I(p,\mu) - \sum_{i=1}^{3} b_i(p)s_i(\mu)\}^2$ as a function of b *and* s. Fortunately minimization of this function, up a linear transform, can be done using singular value decomposition (SVD). This has been first applied to photometric stereo in [13].

Observe that the intensities $\{I(\mu,p)\}$ can be expressed as a $M \times P$ matrix $\mathbf{J}$. Similarly we can express the surface properties $\{b_i(p)\}$ as a $P \times 3$ matrix $\mathbf{B}$ and the light sources $\{s_i(\mu)\}$ as a $3 \times M$ matrix $\mathbf{S}$. SVD implies that we can write $\mathbf{J}$ as:

$$\mathbf{J} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{6}$$

where $\mathbf{\Sigma}$ is a diagonal matrix whose elements are the square roots of the eigenvalues of $\mathbf{J}\mathbf{J}^T$ (or equivalently of $\mathbf{J}^T\mathbf{J}$). The columns of $\mathbf{U}$ correspond to the normalized eigenvectors of the matrix $\mathbf{J}^T\mathbf{J}$. The ordering of these columns corresponds to the ordering of the eigenvalues in $\mathit{\Sigma}$. Similarly, the columns of $\mathbf{V}$ correspond to the eigenvectors of $\mathbf{J}\mathbf{J}^T$.

If our image formation model is correct then there will only be three nonzero eigenvalues of $\mathbf{J}\mathbf{J}^T$ and so $\mathbf{\Sigma}$ will have only three nonzero elements. We do not expect this to be true for our dataset because of shadows, specularities, and noise. But SVD is guaranteed to gives us the best least squares solution in any case. Thus the biggest three eigenvalues of $\mathbf{\Sigma}$, and the corresponding columns of $\mathbf{U}$ and $\mathbf{V}$ represent the Lambertian part of the reflectance function of these objects. We define the vectors $\{\mathbf{f}(\mu) : \mu = 1, ..., M\}$ to be the first three columns of $\mathbf{U}$ and the $\{\mathbf{e}(p) : p = 1, ..., P\}$ to be the first three columns of $\mathbf{V}$.

This assumption enables us to use SVD to solve for $\mathbf{B}$ and $\mathbf{S}$ up to a linear transformation. The solution is:

$$\begin{aligned} \mathbf{s}(\mu) &= \mathbf{P}\mathbf{f}(\mu), \ \forall\ \mu, \\ \mathbf{b}(p) &= \mathbf{Q}\mathbf{e}(p), \ \forall\ p, \end{aligned} \tag{7}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are 3×3 matrices which are constrained to satisfy $\mathbf{P}^T\mathbf{Q} = \mathbf{\Sigma_3}$, where $\mathbf{\Sigma_3}$ is the 3×3 diagonal matrix containing the square roots of the biggest three eigenvalues of $\mathbf{J}\mathbf{J}^T$. There is an ambiguity $\mathbf{P} \mapsto \mathbf{A}\mathbf{P}$, $\mathbf{Q} \mapsto \mathbf{A^{-1}}^{\mathbf{T}}Q$ where $\mathbf{A}$ is an arbitrary invertible matrix.

This means we can determine $\{\mathbf{s}\}$ and $\{\mathbf{b}\}$ up a linear transform. It can be shown [2] that this is sufficient to recognize objects from front-on under

arbitrary illumination. To verify that these linear subspaces are correct we use our knowledge of the light source directions to determine the $\mathbf{P}$ and $\mathbf{Q}$ matrices (i.e. we use least squares to solve $\mathbf{s}(\mu) = \mathbf{Pf}(\mu), \ \forall \ \mu$ for $\mathbf{P}$.) The resulting albedos and surface normals are shown in figure (7). The results are similar to those obtained by using knowledge of the light source directions directly. They appear slightly better than the results without residuals, figure (1), and slightly worse than the results with residuals, figure (4).

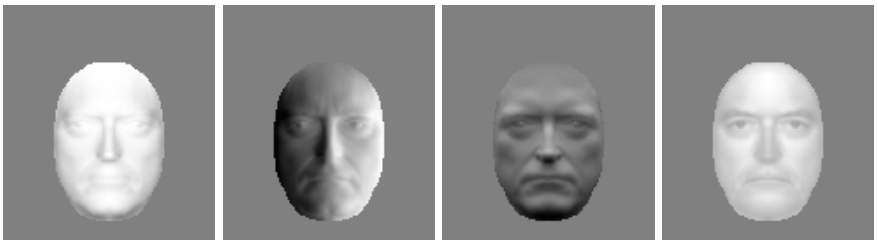

Fig. 7. Normals and albedo calculated directly from SVD using known light source directions to estimate the linear transformations.

4.3 Estimating the linear transformations.

We would like, however, to estimate the true geometry and albedo because this would enable us to predict how the object changes as the viewpoint varies (and to deal with cast shadows). The next subsection discusses ways to use additional information can be used to determine the linear transformation and hence to determine the surface albedo and shape.

Objects of Unknown Class Suppose we have an object of unknown class and we wish to determine the $\mathbf{A}$ matrix.

One plausible assumption is that we have an estimate of the object's albedo. This might consist of an additional image of the object taken under ambient lighting conditions[6]. Alternatively we can assume that the light source directions sample the view hemisphere and so, by taking the mean of our dataset we get an approximation to an ambient image of the object. It should be emphasized that this estimated albedo need only be very approximate.

We use the mean of our dataset to estimate the albedo. This means that, using Equation (7), for each point p in the image we have a constraint on the linear transformations:

$$a(p)^2 = \mathbf{e}^T(p)\mathbf{P}^T\mathbf{Pe}(p), \ \forall \ p = 1, ..., P. \tag{8}$$

[6] Recall that the image of an object under ambient lighting conditions is given by the albedo [14]

We impose these constraints using a least squares goodness of fit criterion. This can be solved using SVD to estimate $\mathbf{P}^T\mathbf{P}$. This yields $\mathbf{P}^T\mathbf{P} = \mathbf{WMW}^T$, where $\mathbf{M}$ is diagonal. We then estimate $\mathbf{P}^* = \mathbf{M}^{1/2}\mathbf{W}$ which is correct up to rotation.

We note that Hayakawa assumes that this rotation matrix is the identity [13]. It can be shown, however, that this is not always the case. Indeed, see [30], it can be shown to hold if the matrices $\sum_p b_i(p)b_j(p)$ and $\sum_\mu s_i(\mu)s_j(\mu)$ are both diagonal. But, for example, it does not hold if $\sum_\mu s_i(\mu)s_j(\mu)$ is diagonal but $\sum_p b_i(p)b_j(p)$ is not. The condition that these matrices are both diagonal can be traced to symmetry assumptions about the dataset. It is straightforward to generate situations for which they fail.

Fortunately, however, this rotation ambiguity can be cured by using the surface integrability constraint, see section 5. The results shown in figures (8,9) are consistent with integrability.

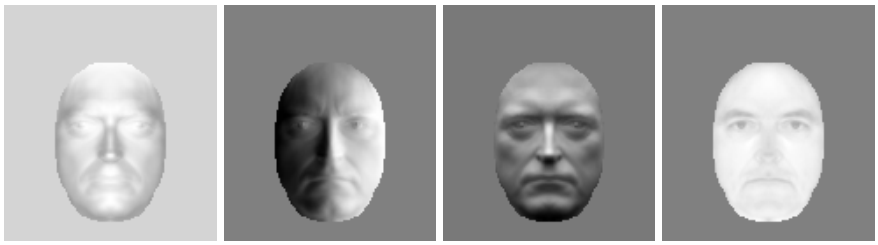

Fig. 8. Normals and albedo calculated from eigenvectors. We used the mean of the dataset as an initial estimate of albedo. The matrix P^TP is then calculated from $a^2(x) = e^T(x)P^TPe(x)\forall x$. SVD on P^TP gives $P^TP = W * M * W^T$, M diagonal. We then take, as an estimate of P, $P^* = sqrt(M) * U$. This is correct up to rotation. In the above results, we take the rotation matrix to be the identity and check consistency with integrability.

Objects of Known Class We can use knowledge about the class of the object to determine the linear transformations $\mathbf{P}$ and $\mathbf{Q}$, and hence determine the surface properties and the light sources uniquely.

To do this all we need is a $\mathbf{b}(p)$ vector from a prototype member of the class. For example, we assume that we know $\mathbf{b}_{Pr}(p)$ for a prototype face Pr. Then when we get the data for a new face image we will estimate its $\mathbf{P}$ and $\mathbf{Q}$ matrices by assuming that it has the same surface properties as the prototype. Thus we estimate $\mathbf{P}$ by minimizing:

$$\sum_p \left|\mathbf{b}_{Pr}(p) - \mathbf{Pe}(p)\right|^2, \tag{9}$$

where the $\mathbf{e}(p)$ are computed from the new dataset. We are minimizing a quadratic function of $\mathbf{P}$ so the result, $\mathbf{P}^*$, can be obtained by linear algebra.

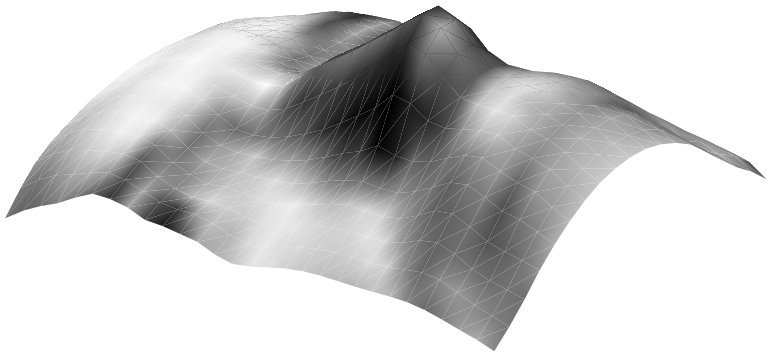

Fig. 9. Surface computed from normals above.

We now solve for the surface properties using:

$$\mathbf{b}(p) = \mathbf{P}^{*}\mathbf{e}(p), \quad \forall\, p. \tag{10}$$

Observe that the prototype is used merely in conjunction with the dataset to solve for the 3×3 matrix $\mathbf{P}$. Our results demonstrate that the surface properties computed using this assumption are good.

Fig. 10. Normals and albedos calculated for a new subject using the results shown in figure 7 as a prototype.

This result has used prior knowledge about object class in the simplest possible form – a prototype model. More sophisticated class knowledge, such as a prior probability distribution for shapes and albedoes, would lead to improved results.

5 Surface Integrability

The surface integrability constraint requires that the normal vectors are consistent with a surface (for a discussion, see [15].) It puts restrictions on the set of normals vectors but it is not sufficient to determine the surface uniquely. We

will show that for Lambertian objects with unknown albedo this leads to an ambiguity including scaling in depth.

The unit normals $\boldsymbol{n}(\boldsymbol{x}) = (n_1(\boldsymbol{x}), n_2(\boldsymbol{x}), n_3(\boldsymbol{x}))$ of a surface must obey the following surface integrability constraint to ensure that they form a consistent surface:

$$\frac{\partial}{\partial y}\left(\frac{n_1(\boldsymbol{x})}{n_3(\boldsymbol{x})}\right) = \frac{\partial}{\partial x}\left(\frac{n_2(\boldsymbol{x})}{n_3(\boldsymbol{x})}\right). \tag{11}$$

This constraint is a necessary and sufficient condition and can be derived from the fact that any surface can be locally parameterized as $z = f(x, y)$ with normals of form:

$$\boldsymbol{n}(\boldsymbol{x}) = \frac{1}{\{\boldsymbol{\nabla} f \cdot \boldsymbol{\nabla} f + 1\}^{(1/2)}}(f_x, f_y, -1). \tag{12}$$

It is straightforward to see that the vector $\boldsymbol{b}(\boldsymbol{x}) = a(\boldsymbol{x})\boldsymbol{n}(\boldsymbol{x})$ also satisfies the same constraint – i.e. we can replace (n_1/n_3) and (n_2/n_3) by (b_1/b_3) and (b_2/b_3) in the constraint equations.

Now recall that the linear algebra in the previous section determined the $\boldsymbol{b}(\boldsymbol{x})$ up to an unknown linear transformation determined by the $\mathbf{P}$ matrix.

The surface integrability constraint will partially determine the $\mathbf{P}$ matrix. It is straightforward to show, and to verify, that the only linear transformations which preserve the integrability constraint are:

$$\begin{aligned} b_1(\boldsymbol{x}) &\mapsto \lambda b_1(\boldsymbol{x}) + \mu b_3(\boldsymbol{x}), \\ b_2(\boldsymbol{x}) &\mapsto \lambda b_2(\boldsymbol{x}) + \nu b_3(\boldsymbol{x}), \\ b_3(\boldsymbol{x}) &\mapsto \rho b_3(\boldsymbol{x}). \end{aligned} \tag{13}$$

Observe that there is a constant scaling factor in this transformation which can never be determined (a dark surface lit with a bright light is indistinguishable from light surface lit by a dark light) so we could set $\rho = 1$ without loss of generality.

If the $\mathbf{A}$ matrix is known up to a rotation ambiguity, as in section 4.3, then integrability determines the remaining part of the transformation.

Moreover, if the albedo is known to be constant, then the class of transformations are reduced to the well known convex/concave (or light up/light down) ambiguity well known in the psychophysics literature. This is because the requirement that $\boldsymbol{b}(\boldsymbol{x})$ has constant magnitude (independent of $\boldsymbol{x}$) puts further restrictions on the transformation.

Thus for objects with unknown albedo, we get a class of perceptual ambiguities corresponding to the transformations given in equation (13). To understand these ambiguities we let the transformed surface be represented by $z = \bar{f}(x, y)$. It is straightforward calculus to see that:

$$\bar{f}(x, y) = \lambda f(x, y) + \mu x + \nu y. \tag{14}$$

In other words, the ambiguity consistent with the integrability constraint consists of scaling the depth by a factor λ and adding a planar surface $z =$

$\mu x + \nu y$. Interestingly, it has been reported [19] that humans appear to differ in their judgement of shape from shading by a scaling in the z direction. This connection is being explored in our current work.

Thus we see that the integrability constraints reduces the ambiguity in reconstructing the surface but it does not eliminate it altogether. To solve the problem uniquely we must impose additional constraints.

6 Learning an Object from a Single View

In previous sections we developed methods for learning object models assuming that we have multiple images of the object. In practice, however, we may only have one image of each object. Moreover, it is important to know how much we can learn about an object from a single image.

A single image, however, gives us little information about the object. Recall that, assuming Lambertian models, we can express the image as $I(\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x}) \cdot \boldsymbol{s}$ where $\boldsymbol{b}(\boldsymbol{x})$ and $\boldsymbol{s}$ are unknown. This equation, without additional assumptions, is not sufficient to determine $\boldsymbol{b}(\boldsymbol{x})$ and $\boldsymbol{s}$[7]. To make progress we must use knowledge about the class of the object. One way to do this would be to do statistics on the class of objects to develop a prior distribution for them[1]. Instead we will determine techniques for learning object models making as few assumptions as possible about the object class. Our assumptions are: (i) a prototype model, $\boldsymbol{b}^p(\boldsymbol{x})$, for the class, and (ii) symmetry assumptions about the object.

For faces the symmetry assumption is valid and we can select a prototype head from our database. It is convenient to use as a prototype one of our previously learnt models shown in figures (8,9). The algorithm proceeds in several stages.

Stage I. We use the prototype model to estimate the light source direction. More precisely, we solve for:

$$\boldsymbol{s}^* = \arg\min_{\boldsymbol{s}} \int d\boldsymbol{x} \left| I(\boldsymbol{x}) - \boldsymbol{b}^p(\boldsymbol{x}) \cdot \boldsymbol{s} \right|^2 . \tag{15}$$

Stage II. The symmetry assumption. We assume that the object is symmetric across the y-axis at $x = 0$. This means that we can express the model as:

$$\begin{aligned} \big(b_1(x,y), b_2(x,y), b_3(x,y)\big) &= \big(h_1(x,y), h_2(x,y), h_3(x,y)\big), \quad x \geq 0, \\ \big(b_1(x,y), b_2(x,y), b_3(x,y)\big) &= \big(-h_1(-x,y), h_2(-x,y), h_3(-x,y)\big), \quad x \leq 0, \end{aligned} \tag{16}$$

where $\big(h_1(x,y), h_2(x,y), h_3(x,y)\big)$ represents the right half of the face.

By using the image of the left and the right part of the face we can observe $s_1 h_1(x,y) + s_2 h_2(x,y) + s_3 h_3(x,y)$ and $-s_1 h_1(x,y) + s_2 h_2(x,y) + s_3 h_3(x,y)$. Therefore, using the fact that we know $\boldsymbol{s}$ from Stage I, we know $s_1 h_1(x,y)$ and $s_2 h_2(x,y) + s_3 h_3(x,y)$. Thus we know two components of $\boldsymbol{h}(x,y)$. It remains to

[7] Current shape from shading algorithms usually assume known light source and constant albedo.

determine the third component $-s_3h_2(x,y)+s_2h_3(x,y)$. Of course, this requires that neither $s_1 = 0$ nor $s_2 = s_3 = 0$. So the lighting cannot be purely front-on or purely from the x-direction.

Stage III. To determine the third component – $-s_3h_2(x,y)+s_2h_3(x,y)$ – we make use of the integrability constraint and, if necessary, the prior model. The integrability constraint is:

$$\frac{\partial}{\partial x}\frac{h_2(x,y)}{h_3(x,y)} = \frac{\partial}{\partial y}\frac{h_1(x,y)}{h_3(x,y)}, \quad \forall x, y. \tag{17}$$

Multiplying this equation by $h_3^2(x,y)$ and expanding it gives:

$$h_3(x,y)\frac{\partial}{\partial x}h_2(x,y) - h_2(x,y)\frac{\partial}{\partial x}h_3(x,y) = h_3(x,y)\frac{\partial}{\partial y}h_1(x,y) - h_1(x,y)\frac{\partial}{\partial y}h_3(x,y). \tag{18}$$

We define two new vectors $p_2(x,y)$ (known) and $p_3(x,y)$ (unknown) by:

$$p_2(x,y) = \frac{s_2h_2(x,y)+s_3h_3(x,y)}{(s_2^2+s_3^2)}, \quad p_3(x,y) = \frac{-s_3h_2(x,y)+s_2h_3(x,y)}{(s_2^2+s_3^2)},$$
$$h_2(x,y) = s_2p_2(x,y) - s_3p_3(x,y), \quad h_3(x,y) = s_3p_2(x,y) + s_2p_3(x,y). \tag{19}$$

Then we express integrability by defining a function $K(x,y)$:

$$K(x,y) = (s_2^2+s_3^2)p_3(x,y)\frac{\partial p_2(x,y)}{\partial x} - (s_2^2+s_3^2)p_2(x,y)\frac{\partial p_3(x,y)}{\partial x}$$
$$-\big(s_3p_2(x,y)+s_2p_3(x,y)\big)\frac{\partial h_1(x,y)}{\partial y} + h_1(x,y)\frac{\partial\big(s_3p_2(x,y)+s_2p_3(x,y)\big)}{\partial y}, \tag{20}$$

and requiring that $K(x,y) = 0 \ \ \forall\ (x,y)$.

Observe that this constraint is linear in the unknown variable $p_3(x,y)$ and we have one constraint for each position (x,y). Thus there may be sufficient information in these constraints to determine $p_3(x,y)$ uniquely, although possibly there are some linear dependencies between the constraints which would prevent uniqueness. It therefore seems wise to impose these constraints by least squares – i.e. write a quadratic cost function for $p_3(x,y)$ by summing the squares of $K(x,y)$ over (x,y) – *and* add an additional prior term. This gives an energy function:

$$E[P_3] = \int d\boldsymbol{x} K^2(\boldsymbol{x}) + \lambda \int d\boldsymbol{x}\{p_3(x,y) - \frac{1}{(s_2^2+s_3^2)}\big(-s_3h_2^p(x,y)+s_2h_3^p(x,y)\big)\}^2, \tag{21}$$

where λ is a constant and $h_2^p(x,y), h_3^p(x,y)$ are the y and z components of the prototype model for the right half of the face.

This completes the three stages. Results are shown in figures (11, 12,13).

7 Object Synthesis

This section briefly shows how to peform recognition by using our learned object models to synthesize images. The methods used are described in [2] which includes further examples.

Fig. 11. Left – the original input image. Center – the estimate of p3. Right – the estimate of the albedo.

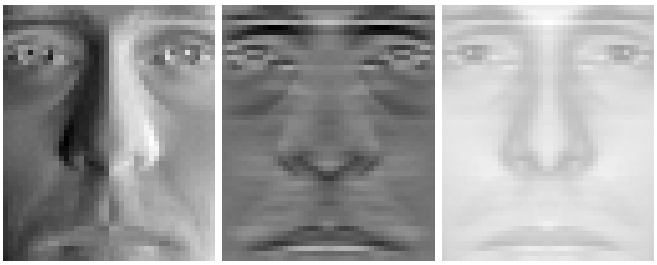

Fig. 12. Estimated b vectors of the face.

Fig. 13. Estimated normals of the face.

We first learned the illumination subspace for each face in the database, by determining b^* up to the $\mathbf{A}$ matrix. We then presented the algorithm with input images of the faces in the database seen under different lighting conditions. The algorithm estimates the best lighting conditions for generating the input assuming a Lambertian model – this is done by finding $\mathbf{s}^*$ to minimize $\sum_{x,y}\{I(x,y) - \mathbf{b}^*(x,y) \cdot \mathbf{s}\}^2$ – and then synthesizes the image using $\mathbf{s}^*$. The algorithm appeared to have no problem in estimating the correct lighting and in synthesizing an image similar to the input, even if the input image was taken under novel lighting conditions and included shadows and specularities, see figures (14,15,16).

Figure (14) shows some of the images used to construct the model. Figure (15) shows four of the input images to the system and figure (16) shows the result of using the algorithm to obtain synthesized images closest to the corresponding inputs. Observe that the synthesized images are similar except for certain shadows and specularities which cannot be synthesized using a purely Lambertian model. Although these shadows and specularities are perceptually salient, they are small in the least squares sense and do not prevent the light sources from being estimated accurately.

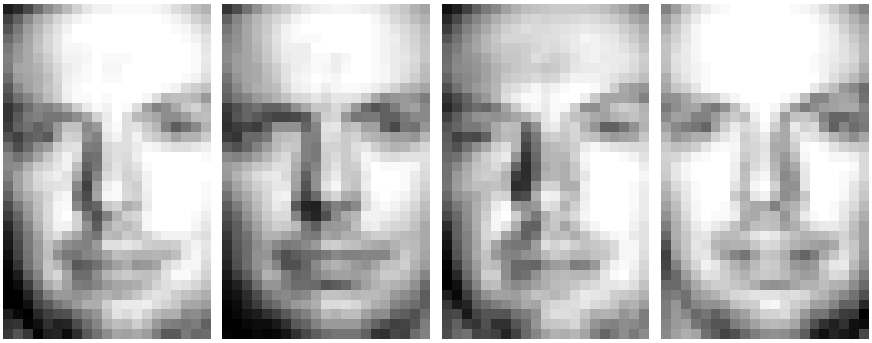

Fig. 14. Five of the original images used to construct the basis.

8 Conclusion

This paper developed a variety of techniques for learning models of the 3D shape and albedoes of objects. We demonstrated, using the dataset of faces constructed in [12], that the resulting models were fairly accurate and that they could be used to synthesize images of objects under arbitrary lighting conditions.

Our four learning schemes used different amounts of knowledge about the light source distribution and the object class. The first learning scheme assumed knowledge of light source directions and was equivalent to standard photometric

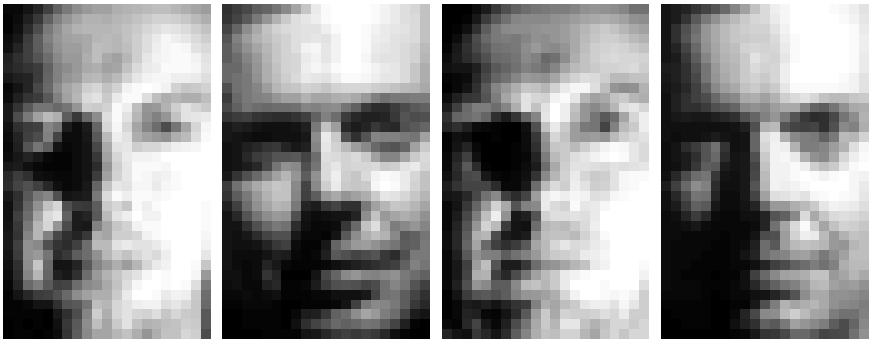

Fig. 15. Four of the input images used to test the the fitting algorithm.

Fig. 16. The synthesized images corresponding to the input images in the previous figure. They are found by first estimating the best principal light source direction and then reconstructing the best fit. Note that estimate of the best light source direction is found only up to an arbitrary invertible linear transformation.

stereo. The remaining three schemes used SVD to estimate light source directions, albedo, and shape up a linear transformation **A**. We discussed why it was uneccessary to know **A** in order to construct the light cone representation [2]. We also described ways to estimate **A** using surface integrability and/or prior knowledge about the object class.

While exploring surface integrability, we found an additional ambiguity in depth estimation which might be related to experimental findings by Koenderink [19]. This is being explored in current work.

We observed that surface integrability could be used to resolve an ambiguity in the SVD approach to photometric stereo [13] and described cases in which Hayawara's model would fail. Our work therefore has relevance to photometric stereo.

In addition, it has been applied to allowing for lighting variations of a moving object and hence improving tracking devices [10]. We are currently working on other applications and attempting to generalize to other reflectance functions.

9 Acknowledgments*

Support was provided by NSF Grant IRI 92-23676 and ARPA/ONR Contract N00014-95-1-1022. We thank David Mumford, Dan Kersten, David Kriegman and Mike Tarr for helpful discussions. Comments from three anonymous referees were also appreciated.

References

1. J.J. Atick, P.A. Griffin and A. N. Redlich. Statistical Approach to Shape from Shading. Preprint. Computational Neuroscience Laboratory. The Rockefeller University. New York. NY. 1995.
2. P. Belhumeur and D. Kriegman. "What is the set of images of an object under all lighting conditions?". In *Proceedings of Conference on Computer Vision and Pattern Recognition.* San Francisco. CA. 1996.
3. P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. "Eignefaces vs. Fisherfaces: Recognition using Class Specific Linear Projection". In *Proceedings of ECCV.* Cambridge, England. 1996.
4. A. Lanitis, C.J. Taylor and T.F. Cootes. A Unified approach to Coding and Interpreting Face Images. In *Proceedings of ICCV.* Boston. MA. 1995.
5. R. Epstein, P.W. Hallinan and A.L. Yuille. "5 ± Eigenimages Suffice: An Empirical Investigation of Low-Dimensional Lighting Models". In *Proceedings of IEEE WORKSHOP ON PHYSICS-BASED MODELING IN COMPUTER VISION.* 1995.
6. R. Epstein and A.L. Yuille. In preparation. 1996.
7. R.A. Fisher. The use of multiple measures in taxonomic problems. *Ann. Eugenics*, 7: pp 179-188. 1936.
8. U. Grenander, Y. Chow, and D.M. Keenan. **Hands: A Pattern Theroetic Study of Biological Shapes**. New York: Springer-Verlag. 1991.
9. W.E.L. Grimson. **Object Recognition by Computer.** MIT Press. Cambrige, MA. 1990.

10. G. Hager and P.N. Belhumeur. In *Proc ECCV*. Cambridge, England. 1996.
11. P.W. Hallinan. "A low-dimensional lighting representation of human faces for arbitrary lighting conditions". In. *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pp 995-999. 1994.
12. P.W. Hallinan. **A Deformable Model for Face Recognition under Arbitrary Lighting Conditions.** PhD Thesis. Division of Applied Sciences. Harvard University. 1995.
13. K. Hayakawa. "Photometric Stereo under a light source with arbitrary motion". *Journal of the Optical Society of America* A, 11(11). 1994.
14. B.K.P. Horn. **Computer Vision**. MIT Press, Cambridge, Mass. 1986.
15. B.K.P. Horn and M. J. Brooks, Eds. **Shape from Shading**. Cambridge MA, MIT Press, 1989.
16. P.J. Huber. **Robust Statistics**. John Wiley and Sons. New York. 1980.
17. Y. Iwahori, R.J. Woodham and A Bagheri "Principal Components Analysis and Neural Network Implementation of Photometric Stereo". In *Proceedings of the IEEE Workshop on Physics-Based Modeling in Computer Vision.* pp117-125 (1995).
18. J.J. Koenderink and A.J. van Doorn. "The internal representation of solid shape with respect to vision". *Biological Cybernetics*, Vol. 32. pp 211-216. 1979.
19. J.J. Koenderink, A.J. van Doorn, A.M.L. Kappers "Surface Perception in Pictures".*Perception and Psychophysics*, vol 52, pp487-496. 1992.
20. Y. Moses, Y. Adini, and S. Ullman. "Face Recognition: The problem of compensating for changes in the illumination direction". In *European Conf. on Comp. Vision.*, pp 286-296. 1994.
21. Y. Moses, S. Ullman, and S. Edelman. "Generalization to Novel Images in Upright and Inverted Faces". Preprint. Dept. of Applied Mathematics and Computer Science. The Weizmann Institute of Science. Israel. 1995.
22. J.L. Mundy and A. Zisserman (eds.) **Geometric Invariance in Computer Vision**. MIT Press. Cambridge, MA. 1992.
23. H. Murase and S. Nayar. "Visual learning and recognition of 3-D objects from appearance". *Int. Journal of Computer Vision.* 14. pp 5-24. 1995.
24. S.K. Nayar, K. Ikeuchi and T. Kanade "Surface reflections: physical and geometric perspectives" *IEEE trans. on Pattern Analysis and Machine Intelligence*, vol 13 p611-634. 1991.
25. A. Shashua. **Geometry and Photometry in 3D Visual Recognition**. PhD Thesis. MIT. 1992.
26. W. Silver. *Determining Shape and Reflectance Using Multiple Images.* PhD Thesis. MIT, Cambridge, MA. 1980.
27. M. Turk and A. Pentland. "Faces recognition using eigenfaces". In *Proceedings of IEEE Conf. on Comp. Vision and Pattern Recognition.* pp 586-591. 1991.
28. R. Woodham. "Analyszing Images of Curved Surfaces". *Artificial Intelligence*, 17, pp 117-140. 1981.
29. A.L. Yuille, M. Ferraro, and T. Zhang. "Shape from Warping". Submitted to CVGIP. 1996.
30. A.L. Yuille. "A Mathematical Analysis of SVD applied to lighting estimation and image warping". Preprint. Smith-Kettlewell Eye Research Institute. San Francisco. 1996.
31. Song Chun Zhu and A.L. Yuille. "A Flexible Object Recognition and Modelling System". In *Proceedings of the International Conference on Computer Vision.* Boston 1995.