International Journal of Computer Vision 21(1/2), 99–122 (1997)

On Photometric Issues in 3D Visual Recognition from a Single 2D Image

AMNON SHASHUA
The Hebrew University of Jerusalem, Institute for Computer Science, Jerusalem 91904, Israel
shashua@cs.huji.ac.il

Abstract. We describe the problem of recognition under changing illumination conditions and changing viewing positions from a computational and human vision perspective. On the computational side we focus on the mathematical problems of creating an equivalence class for images of the same 3D object undergoing certain groups of transformations—mostly those due to changing illumination, and briefly discuss those due to changing viewing positions. The computational treatment culminates in proposing a simple scheme for recognizing, via alignment, an image of a familiar object taken from a novel viewing position and a novel illumination condition. On the human vision aspect, the paper is motivated by empirical evidence inspired by Mooney images of faces that suggest a relatively high level of visual processing is involved in compensating for photometric sources of variability, and furthermore, that certain limitations on the admissible representations of image information may exist. The psychophysical observations and the computational results that follow agree in several important respects, such as the same (apparent) limitations on image representations.

1. Introduction

The problem of visual recognition is one of the well known challenges to researchers in human and machine vision. The task seems very easy and natural for biological systems, yet has proven to be very difficult to place within a comprehensive analytic framework. Some of the difficulties arise due to a lack of a widely accepted definition of what the problem is. For example, one can easily recognize scenes (such as a highway scene, city scene, restaurant scene, and so forth) (Potter, 1975) without an apparent need to recognize individual objects in the scene, nor to have a detailed recollection of the spatial layout of "things" in the scene. In this case it seems that some form of statistical regularity of scenes of a specific type is exploited, rather than what we normally associate with the task of "recognizing an object". Other difficulties arise due to hard mathematical problems in understanding the relationship between 3D objects and their images. For example, as we move our eyes, change position relative to the object, or move the object relative to ourselves, the image of the object undergoes change. Some of these changes are intuitive and include displacement and/or rotation in the image plane, but in general the changes are far from obvious. If the illumination conditions change, that is, the level of illumination, as well as the positions and distributions of light sources, then the image of the object changes as well. The light intensity distribution changes, and shadows and highlights may change their position.

In this paper we focus on the mathematical problems of creating an equivalence class for images of the same 3D object undergoing a certain group of changes. Before narrowing further the scope of discussion it may be worthwhile to consider further the types of "sources of variability" that are of general interest in recognition of individual objects. Following the seminal work of (Ullman, 1986), we distinguish four general sources of variability:

- *Photometric*: Changes in the light intensity distribution as a result of changing the illumination conditions.
- *Geometric*: Changes in the spatial location of image information as a result of a relative change of viewing position.

- *Varying Context*: Objects rarely appear in isolation and a typical image contains multiple objects that are next to each other or partially occluding each other. Changes in the image can, therefore, occur by changing the context without applying any transformation to the object itself.
- *Non-rigid Object Characteristics*: These include objects changing shape (such as facial expressions), objects having movable parts (like scissors), and so forth.

The photometric source of variability has to do with the relation between objects and the images they produce under changing conditions of illumination, i.e., changing the level of illumination, direction and number of light sources. This has the effect of changing the light intensity distribution in the image and the location of shadows and highlights. We will examine this issue later in the paper.

The geometric source of variability has to do with the geometric relation between rigid objects and their perspective images produced under changing viewing positions (relative motion between the viewer and the object). This is probably the most emphasized source of variability in computer vision circles and has received much attention in the context of recognition, structure from motion, visual navigation, and recently in the body of research on geometric invariants (e.g., (Mundy and Zisserman, 1992; Mundy et al., 1994)). We note that even relatively small changes in viewing position between two images of the same object often create a real problem in matching the two against each other. Figure 1 illustrates this point by superimposing two edge images of a face separated by a relatively small rotation around the vertical axis. We will discuss geometric issues of recognition later in Section 6, but for the most part of this paper we will focus only on the photometric source of variability.

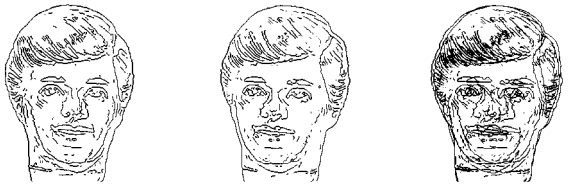

Figure 1. Demonstrating the effects of changing viewing position on the matching process. The difficulty of matching two different views can be illustrated by superimposing the two. One can see that, even for relatively small changes in viewing position, it could be very difficult to determine whether the two views come from the same face without first compensating for the effects of viewing transformation.

The third source of variability has to do with the effect of varying context. A typical image often contains multiple objects that are next to each other, or partially occluding each other. If we attempt to compare the entire image (containing a familiar object) to the model representation of an object in question, then we are unlikely to have a match between the two. The problem of varying context is, therefore, a question of how the image representation of an object (say its contours) can be separated from the rest of the image before we have identified the object. The problem is difficult and is often referred to as the problem of "segmentation", "grouping" or "selection". In the context of achieving recognition the crucial question is whether the problem of context can be approached in a bottom-up manner, i.e., irrespective of the object to be recognized, or whether it requires top-down processes as well. It appears that in some cases in human vision the processes for performing grouping and segmentation cannot be isolated from the recognition process. In some well known examples, such as R.C. James' image of a Dalmation dog (see, (Marr, 1982)), it appears unlikely that the image of the object can be separated from the rest of the image based on image properties alone and, therefore, some knowledge about the specific class of objects is required to interpret the image.

Human vision, however, appears also to contain relatively elaborate processes that perform grouping and segmentation solely on a data-driven basis independent of subsequent recognition processes. For example, Kinsbourne and Warrington (1962), cited in (Farah, 1990) report that patients with lessions in the left inferior temporo-occipital region are generally able to recognize single objects, but do poorly when more than one object is present in the scene. Another line of evidence comes from displays containing occlusions. The occluding stimuli, when made explicit, seem to stimulate an automatic 'grouping' process that groups together different parts of the same object (Nakayama et al., 1989). The third line of evidence comes from 'saliency' displays in which structures, not necessarily recognizable ones, are shown against a complex background. Some examples are shown in Fig. 2. In these displays, the figure-like structures seem to be detected immediately despite the lack of any apparent local distinguishing cues, such as local orientation, contrast

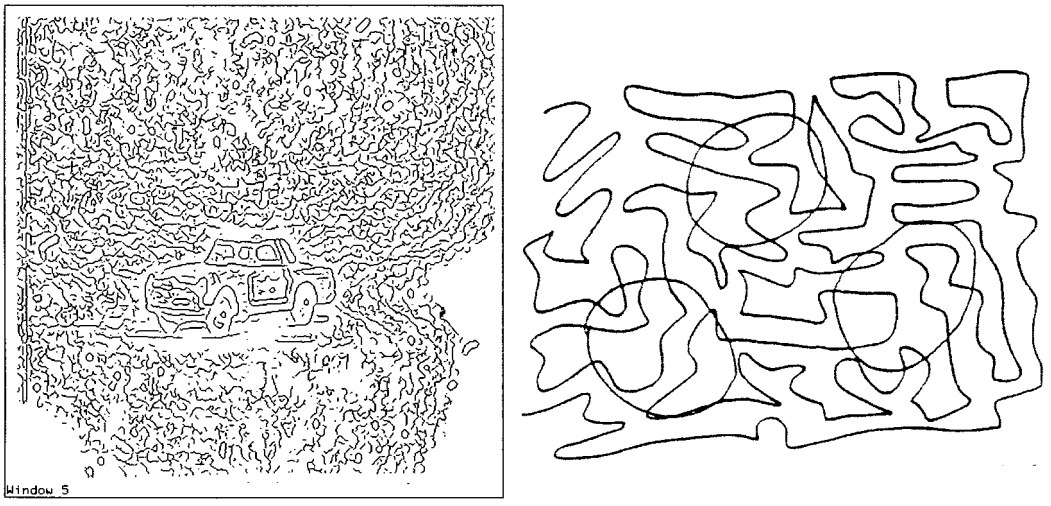

Figure 2. Structural-saliency displays. The figure like structures seem to 'pop-up' from the display, despite the lack of any apparent local distinguishing cues, such as local orientation, contrast and curvature (Shashua and Ullman, 1988).

and curvature (Shashua and Ullman, 1988, 1991). We will not consider further the problem of varying context, and assume instead that the region in the image containing the object has been isolated for purposes of recognition.

The fourth source of variability has to do with objects changing their shape. These include objects with movable parts (such as the human body) and flexible objects (for example, a face where the changes in shape are induced by facial expressions). This source of variability is geometrical, but unlike changing viewing positions, the geometric relation between objects and their images has less to do with issues of projective geometry and more to do with defining the space of admissible transformations in object space.

Our primary focus will be on the photometric source of variability, with some discussion on the geometric source. We discuss next in more detail the issues related to changing illumination in visual recognition.

2. The Photometric Source of Variability and Its Impact on Visual Recognition

The problem of varying illumination conditions, or the photometric source of variability as we refer to it here, raises the question of whether the problem can be isolated and dealt with independently of subsequent recognition processes, or whether it is coupled with the recognition process.

It appears that in some cases in human vision the effects of illumination are factored out at a relatively early stage of visual processing and independently of subsequent recognition processes. A well known example is the phenomenon of lightness and color constancy. In human vision the color of an object, or its greyness, is determined primarily by it's reflectance curve, not by the actual wavelengths that reach the observer's eye. This property of the visual system is not completely robust as it is known, for example, that fluorescent lighting alters our perception of colors. Nevertheless, this property appears to suggest that illumination is being factored out at an early stage prior to recognition. Early experiments that were used to demonstrate this used simple displays such as a planar ensemble of rectangular color patches, named after Mondrians' paintings, or comparisons among Munsel chips (Land and McCann, 1971). More recent psychophysical experiments demonstrated the effect of 3D structure on the perception of color and lightness (Gilchrist, 1979; Knill and Kersten, 1991). These experiments show that the perception of lightness changes with the perceived shape of the object. The objects that were used for these experiments are relatively simple, such as cylinders, polyhedrons and so forth. It is therefore conceivable that the 3D structure of the object displayed in these kinds of experiments can be re-constructed on the basis of image properties alone after which illumination effects can be factored out.

Another example of factoring out the illumination at an early stage, prior to and independently of recognition, is the use of edge detection. Edge detection is the most dominant approach to the problem of changing illumination and is based on recovering features from the image that are invariant to changes of illumination. The best known example of such features are step edges, i.e., contours where the light intensity changes relatively abruptly from one level to another. Such edges are often associated with object boundaries, changes in surface orientation, or material properties (Marr, 1976; Marr and Hildreth, 1980). Edge images contain most of the relevant information in the original grey-level image in cases where the information is mostly contained in changing surface material, in sharp changes in surface depth and/or orientation, and in surface texture, color, or greyness. In terms of 3D shape, these are characteristics of relatively simple objects. Therefore, the edges of simple objects are relatively informative (or recognizable) and will change only slightly when the illumination conditions change.

Many natural objects have a more complex structure, however: surface patches do not change orientation abruptly but rather smoothly. In this case, step edges may not be an ideal representation for two reasons: the edge image may not necessarily contain most of the relevant information in the grey-level image, and not all edges are stable with respect to changing illumination. For example, edges that correspond to surface inflections in depth are actually "phantom" edges and depend on the direction of light source (Moses, 1993).

Alternative edge detectors prompted by the need for more recognizable or more stable contour images search instead for extremal points of the light intensity distribution, known as valleys and ridges, or build up a "composite" edge representation made out of the union of step edges, valleys, and ridges (Freeman and Adelson, 1991; Morrone and Burr, 1988; Pearson et al., 1986; Perona and Malik, 1990). The composite edge images do not necessarily contain the subset of edges that are stable against changing illumination; they generally look better than step edges alone, but that varies considerably depending on the specific object.

The process of edge detection, producing step edges, ridges, valleys, and composite edge images, is illustrated in Figs. 3 and 4. In Fig. 3 three 'Ken' doll images are shown, each taken under a distinct illumination condition, with their corresponding step edges. In Fig. 4 the ridges, valleys, and the composite edge images of the three original images are shown (produced by Freeman and Adelson's (1991) edge and line detector). These results show the invariance of edges are not complete; some edges appear or disappear, some change location, and spurious edges result from shadows (especially attached shadows), specularities, and so forth.

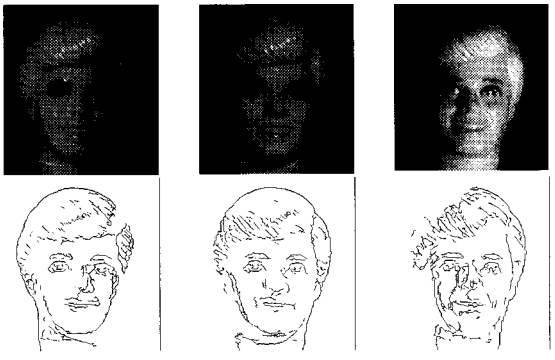

Figure 3. Grey-scale images of 'Ken' taken from three different illumination conditions. The bottom row shows the step edges detected by local energy measures followed by hysteresis (Freeman and Adelson, 1991). The step edges look very similar to the ones produced by Canny's edge detection scheme.

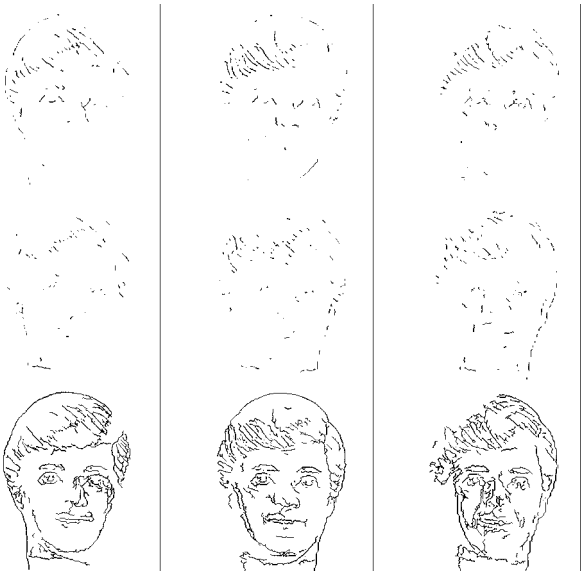

Figure 4. Valleys, ridges, and composite contour images produced by Freeman's contour detection method applied to the three images of the previous figure.

The 'Ken' images and their edge representations also demonstrate the practical side of the problem of recognition under changing illumination conditions. The

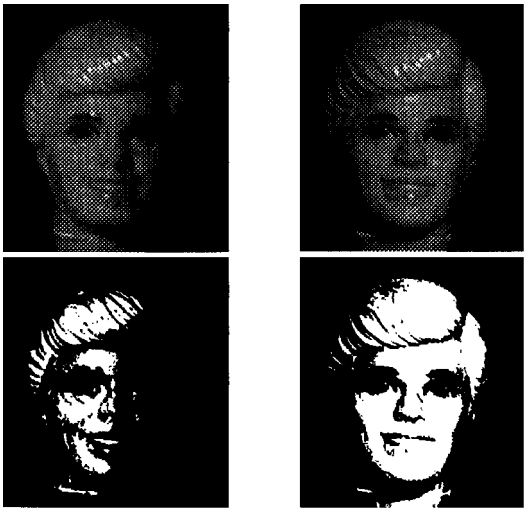

Figure 5. Images of 'Ken' taken from different illumination conditions followed by a thresholding operation. The recognizability of the thresholded images suggests that some knowledge about objects is required in order to factor out the illumination, and specifically that the image we are looking at is an image of a face.

images appear different to the degree that a template match between any two of them is not likely to succeed without first compensating for the changing illumination.

Human vision appears also to contain processes that factor out the effect of illumination during the recognition process. In other words, the image and the model are coupled together early on in the stages of visual processing. Consider, for example, the images displayed in Fig. 5. The images are of 'Ken' lit by two different illumination conditions, and thresholded by an arbitrary value. The thresholded images appear to be recognizable, at least in the sense that one can clearly identify the image as containing a face. Because the appearance of the thresholded images critically rely on the illumination conditions, it appears unlikely that recognition in this case is based on the input properties alone. Some knowledge about objects (specifically that we are looking at the image of a face) may be required in order to factor out the illumination.

Thresholded images are familiar in psychological circles, but less so in computational. A well-known example is the set of thresholded face images produced by Mooney (1960) for clinical recognizability tests, known as the closure faces test, in which patients had to sort the pictures into general classes that include: boy, girl, grown-up man or woman, old man or woman, and so forth. An example of Mooney's pictures are shown

Figure 6. Mooney faces and their level-crossings.

Figure 7. A less interpretable Mooney picture and its level-crossings.

in Fig. 6. Most of the control subjects could easily label most of the pictures correctly. Some of Mooney's pictures are less interpretable (for example, Fig. 7), but as a general phenomenon it seems remarkable that a vivid visual interpretation is possible from what seems an ambiguous collection of binary patches that do not bear a particularly strong relationship to surface structure or other surface properties.

Mooney images are sometimes referred to as representing the phenomenon of "shape from shadows" (Cavanagh, 1990). Although some Mooney images do contain cast shadows, the phenomenon is not limited to the difficulty of separating shadow borders from object contours. The thresholded image shown in Fig. 8, for example, is not less difficult to account for in computational terms, yet the original image was not lit in a way to create cast or attached shadows.

These kind of images appear also to indicate that in some cases in human vision the interpretation process involves more than just contours. It is evident that the contours (level-crossings) alone are not interpretable, as can be seen with the original Mooney pictures and

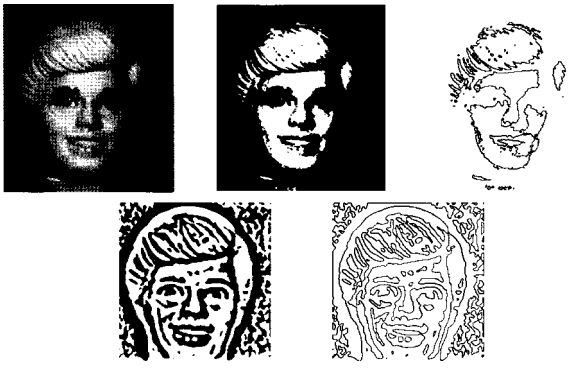

Figure 8. *Top row*: a 'Ken' image represented by grey-levels, the same image followed by a threshold, the level-crossings of the thresholded image. The thresholded image shown in the center display is difficult to account for in computational terms, yet the original image was not lit in a way to create cast or attached shadows. *Bottom row*: the sign-bits of the Laplacian of Gaussian operator applied to the original image, and its zero-crossings (step edges). Interpretability of the sign-bit image is considerably better than the interpretability of the zero-crossings.

with the level-crossing image in Fig. 8. It seems that only when the distinction of what regions are above the threshold and what are below the threshold is made clear (we refer to that as adding "sign-bits") does the resulting image become interpretable. This appears to be true not only for thresholded images but also for step edges and their sign-bits (see Fig. 8, bottom row).

It appears, therefore, that in some cases in human vision the illumination is factored out within the recognition process using top-down information and that the process responsible apparently requires more than just contours—but not much more. We refer from here on to the Mooney-kind of images as reduced images. From a computational standpoint we will be interested not only in factoring out the illumination, in a model-based approach, but also in doing so from reduced images.

3. Problem Scope

The recognition problem we consider is that of identifying an image of an arbitrary individual 3D object. We allow the object to be viewed from arbitrary viewing positions and to be illuminated by an arbitrary setting of light sources. We assume that the image of the object is already separated from the rest of the image, but may have missing parts (for example, as caused by occlusion).

We adopt the alignment methodology, which defines "success" as the ability to exactly re-construct the input image representation of the object (possibly viewed under novel viewing and illumination conditions) from the model representation of the object stored in memory. Alignment has been studied in the past for compensating for changes of viewing positions and the method is typically realized by storing a few number of "model" views (two, for example), or a 3D model of the object, and with the help of corresponding points between the model and any novel input view, the object is "re-projected" onto the novel viewing position. Recognition is achieved if the re-projected image is successfully matched against the input image (Fischler and Bolles, 1981; Huttenlocher and Ullman, 1990; Lowe, 1985; Ullman, 1986; Ullman and Basri, 1989). Our approach is to apply the basic concept of alignment onto the photometric domain, and then combine both sources of variability into a single alignment framework that can deal with both changes due to geometry and photometry occurring simultaneously.

The basic method, we call *photometric alignment*, for compensating for the effects of illumination during recognition is introduced next. The method is based on a result that three images of the surface provide a basis that spans all other images of the surface (same viewing position, but changing illumination conditions). The photometric problem of recognition is, therefore, reduced to the problem of determining the linear coefficients—which is conceptually similar to the idea of Ullman and Basri (1989) in the geometric context. We then extend the basic method to deal with situations of recognition from reduced image representations with results that appear to agree with the empirical observation made earlier that sign-bits appear to be sufficient for visual interpretation, whereas edges alone do not.

4. Photometric Alignment

The basic approach is based on finding an algebraic connection between all images of an object taken under varying illumination conditions. We start by defining the family of surface reflectance functions for which our results will hold:

Definition. An order k Linear Reflectance Model is defined as the scalar product $\mathbf{x} \cdot \mathbf{a}$, where $\mathbf{x}$ is a vector in k-dimensional Euclidean space of invariant surface properties (such as surface normal, surface albedo, and so forth), and $\mathbf{a}$ is an arbitrary vector (of the same dimension).

The Lambertian model of reflection is an obvious case of an order 3 linear reflectance model. The grey-value, $I(p)$, at location p in the image can be represented by the scalar product of the surface normal vector and the light source vector,

$$I(p) = \mathbf{n}_p \cdot \mathbf{s}.$$

Here the length of the surface normal $\mathbf{n}_p$ represents the surface albedo (a scalar ranging from zero to one). The length of the light source vector $\mathbf{s}$ represents a mixture of the spectral response of the image filters, and the spectral composition of light sources—both of which are assumed to be fixed for all images of the surface (we assume for now that light sources can change direction and level of intensity but not spectral composition).

Another example of a linear reflectance model is the image irradiance of a tilted Lambertian surface under a hemispherical sky. Horn (1986, p. 234) shows that the image irradiance equation is $E\delta_p \cos^2 \frac{\alpha}{2}$, where α is the angle between the surface normal and the zenith, E is the intensity of light source, and δ_p is the surface albedo. The equation is an order 4 linear reflectance function:

$$\begin{aligned} I(p) &= \frac{1}{2} E\delta_p(1+\cos\alpha) = \mathbf{n}_p \cdot \mathbf{s} + |\mathbf{n}_p| \cdot |\mathbf{s}| \\ &= (\mathbf{n}_p, |\mathbf{n}_p|)^t(\mathbf{s}, |\mathbf{s}|), \end{aligned}$$

where $\mathbf{s}$ represents the direction of zenith, whose length is $\frac{E}{2}$.

Proposition 1. *An image of an object with an order k linear reflection model $I(p) = \mathbf{x}(p) \cdot \mathbf{a}$ can be represented as a linear combination of a fixed set of k images of the object.*

Proof: Let $\mathbf{a}_1, \ldots, \mathbf{a}_k$ be some arbitrary set of basis vectors that span k-dimensional Euclidean space. The image intensity $I(p) = \mathbf{x}(p) \cdot \mathbf{a}$ is therefore represented by

$$\begin{aligned} I(p) &= \mathbf{x}(p)[\alpha_1\mathbf{a}_1 + \cdots + \alpha_k\mathbf{a}_k] \\ &= \alpha_1 I_1(p) + \cdots + \alpha_k I_k(p), \end{aligned}$$

where $\alpha_1, \ldots, \alpha_k$ are the linear coefficients that represent $\mathbf{a}$ with respect to the basis vectors, and $I_1, \ldots, I_k$ are the k images $I_k(p) = \mathbf{x}(p) \cdot \mathbf{a}_k$. □

To see the relevance of this proposition to visual recognition, consider the case of a Lambertian surface under a point light source (or multiple point light sources). Assume we take three pictures of the object I_1, I_2, I_3 from light source directions $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$, respectively. The linear combination result is that any other image I of the object, taken from a novel setting of light sources, is simply a linear combination of the three pictures,

$$I(p) = \alpha_1 I_1(p) + \alpha_2 I_2(p) + \alpha_3 I_3(p),$$

for some coefficients $\alpha_1, \alpha_2, \alpha_3$. The coefficients can be solved by observing the grey-values of three points providing three equations. Using more than three points will provide a least squares solution. The solution is unique provided that $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$ are linearly independent, and that the normal directions of the three sampled points span all other surface normals (for a general 3D surface, for example, the three normals should be linearly independent).

Alignment-based recognition under changing illumination can proceed in the following way. The images $I_1, \ldots, I_k$ are the model images of the object (three for Lambertian under point light sources). For any new input image I, rather than matching it directly to previously seen images (the model images), we first select a number of points (at least k) to solve for the coefficients, and then synthesize an image $I' = \alpha_1 I_1 + \cdots + \alpha_k I_k$. If the image I is of the same object, and the only change is in illumination, then I and I' should perfectly match (the matching is not necessarily done at the image intensity level, one can match the edges of I against the edges of I', for example). This procedure has factored out the effects of changing illumination from the recognition process without recovering scene information, i.e., surface albedo or surface normal, and without assuming knowledge of direction of light sources. Another property of this method is that one can easily find a least squares solution for the reconstruction of the synthesized image, thereby being less sensitive to errors in the model, or in the input.

We address below the problems that arise when some of the objects points are occluded from some of the light sources, and when the surface reflects light specularly. We then extend the results to deal with cases of changing spectral composition of light sources.

4.1. Attached and Cast Shadows

We have practically assumed that surfaces are convex because the linear combination result requires that

points be visible to the light sources. In a general non-convex surface, object points may be occluded from some, or from all, the light sources. This situation generally leads to two types of shadows known as attached and cast shadows. A point P is in an attached shadow if the angle between the surface normal and the direction of light source is obtuse ($\mathbf{n}_p \cdot \mathbf{s} < 0$). An object point P is in a cast shadow if it is obstructed from the light source by another object or by part of the same object. An attached shadow, therefore, lies directly on the object, whereas cast shadows are thrown from one object onto another, or from one part onto another of the same object (such as when the nose casts a shadow on the cheek under oblique illumination).

In the case of attached-shadows, a correct reconstruction of the image grey-value at p does not require that the object point P be visible to the light source $\mathbf{s}$, but only that it be visible to the light sources $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$. If P is not visible to $\mathbf{s}$, then the linear combination will produce a negative grey-value (because $\mathbf{n}_p \cdot \mathbf{s} < 0$), which can be set to 0 for purposes of display or recognition.

If P is not visible to one of the model light sources, say $\mathbf{s}_1$, then the linear combination of the three model images will produce $I'(p)$ under a light source $\mathbf{s}'$ which is the projection of $\mathbf{s}$ onto the sub-space spanned by $\mathbf{s}_2, \mathbf{s}_3$. This implies that photometric alignment would perform best in the case where the novel direction of light source $\mathbf{s}$ is within the cone of directions $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$.

The remaining case is when the object point P is in a cast shadow region with respect to the novel light direction $\mathbf{s}$. In this case there is no way to predict a low, or zero, grey-value for $I'(p)$ and the reconstruction will not match $I(p)$ in that region. Therefore, cast shadow regions in the novel image are not modeled in this framework, and hence, the performance degrades with increasing number and extent of cast-shadows in the novel image.

With regard to human vision, there appears to be a marked increase in difficulty in interpreting cast shadows compared to attached shadows. Arnheim (1954) discusses the effect of cast shadows on visual perception, its relation to chiaroscuro in Renaissance art, and its symbolism in various cultures. He points out that cast shadows often interfere with the object's integrity, whereas attached shadows are often perceived as an integral part of the object. The general observation is that the more the cast-shadow extends from the part that throws it, the less meaningful is the connection made with the object. The interpretability of cast shadows can be illustrated by 'Ken' images displayed in Fig. 3. The three model images have extensive attached shadows that appear naturally integrated with the object. The cast shadow region thrown from the nose in the image on the right appears less integrated with the overall composition of the image.

In conclusion, attached shadows in the novel image, or shadows in general in the model images, do not have significant adverse effects on the photometric alignment scheme. Cast shadows in the novel image, cannot be reconstructed or even approximated, and therefore are not modeled in this framework. It may be noted that apparently there is a perceptual difference between attached and cast shadows, whereby the latter may appear to be disconnected from the object upon which they are cast.

4.2. Detecting and Removing Specular Reflections

The linear combination result and the photometric alignment scheme that followed assume that objects are matte. In general, inhomogeneous surfaces are dominantly Lambertian, except for isolated regions that are specularly reflecting light. In practice, if the specular component is ignored, the reconstructed image contains the specular regions of all three model images combined together, and the specular regions of the novel image are not reconstructed. For purposes of recognition, as long as the specular regions are relatively small, they do not seem to have a significant adverse effect on the overall photometric alignment scheme. Nevertheless, the alignment method can be used to detect the specular regions and replace them with the Lambertian reflectance provided that four images are used.

The detection of specular points is based on the observation that if a point is in the specular lobe, then it is likely to be so only in one of the images at most. This is because the specular lobe occupies a region that falls off exponentially from the specular direction. In general we cannot detect the specular points by simply comparing grey-values in one image with the grey-values of the same points in the other images because the intensity of the light source may arbitrarily change from one image to another.

By using Proposition 1, that is, the result that three images uniquely determine the Lambertian component of the fourth image, we can, thereby, compare the reconstructed intensity of the fourth image with the

observed intensity, and check for significant deviations. For every point p, we select the image with the highest intensity, call it I_s, and reconstruct $I'_s(p)$ from the other three images (we recover the coefficients once, based on points that are not likely to be specular or shadowed, i.e., do not have an especially high or low intensity). If $I_s(p)$ is in the specular lobe, then $I'_s(p) \ll I_s(p)$. To avoid deviations that are a result of shadowed points, we apply this procedure to points for which none of the images has an especially low grey-value.

In practice we observe that the deviations that occur at specular points are of an order of magnitude higher than deviations anywhere else, which makes it relatively easy to select a threshold for deciding what is specular and what is not. A similar approach for detecting specular points was suggested by (Coleman and Jain, 1982) based on photometric stereo (Woodham, 1980). The idea is to have four images and to reconstruct the normal at each point from every subset of three images. If the point in question is not significantly specular, then the reconstructed normals should have the same direction and length, otherwise the point is likely to be specular. Their method, however, requires knowledge of direction and intensity of light sources, whereas in our method we do not.

4.3. *Some Experiments*

We used the three 'Ken' images displayed in Fig. 3 as model images for the photometric alignment scheme. The surface of the doll is non-convex and not purely matte which gives rise to specular reflections and shadows. The novel image (shown in Fig. 9) was taken using light source directions that were within the cone of directions used to create the model images. In principle, one can use novel light source directions that are outside the cone of directions, but that will increase the likelihood of creating new cast shadow regions. The reconstruction was based on a least squares solution using eight points. The points were chosen automatically by searching for smooth regions of image intensity. The search was restricted to the area of the face, not including the background. To minimize the chance of selecting shadowed or specular points, a point was considered as an admissible candidate if it was contained in an 8×8 sized smooth area, and its intensity was not at the low or high end of the spectrum. We then selected eight points that were widely separated from each other. The reconstructed image (linear combination of the three model images) is displayed in Fig. 9

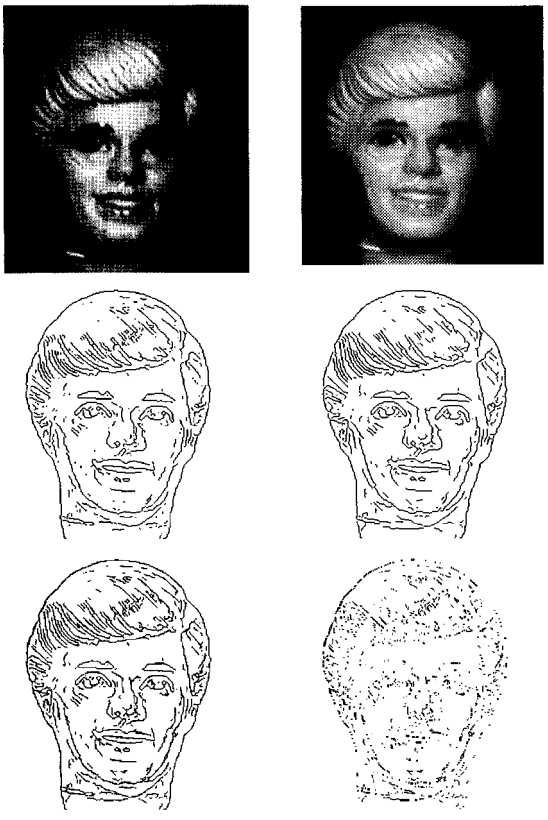

Figure 9. Reconstructing a novel image. *Row* 1 (*left to right*): a novel image taken from two point light sources, and the reconstructed image (linear combination of the three model images). *Row* 2: step edges of the novel and reconstructed images. *Row* 3: overlaying both edge maps, and subtracting (*xor* operation) the edge maps from each other. The difference between the images both at the grey-scale and edge level is hardly noticeable.

together with its step edges. The novel and reconstructed image are visually very similar at the grey-value level, and even more so at the edge-map level. The difference between the two edge maps is negligible and is mostly due to quantization of pixel locations.

In conclusion, this result shows that for the purposes of recognition, the existence of shadows and (small) specular regions in the model images do not have a significantly adverse effect on the reconstruction. Moreover, we did not use a matte surface for the experiment, illustrating the point that plastic surfaces are dominantly Lambertian, and therefore sufficiently applicable to this method.

Figure 10 demonstrates the specular detection scheme. The method appears to be successful in identifying small specular regions. Other schemes for

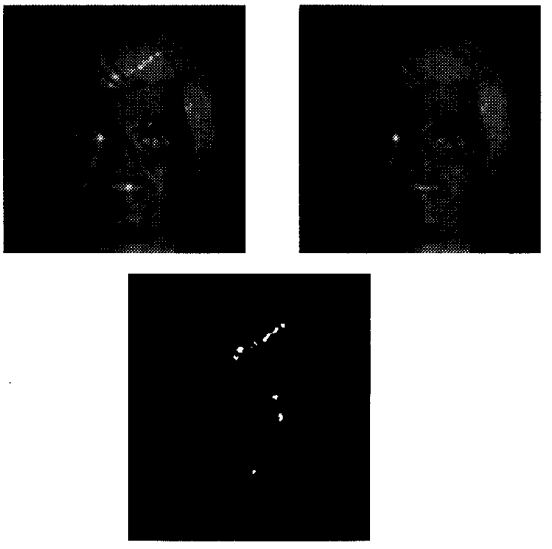

Figure 10. Detecting and removing specular regions. *Row* 1: the image on the left is a novel image, and the one on the right is the same image following the procedure for detecting and removing the specular regions. The specular regions are replaced with the reconstructed grey-value from the model images. *Row* 2: the specular regions that were detected from the image.

detecting specular regions using the dichromatic model of reflection often require a relatively large region of analysis and, therefore, would have difficulties in detecting small specular regions (Klinker et al., 1990; Shafer, 1985).

4.4. The Linear Combination of Color Bands

The photometric problem considered so far involved only changes in direction and intensity of light sources, but not changes in their spectral compositions. Light sources that change their spectral composition are common as, for example, sunlight changes its spectral composition depending on the time of day (because of scattering). The implication for recognition, however, is not entirely clear because there may be an adaptation factor involved rather than an explicit process of eliminating the effects of illumination. Adaptation is not a possibility when it comes to changing direction of light source, because objects are free to move in space and hence change their positions with respect to the light sources. Nevertheless, it is of interest to explore the possibility of compensating for changing spectral composition as well as direction of light sources.

We assume, for reasons that will be detailed below, that our surface is either *neutral*, or is of the same color, but may change in luminosity. A neutral surface is a grey-scale surface only affecting the scale of light falling on the surface, but not its spectral composition. For example, the shades of grey from white to black are all neutral. Note that the assumption is weaker than the uniform albedo assumption because we allow change in luminosity, but is less general than what we had previously because we do not allow changes in hue or saturation to occur across the surface. We also assume that our model of the object consists of a single color image obtained by overlaying three color images of the object each taken from a distinct direction of light source having a distinct spectral composition.

Let I_r, I_g, I_b be the three color bands that together define the color picture. Let $\delta_p \rho(\lambda)$ be the surface reflectance function, where δ_p is the surface albedo and $\rho(\lambda)$ is the spectral reflectance function of wavelength λ. Note that the neutral surface assumption means that across the surface $\rho(\lambda)$ is fixed, but δ_p may change arbitrarily. Let $S_1(\lambda), S_2(\lambda), S_3(\lambda)$ be the spectral composition of the three light sources, and $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$ be their directions. As before, we require that the directions be non-coplanar, and that the spectral compositions be different from each other. This, however, does not mean that the three spectral functions should form a basis (such as required in some color constancy models (Maloney and Wandell, 1986)). Finally, let $R_r(\lambda), R_g(\lambda), R_b(\lambda)$ be the spectral sensitivity functions of the three CCD filters (or retinal cones). The composite color picture (taking the picture separately under each light source, and then combining the results) is, therefore, determined by the following equation:

$$\begin{pmatrix} I_r(p) \\ I_g(p) \\ I_b(p) \end{pmatrix} = \begin{pmatrix} \int S_1(\lambda)\rho(\lambda)R_r(\lambda)d\lambda \\ \int S_1(\lambda)\rho(\lambda)R_g(\lambda)d\lambda \\ \int S_1(\lambda)\rho(\lambda)R_b(\lambda)d\lambda \end{pmatrix} \mathbf{n}_p \cdot \mathbf{s}_1$$

$$+ \begin{pmatrix} \int S_2(\lambda)\rho(\lambda)R_r(\lambda)d\lambda \\ \int S_2(\lambda)\rho(\lambda)R_g(\lambda)d\lambda \\ \int S_2(\lambda)\rho(\lambda)R_b(\lambda)d\lambda \end{pmatrix} \mathbf{n}_p \cdot \mathbf{s}_2$$

$$+ \begin{pmatrix} \int S_3(\lambda)\rho(\lambda)R_r(\lambda)d\lambda \\ \int S_3(\lambda)\rho(\lambda)R_g(\lambda)d\lambda \\ \int S_3(\lambda)\rho(\lambda)R_b(\lambda)d\lambda \end{pmatrix} \mathbf{n}_p \cdot \mathbf{s}_3,$$

where the length of $\mathbf{n}_p$ is δ_p. This can re-written in matrix form, as follows:

$$\begin{pmatrix} I_r(p) \\ I_g(p) \\ I_b(p) \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \mathbf{n}_p \cdot \mathbf{s}_1 + \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \mathbf{n}_p \cdot \mathbf{s}_2 + \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} \mathbf{n}_p \cdot \mathbf{s}_3 = [\mathbf{v}, \mathbf{u}, \mathbf{w}] \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \mathbf{s}_3 \end{bmatrix} \mathbf{n}_p = A\mathbf{n}_p.$$

The 3×3 matrix $[\mathbf{v}, \mathbf{u}, \mathbf{w}]$ is assumed to be non-singular (for that reason we required that the spectral composition of light sources be different from one another), and therefore the matrix A is also non-singular. Note that because of the assumption that the surface is neutral, the matrix A is independent of position. Consider any novel image of the same surface, taken under a new direction of light source with a possible different spectral composition. Let the novel picture be J_r, J_g, J_b. The red color band, for instance, can be represented as a linear combination of the three color bands I_r, I_g, I_b, as follows:

$$\begin{aligned} J_r(p) &= \left[\int S(\lambda)\rho(\lambda)R_r(\lambda)d\lambda\right]\mathbf{n}_p \cdot \mathbf{s} \\ &= \mathbf{n}_p \cdot (\alpha_1 A_1 + \alpha_2 A_2 + \alpha_3 A_3) \\ &= \alpha_1 I_r(p) + \alpha_2 I_g(p) + \alpha_3 I_b(p) \end{aligned}$$

where A_1, A_2, A_3 are the rows of the matrix A. Because A is non-singular, the row vectors form a basis that spans the vector $[\int S(\lambda)\rho(\lambda)R_r(\lambda)d\lambda]\mathbf{s}$ with some coefficients $\alpha_1, \alpha_2, \alpha_3$. These coefficients are fixed for all points in the red color band because the scale $\int S(\lambda)\rho(\lambda)R_r(\lambda)d\lambda$ is independent of position (the neutral surface albedo δ_p is associated with the length of $\mathbf{n}_p$). Similarly the remaining color bands J_g, J_b are also represented as a linear combination of I_r, I_g, I_b, but with different coefficients. We have, therefore, arrived at the following result:

Proposition 2. *An image of a Lambertian object with a neutral surface reflectance* (*grey-scale surface*) *taken under an arbitrary point light source condition* (*intensity, direction and spectral composition of light source*) *can be represented as a linear combination of the three color bands of a model picture of the same object taken under three point light sources having different* (*non-coplanar*) *directions and different spectral composition.*

For a neutral surface, the linear combination of color bands can span only images of the same surface with the same hue and saturation under varying illumination conditions. The combination of color bands of a non-neutral surface spans the space of illumination *and* color (hue and saturation). That is, two surfaces with the same structure but with different hue and saturation levels, are considered the same under the photometric alignment scheme.

5. Photometric Alignment with Reduced Images

We have seen in Section 2 empirical evidence to suggest that in some cases the process responsible for factoring out the illumination during the recognition process appears to require more than just contour information, but just slightly more. So far we proposed a scheme which can directly factor out the illumination during the model-to-image matching stage by using the information contained in the grey-values of the model and novel images.

In this section we explore the possibilities of using less than grey-values for purposes of factoring out the illumination. In other words, since the photometric alignment method is essentially about recovering the linear coefficients that represent the novel image as a linear combination of the three model images, then the question is whether those coefficients can be recovered by observing more reduced representations of the novel image, such as edges, edges and gradients, sign-bits, and so forth. Specifically, we are most interested in making a computational connection with the empirical observation that sign-bits appear to be sufficient for visual interpretation, whereas edges alone are not.

The proposition below shows that in principle the level-crossing or zero-crossing contours of the novel image are theoretically sufficient for recovering the linear coefficients for combining the model images.

Proposition 3. *The coefficients that span an image I from three model images, as described in Proposition* 1 *can be solved, up to a common scale factor, from just the contours of I, zero-crossings or level-crossings.*

Proof: Let α_j be the coefficients that span I by the basis images I_j, $j = 1, 2, 3$, i.e., $I = \sum_j \alpha_j I_j$. Let

f, f_j be the result of applying a Laplacian of Gaussian (LOG) operator, with the same scale, on images I, I_j, $j = 1, 2, 3$. Since LOG is a linear operator we have $f = \sum_j \alpha_j f_j$. Since $f(p) = 0$ along zero-crossing points p of I, then by taking three zero-crossing points, which are not on a cast shadow border and whose corresponding surface normals are non-coplanar, we get a homogeneous set of equations from which α_j can be solved up to a common scale factor.

Similarly, let k be an unknown threshold applied to I. Therefore, along level crossings of I we have $k = \sum_j \alpha_j I_j$; hence four level-crossing points that are visible to all four light sources are sufficient for solving α_j and k. □

The result is that in principle we could cancel the effects of illumination directly from the zero-crossings (or level-crossings) of the novel image instead of from the raw grey-values of the novel image. Note that the model images are represented as before by grey-values. Because the model images are taken only once, it is not unreasonable to assume more strict requirements on the quality of those images. We therefore make a distinction between the model acquisition, or learning, phase for which grey-values are used and the recognition phase for which a reduced representation of the novel image is being used.

The result that contours may be used instead of grey-values is not surprising at a theoretical level, considering the literature of image compression. Under certain restrictions on the class of signals, it is known that the zero-crossings form a complete representation of an arbitrary signal of that class. The case of one-dimensional bandpass signals, with certain conditions on the signals' Hilbert transform, is provided by (Logan, 1977). The more general case is approached by assuming the signal can be represented as a finite complex polynomial (Curtis et al., 1985; Sanz and Huang, 1989). Complex polynomials have the well known property that they are fully determined by their analytic varieties (curves in the one-dimensional case) using analytic continuation methods (see for example, (Saff and Snider, 1976)). It is well known that analytic continuation is an unstable process (Hille, 1962) and therefore, the reconstruction of the image from its zero-crossings is likely to be unstable. Curtis et al. (1985) report, for instance, that zero-crossings must be recorded with great precision, at sub-pixel accuracy of 14 digits.

The result of Proposition 3 can be viewed as a model-based reconstruction theorem, that applies to a much less restricted class of signals (images do not have to be bandpass, for instance). The process is much simpler, but on the other hand it is restricted to a specific model undergoing a restricted group of transformations (changing illumination). The simplicity of the model-based reconstruction, however, is not of great help in circumventing the problem of instability. Stability depends on whether contours are recorded accurately and whether those contours are invariant across the model images.

The assumption that the value of f at a zero-crossing location p is zero, is true for a subpixel location p. In other words, it is unlikely that $f(p) = 0$ for some integral location p. This introduces, therefore, a source of error whose magnitude depends on the 'strength' of the edge that gives rise to the zero-crossing in the signal f, that is, the sharper and stronger the discontinuity in image intensities along an edge in the image I is, the larger the variance around $f(p)$. This suggests that 'weak' edges should be sampled, with more or less the same strength, so that by sampling more than the minimum required number of points, the error could be canceled by a least squares solution.

The second source of error has to do with the stability of the particular edge under changing illumination. Assume, for example, that the zero-crossing at p (recorded accurately) is a result of a sharp change in surface reflectance. Although the image intensity distribution around p changes across the model images, the location of the discontinuity does not, i.e., the zero-crossing is stable. In this case we have that $f(p) = f_j(p) = 0$, $j = 1, 2, 3$. Therefore, such a point will not contribute any information if recorded accurately and will contribute pure noise if recorded with less than the required degree of accuracy. This finding suggests, therefore, that zero-crossings should be sampled along attached shadow contours or along valleys and ridges of image intensities (a valley or a ridge gives rise to two unstable zero-crossings (Moses, 1993)).

The situation with reconstruction from level-crossings is slightly different. The first source of error, related to the accuracy in recording the location of level-crossings, still applies, but the second source does not. In general, the variance in intensity around a level crossing point p is not as high as the variance around an edge point. A random sampling of points for a least squares solution is not likely to have a zero mean

error, however, and the mean error would therefore be absorbed in the unknown threshold k. The least squares solution would be biased towards a zero mean error solution that will affect both the recovered threshold and the linear coefficients α_j. The solution, therefore, does not necessarily consist of a correct set of coefficients and a slightly off threshold k, but a mixture of both inaccurate coefficients and an inaccurate threshold. This implies that level-crossings should be sampled at locations that do not correspond to zero-crossings in order to minimize the magnitude of errors.

To summarize, the reconstruction of the novel image from three model images and the contours of the novel image is possible in principle. In the case of both zero-crossings and level-crossings, the locations of the contours must be recorded at sub-pixel accuracy. In the case of zero-crossings, another source of potential error arises, which is related to the stability of the zero-crossing location under changing illumination. Therefore, a stable reconstruction requires a sample of points along weak edges that correspond to attached shadow contours or to ridges and valleys of intensity. Alternatively, the locations of contour points must be recorded at sub-pixel accuracy, given also that the sample is large enough to contain unstable points with respect to illumination. Experimental results show that a random sample of ten points (spread evenly all over the object) with accuracy of two digits for zero-crossings and one digit for level-crossings is sufficient to produce results comparable to those produced from sampling image intensities directly. The performance with integral locations of points sampled over edges p that have no corresponding edges in a 3×3 window around p in any of the model images was not satisfactory.

These results show that reconstruction from contours does not appear to be generally useful for the photometric alignment scheme because of its potential instability. It is also important to note that in these experiments the viewing position is fixed, thereby eliminating the correspondence problem that would arise otherwise and would most likely increase the magnitude of errors.

5.1. *Photometric Alignment from Contours and Gradients*

When zero-crossings are supplemented with gradient data, the reconstruction does no longer suffer from the two sources of errors that were discussed in the previous section. We can use gradient data to solve for the coefficients, because the operation of taking derivatives (continuous and discrete) is linear and therefore leaves the coefficients unchanged. The accuracy requirement is relaxed because the gradient data is associated with the integral location of contour points, not with their sub-pixel location. Stable zero-crossings do not affect the reconstruction, because the gradient depends on the distribution of grey-values in the neighborhood of the zero-crossing, and the distribution changes with a change in illumination (even though the location of the zero-crossing may not change).

Errors, however, may be more noticeable once we allow changes in viewing positions in addition to changes in illumination. Changes in viewing positions may introduce errors in matching edge points across images. Because the change in image intensity distribution around an edge point is localized and may change significantly at nearby points, then errors in matching edge points across the model images may lead to significant errors in the contribution those points make to the system of equations.

5.2. *Photometric Alignment from Sign-Bits*

Reconstruction from contours, general or model-based, appears to rely on the accurate location of contours. This reliance, however, seems to be at odds with the intuitive interpretation of Mooney-type pictures, like those in Figs. 6. These images suggest that, instead of contours being the primary vehicle for shape interpretation, the regions bounded by the contours (the sign-bit regions) are primarily responsible for the interpretation process. Thus instead of a contour-based technique we investigate below an area-based technique arising from the use of sign-bits. It is worth noting that the "sign-bit correlation" method for stereo matching proposed by Nishihara (1984) was advocated on similar grounds. Nishihara's conclusion was that the sign-bits contributed to increased stability (because regions change less than contours do)—conclusions that are similar to what we find below. It is also worthwhile noting that, theoretically speaking, only one bit of information is added in the sign-bit displays. This is because zero-crossings and level-crossings form nested loops (Koenderink and Van Doorn, 1980), and therefore the sign-bit function is completely determined up to a common sign flip. In practice, however, this property of contours does not emerge from edge detectors because weak contours are often thresholded out as they tend to be the most sensitive to noise (see, for example, Fig. 3). This may also explain why our visual system

apparently does not use this property of contours. We therefore do not make use of the global property of the sign-bit function; rather, we treat it as a local source of information, i.e., one bit of information per pixel.

Because the location of contours is an unreliable source of information, especially when the effects of changing viewing positions are considered, we propose to rely instead only on the sign-bit source of information. From a computational standpoint, the only information that a point inside a region can provide is whether the function to be reconstructed (the filtered image f, or the thresholded image I) is positive or negative (or above/below threshold). This information can be incorporated in a scheme for finding a separating hyperplane, as suggested in the following proposition:

Proposition 4. *Solving for the coefficients from the sign-bit image of I is equivalent to solving for a separating hyperplane in 3D or 4D space in which image points serve as "examples".*

Proof: Let $\mathbf{z}(p) = (f_1, f_2, f_3)^T$ be a vector function and $\boldsymbol{\omega} = (\alpha_1, \alpha_2, \alpha_3)^T$ be the unknown weight vector. Given the sign-bit filtered image $\hat{f}$ of I, we have that for every point p, excluding zero-crossings, the scalar product $\boldsymbol{\omega}^T \mathbf{z}(p)$ is either positive or negative. In this respect, points in $\hat{f}$ can be considered as "examples" in 3D space and the coefficients α_j as a vector normal to the separating hyperplane. Similarly, the reconstruction of the thresholded image $\hat{I}$ can be represented as a separating hyperplane problem in 4D space, in which $\mathbf{z}(p) = (I_1, I_2, I_3, -1)^T$ and $\boldsymbol{\omega} = (\alpha_1, \alpha_2, \alpha_3, k)^T$. □

The contours lead to a linear system of equations, whereas the sign-bits lead to a linear system of *inequalities*. The solution of a linear system of inequalities $A\mathbf{w} < \mathbf{b}$ can be approached using Linear Programming techniques or using Linear Discriminant Analysis techniques (see (Duda and Hart, 1973) for a review). Geometrically, the unknown weight vector $\mathbf{w}$ can be considered as the normal direction to a plane, passing through the origin, in 3D Euclidean space, and a solution is found in such a way that the plane separates the "positive" examples, $\boldsymbol{\omega}^T \mathbf{z}(p) > 0$, from the "negative" examples, $\boldsymbol{\omega}^T \mathbf{z}(p) < 0$. In the general case, where $\mathbf{b} \neq 0$, the solution is a point inside a polytope whose faces are planes in 3D space.

The most straightforward solution is known as the *perceptron* algorithm (Rosenblatt, 1962). The basic perceptron scheme proceeds by iteratively modifying the estimate of $\mathbf{w}$ by the following rule:

$$\mathbf{w}^{n+1} = \mathbf{w}^n + \sum_{i \in M} \mathbf{z}^i$$

where $\mathbf{w}^n$ is the current estimate of $\mathbf{w}$, and M is the set of examples $\mathbf{z}^i$ that are incorrectly classified by $\mathbf{w}^n$. The critical feature of this scheme that it is guaranteed to converge to a solution, irrespective of the initial guess $\mathbf{w}^0$, provided that a solution exists (examples are linearly separable). Another well known method is to reformulate the problem as a least squares optimization problem of the form

$$\min_{\mathbf{w}} |A\mathbf{w} - \mathbf{b}|^2$$

where the i'th row of A is $\mathbf{z}^i$, and $\mathbf{b}$ is a vector of arbitrarily specified positive constants (often $\mathbf{b} = 1$). The solution $\mathbf{w}$ can be found using the pseudoinverse of A, i.e.,

$$\mathbf{w} = A^+\mathbf{b} = (A^tA)^{-1}A^t\mathbf{b},$$

or iteratively through a gradient descent procedure, which is known as the Widrow-Hoff procedure. The least squares formulation is not guaranteed to find a correct solution but has the advantage of finding a solution even when a correct solution does not exist (a perceptron algorithm is not guaranteed to converge in that case).

By using the sign-bits instead of the contours, we are trading a unique, but unstable, solution for an approximate, but stable, solution. The stability of reconstruction from sign-bits is achieved by sampling points that are relatively far away from the contours. This sampling process also has the advantage of tolerating a certain degree of misalignment between the images as a result of less than perfect correspondence due to changes in viewing position (this feature is discussed further in Section 7.4). Experimental results (see Figs. 11 and 12) demonstrate that 10 to 20 points, distributed over the entire object, are sufficient to produce results that are comparable to those obtained from an exact solution. The experiments were done on images of 'Ken' and on another set of face images taken from a plaster bust of Roy Lamson (courtesy of the M.I.T Media Laboratory). Both the perceptron algorithm and the least-squares approach were implemented and both yielded practically the same results. The sample points were chosen manually, and over several trials we found that the reconstruction is not sensitive to the particular

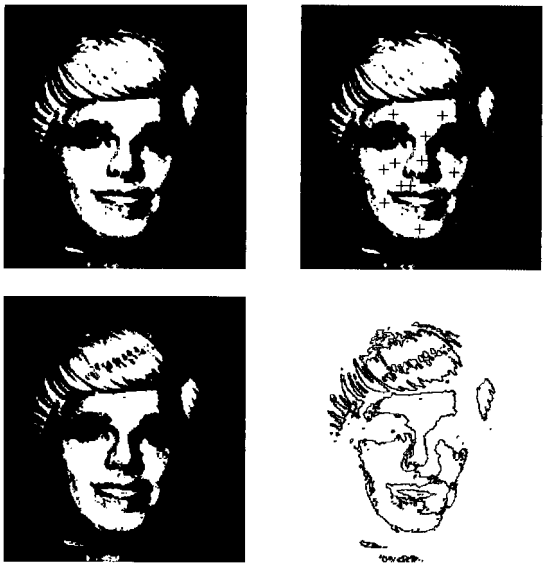

Figure 11. Reconstruction from sign-bits. *Top row* (*left to right*): the input novel image; the same image but with the sample points marked for display. *Bottom row*: the reconstructed image; the overlay of the original level-crossings and the level-crossings of the reconstructed thresholded image.

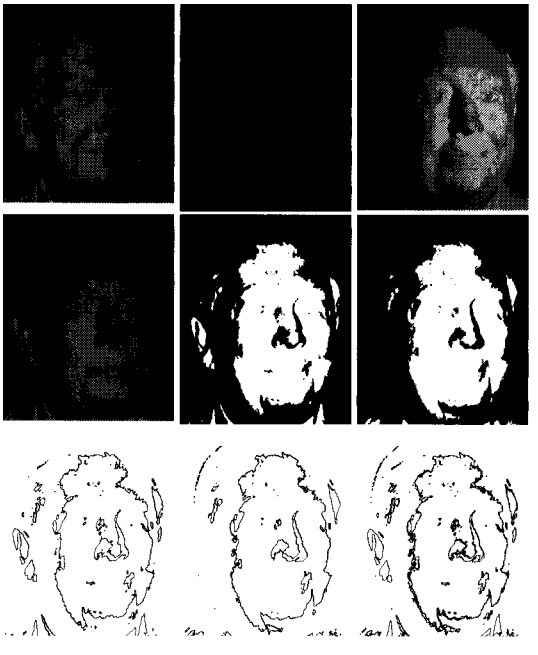

Figure 12. Reconstruction from sign-bits. *Row* 1: three model images. *Row* 2: novel image; thresholded input; reconstructed image (same procedure as described in the previous figure). Note that the left ear has not been reconstructed; this is mainly because the ear is occluded in two of the three model images. *Row* 3: the level-crossings of the novel input; level-crossings of the reconstructed image; the overlay of both level-crossing images.

choice of sample points, as long as they are not clustered in a local area of the image and are sampled a few pixels away from the contours. The results show the reconstruction of a novel thresholded images from three model images. The linear coefficients and the threshold are recovered from the system of inequalities using a sample of 16 points; the model images are then combined and thresholded with the recovered threshold to produce a synthesized thresholded image. Recognition then proceeds by matching the novel thresholded image given as input against the synthesized image.

6. The Geometric Source of Variability

So far, we have assumed that the object is viewed from a fixed viewing position, and allowed only photometric changes to occur. This restriction was convenient because that allowed us to combine the model images in a very simple manner. When changes in viewing positions are allowed to occur, the same image point across different projections does no longer correspond to the same object point. The simplest example is translation and rotation in the image plane which occur when the object translates, rotates around the line of sight, and then orthographically projects onto the image. This transformation can easily be undone if we observe two corresponding points between the novel input image and the model images. In general, however, the effects of changing viewing positions may not be straightforward as happens when the object rotates in depth and when the projection is perspective.

Because of the Lambertian assumption, we can treat the photometric and geometric sources of variability independently of each other. In other words, we have assumed photometric changes occurring in the absence of any geometric changes, and now we will assume that the different views are taken under identical illumination conditions. We can later combine the two sources of variability into a single framework which will allow us to compensate for both photometric and geometric changes occurring simultaneously.

The geometric source of variability raises two related issues. First, is establishing point-to-point correspondence between the model images. Second, given correspondence between the model images, undo the effects of viewing transformation based on a small number of corresponding points between the novel view and the model views.

The process of "undoing" the effects of changing viewing transformation between views is known as

the "alignment" approach in recognition. Given a 3D model, or at least two model views in full correspondence, one can "re-project" the object onto the novel viewing position with the help of a small number of corresponding points. Recognition is achieved if the re-projected image is successfully matched against the input image. We refer from hereafter to the problem of predicting a novel view from a set of model views using a limited number of corresponding points, as the problem of *re-projection*. The problem of finding a small number of corresponding points between two views is often referred to as *Minimal Correspondence* (Huttenlocher and Ullman, 1987).

The problem of establishing full correspondence between the model images requires not only undoing the effects of viewing transformation, but knowledge of object structure as well. Given two views, there is no finite number of corresponding points that would determine uniquely all other correspondences, unless the object is planar.

We discuss these issues briefly in the next section, restricting the discussion to the case of orthographic views. A more detailed treatment of these issues, including perspective views, can be found in Shashua (1991, 1992, 1994, 1995; Shashua and Navab, 1996; Shashua and Toelg, 1994).

6.1. Re-Projection and Correspondence

Let O, P_1, P_2, P_3 be four non-coplanar object points, referred to as reference points, and let O', P'_1, P'_2, P'_3 be the coordinates of the reference points from the second camera position. Let b_1, b_2, b_3 be the affine coordinates of an object point of interest P with respect to the basis OP_1, OP_2, OP_3, i.e.,

$$OP = \sum_{j=1}^{3} b_j (OP_j),$$

where the OP denotes the vector from O to P. Under parallel projection the viewing transformation between the two cameras can be represented by an arbitrary affine transformation, i.e., $O'P' = T(OP)$ for some linear transformation T. Therefore, the coordinates b_1, b_2, b_3 of P remain fixed under the viewing transformation, i.e.,

$$O'P' = \sum_{j=1}^{3} b_j (O'P'_j).$$

Since depth is lost under parallel projection, we have a similar relation in image coordinates:

$$op = \sum_{j=1}^{3} b_j (op_j) \tag{1}$$

$$o'p' = \sum_{j=1}^{3} b_j (o'p'_j). \tag{2}$$

Given the corresponding points p, p' (in image coordinates), the two formulas 1, 2 provide four equations for solving for the three affine coordinates associated with the object point P that projects to the points p, p'. Furthermore, since the affine coordinates are fixed for all viewing transformations, we can predict the location p'' on a novel view by first recovering the affine coordinates from the two model views and then substituting them in the following formula:

$$o''p'' = \sum_{j=1}^{3} b_j (o''p''_j).$$

We have, therefore, a method for recovering affine coordinates from two views and a method for achieving re-projection given two model views (in full correspondence) and four corresponding points across the three views.

Assume we would like to find the corresponding point p' given we know the correspondences due to the four reference points. It is clear that this cannot be done with the available information because a dimension is lost due to the projection from 3D to 2D (in fact, any number of corresponding points n would not be sufficient for determining the correspondence of the $n+1$ point (Aloimonos and Brown, 1989; Huang and Lee, 1989)). Since we do not have a sufficient number of observations to recover the affine coordinates, we look for an additional source of information.

We assume that both views are taken under similar illumination conditions:

$$I(x + \Delta x, y + \Delta y, t + \Delta t) = I(x, y, t),$$

where $\mathbf{v} = (\Delta x, \Delta y)$ is the displacement vector, i.e., $p' = p + \mathbf{v}$. We assume the convention that the two views were taken at times t and $t + \Delta t$. A first order approximation of a Taylor series expansion leads to the following equation which describes a linear approximation to the change of image grey-values at p

due to motion:

$$\nabla I \cdot \mathbf{v} + I_t = 0, \tag{3}$$

where ∇I is the gradient at point p, and I_t is the temporal derivative at p. Equation (3) is known as the "constant brightness equation" and was introduced by Horn and Schunk (1981). In addition to assuming that the change in grey-values is due entirely to motion, we have assumed that the motion (or the size of view separation) is small, and that the surface patch at P is locally smooth.

The constant brightness equation provides only one component of the displacement vector $\mathbf{v}$, the component along the gradient direction, or normal to the isobrightness contour at p. This "normal flow" information is sufficient to uniquely determine the affine coordinates b_j at p, as shown next. By subtracting Eq. (1) from Eq. (2) we get the following relation:

$$\mathbf{v} = \sum_{j=1}^{3} b_j \mathbf{v}_j + \left(1 - \sum_j b_j\right)\mathbf{v}_o, \tag{4}$$

where $\mathbf{v}_j$ ($j = 0, \ldots, 3$) are the known displacement vectors of the points o, p_1, p_2, p_3. By substituting Eq. (4) in the constant brightness equation, we get a new equation in which the affine coordinates are the only unknowns:

$$\sum_j b_j[\nabla I \cdot (v_j - v_o)] + I_t + \nabla I \cdot v_o = 0. \tag{5}$$

Equations (1) and (5) provide a complete set of linear equations to solve for the affine coordinates at all locations p that have a non-vanishing gradient, which is not perpendicular to the direction of the epipolar line passing through p'. Once the affine coordinates are recovered, the location of p' immediately follows. We have, therefore, derived a scheme for obtaining full correspondence given a small number of known correspondences, and a scheme for re-projecting the object onto any third view, given four corresponding points with the third view (the affine coordinates b_1, b_2, b_3 are view independent). Both schemes can be considerably simplified by expressing the problems in terms of one unknown per image point (instead of three) as follows.

Let A and $\mathbf{w}$ be the six affine parameters determined (uniquely) from the three corresponding vectors $op_j \leftrightarrow o'p'_j$ $j = 1, 2, 3$, i.e.,

$$o'p'_j = A(op_j) + \mathbf{w}, \qquad j = 1, 2, 3. \tag{6}$$

For an arbitrary pair of corresponding points p, p', we have the following relation:

$$o'p' = \sum_{j=1}^{3} b_j(o'p'_j) = \sum_{j=1}^{3} b_j(A(op_j) + \mathbf{w})$$
$$= A(op) + \left(\sum_j b_j\right)\mathbf{w},$$

or equivalently:

$$p' = [A(op) + o' + w] + \gamma_p \mathbf{w}. \tag{7}$$

where $\gamma_p = \sum b_j - 1$ is view-point invariant and is unknown. Note that $\gamma_p = 0$ if the object point P is coplanar with the plane $P_1P_2P_3$ (reference plane), and therefore γ_p represents the relative deviation of P from the reference plane (Shashua, 1991; Koenderink and Van Doorn, 1991). A convenient way to view this result is that the location of the corresponding point p' is determined by a "nominal component", described by $A(op) + o' + w$ and a "residual parallax component", described by $\gamma_p \mathbf{w}$. The nominal component is determined from the four known correspondences, and the residual component can be determined using the constant brightness Eq. (3) (for more details and discussion see (Shashua, 1991, 1992)):

$$\gamma_p = \frac{-I_t - \nabla I \cdot [A - I](op)}{\nabla I \cdot \mathbf{w}}.$$

The "affine depth" γ_p can be also used to simplify the re-projection scheme onto a third view: since γ_p is invariant, then it can be computed from the correspondence between the two model views and substituted in the equation describing the epipolar relation between the first and third (novel) view. The re-projection scheme can be further simplified (Shashua, 1992) to yield the "linear combination of views" of (Ullman and Basri, 1989) (also (Poggio, 1990)):

$$x'' = \alpha_1 x' + \alpha_2 x + \alpha_3 y + \alpha_4, \tag{8}$$
$$y'' = \beta_1 y' + \beta_2 x + \beta_3 y + \beta_4, \tag{9}$$

where the coefficients α_j, β_j are functions of the affine viewing transformations between the three views. The coefficients can be recovered from four corresponding

points across the three views, and then used to generate p'' for every corresponding pair $p \leftrightarrow p'$.

These techniques extend to the general case of perspective views. Instead of affine coordinates one can recover projective coordinates from two views and eight corresponding points (Faugeras, 1992; Hartley et al., 1992; Shashua, 1994). The affine depth invariant γ_p turns into a projective invariant ("relative affine structure") (Shashua and Navab, 1996). Finally, the linear combination of views result of Ullman and Basri (1989) turns into a trilinear relation requiring seven matching points in general (instead of four), or bilinear in case only the model views are orthographic—requiring five matching points (Shashua, 1995).

7. Combining Changes in Illumination with Changes in Viewing Positions: Experimental Results

We have described so far three components that are necessary building blocks for dealing with recognition via alignment under the geometric and photometric sources of variability. First, is the component describing the photometric relation between three model images and a novel image of the object. Second, is the component describing the geometric relation between two model views and a novel view of an object of interest. Third, is the correspondence component with which it becomes possible to represent objects by a small number of model images. The geometric and photometric components were treated independently of each other. In other words, the photometric problem assumed the surface is viewed from a fixed viewing position. The geometric problem assumed that the views are taken under a fixed illumination condition, i.e., the displacement of feature points across the different views is due entirely to a change of viewing position. In practice, the visual system must confront both sources of variability at the same time. The combined geometric and photometric problem is defined below:

We assume we are given three model images of a 3*D matte object taken under different viewing positions and illumination conditions. For any input image, determine whether the image can be produced by the object from some viewing position and by some illumination condition.*

The combined problem definition suggests that the problem be solved in two stages: first, changes in viewing positions are compensated for, such that the three model images are aligned with the novel input image. Second, changes of illumination are subsequently compensated for, by using the photometric alignment method. In the following sections we describe several experiments with 'Ken' images starting from the procedure that was used for creating the model images, followed by three recognition situations: (i) the novel input image is represented by its grey-levels, (ii) the input representation consists of sign-bits, and (iii) the input representation consists of grey-levels, but the model images are taken from a fixed viewing position (different from the viewing position of the novel image). In this case we make use of the sign-bits in order to achieve photometric alignment although the novel image is taken from a different viewing position.

7.1. Creating a Model of the Object

The combined recognition problem implies that the model images represent both sources of variability, i.e., be taken from at least two distinct viewing positions and from three distinct illumination conditions. The three model images displayed in the top row of Fig. 13 were taken under three distinct illumination conditions, and from two distinct viewing positions (23° apart, mainly around the vertical axis). In order to apply the correspondence method described in the previous section, we took an additional image in the following way. Let the three illumination conditions be denoted by the symbols S_1, S_2, S_3, and the two viewing positions be denoted by V_1, V_2. The three model images, from left to right, can be described by $\langle V_1, S_1\rangle$, $\langle V_2, S_2\rangle$ and $\langle V_1, S_3\rangle$, respectively. Since the first and third model images are taken from the same viewing position, the two images are already aligned. In order to achieve full correspondence between the first two model images, a fourth image $\langle V_2, S_1\rangle$ was taken. Correspondence between $\langle V_1, S_1\rangle$ and $\langle V_2, S_1\rangle$ was achieved via the correspondence method described in the previous section. Since $\langle V_2, S_1\rangle$ and $\langle V_2, S_2\rangle$ are from the same viewing position, then the correspondence achieved previously holds also between the first and second model images. The fourth image $\langle V_2, S_1\rangle$ was then discarded and did not participate in subsequent recognition experiments.

7.2. Recognition from Grey-Level Images

The method for achieving recognition under both sources of variability is divided into two stages: first,

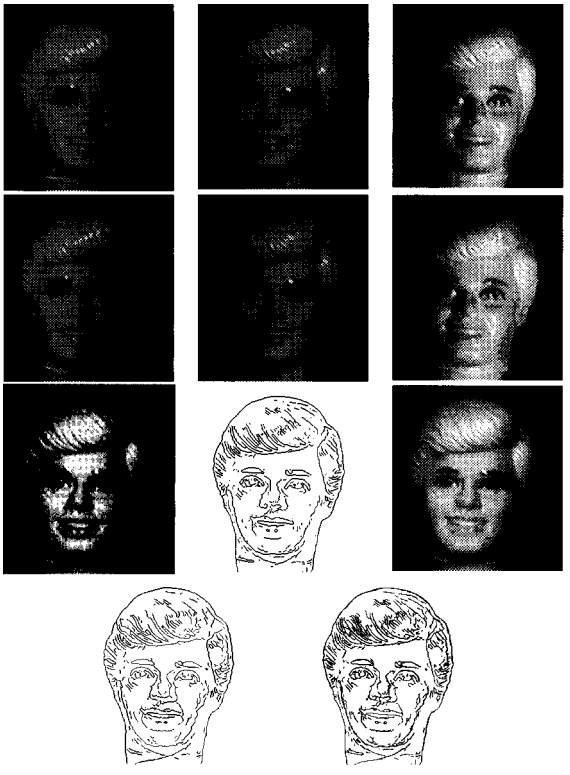

Figure 13. Recognition from full grey-level novel image (see text for more detailed description). *Row* 1 (*left to right*): three model images (the novel image is shown third row lefthand display). *Row* 2: view-compensated model images—all three model images are transformed (using four points) as if viewed from the novel viewing position. *Row* 3: novel image, edges of novel image, photometric alignment of the three view-compensated model images (both view and illumination compensated). *Row* 4: edges of the resulting synthesized image (third row righthand), overlay of edges of novel and synthesized image.

the three model images are re-projected onto the novel image. This is achieved by first assuming minimal correspondence between the novel image and one of the model images. With minimal correspondence of four points across the images (model and novel) we can predict the new locations of model points that should match with the novel image (assuming orthographic projection). Second, photometric alignment is subsequently applied by selecting a number of points (no correspondence is needed at this stage because all images are now view-compensated) to solve for the linear coefficients. The three model images are then linearly combined to produce a synthetic image that is both view and illumination compensated, i.e., should match the novel image.

Figure 13 illustrates the chain of alignment transformations. The novel image, displayed in the third row left image, is taken from an in-between viewing position and illumination condition. Although, in principle, the recognition components are not limited to in-between situations, there are few practical limitations. The more extrapolated the viewing position is, the more new object points appear and old object points disappear, and similarly, the more extrapolated the illumination condition is, the more new cast shadows are created (see Section 4.1). Minimal correspondence was achieved by manually selecting four points that corresponded to the far corners of the eyes, one eyebrow corner, and one mouth corner. The model views were re-projected onto the novel view, and their original grey-values retained. As a result, we have created three synthesized model images (shown in Fig. 13, second row) that are from the same viewing position as the novel image, but have different image intensity distributions due to changing illumination. The photometric alignment method was then applied to the three synthesized model images and the novel image, without having to deal with correspondence because all four images were already aligned. The sample points for the photometric alignment method were chosen automatically by searching over smooth regions of image intensity (as described in Section 4.3). The resulting synthesized image is displayed in Fig. 13, third row right image. The similarity between the novel and the synthesized image is illustrated by superimposing the step edges of the two images (Fig. 13, bottom row right image).

7.3. Recognition from Reduced Images

A similar procedure to the one described above can be applied to recognize a reduced novel image. In this case the input image is taken from a novel viewing position and illumination condition, followed by a thresholding operator (unknown to the recognition system). Figure 14 illustrates the procedure. We applied the linear combination method of re-projection (Ullman and Basri, 1989) and used more than the minimum required four points. In this case it is more difficult to reliably extract corresponding points between the thresholded input and the model images. Therefore, seven points were manually selected and their corresponding points were manually estimated in the model images. The linear combination method was then applied using a least squares solution for the linear coefficients to

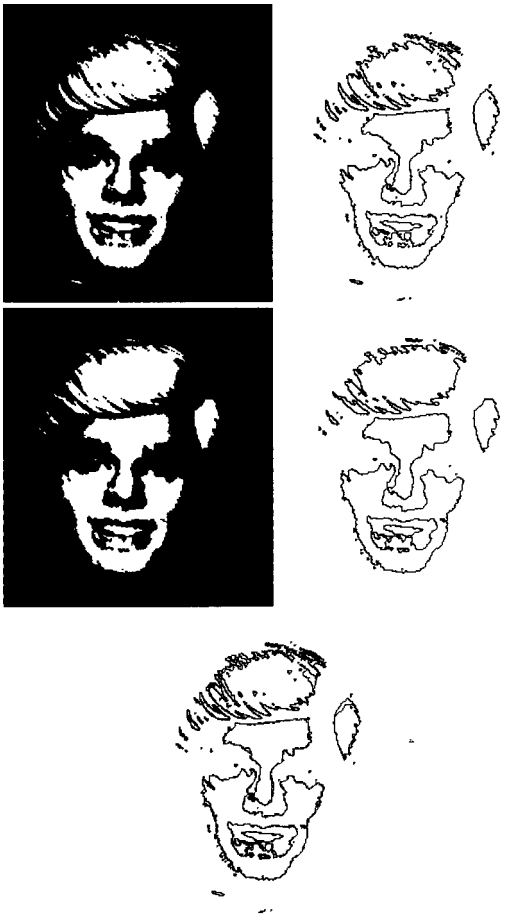

Figure 14. Recognition from a reduced image. *Row* 1 (*left to right*): novel thresholded image; its level-crossings (the original grey-levels of the novel image are shown in the previous figure, third row on the left). *Row* 2: the synthesized image produced by the recognition procedure; its level-crossings. *Row* 3: overlay of both level-crossings for purposes of verifying the match.

produce three synthesized view-compensated model images. The photometric alignment method from sign-bits was then applied (Section 5.2) using a similar distribution of sample points as shown in Fig. 11.

We consider next another case of recognition from reduced images, in which we make use of the property that exact alignment is not required when using sign-bits.

7.4. Recognition from a Single Viewing Position

Photometric alignment from sign-bits raises the possibility of compensating for changing illumination without an exact correspondence between the model images and the novel image. The reason lies in the way points are sampled for setting the system of inequalities; that is, points are sampled relatively far away from the contours (see Section 5.2). In addition, the separation of image displacements into nominal and residual components (Section 6.1) suggests that in an area of interest bounded by at least three reference points, the nominal component alone may be sufficient to bring the model images close enough to the novel image so that we can apply the photometric alignment from sign bits method.

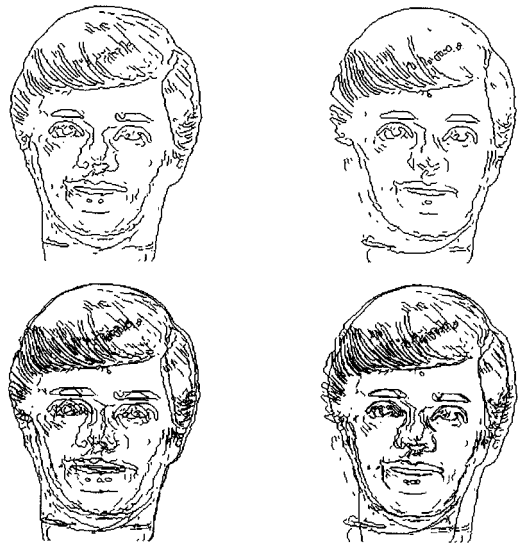

Figure 15. Demonstrating the effect of applying only the nominal transformation between two distinct views. *Row* 1: edges of two distinct views. *Row* 2: overlay of both edge image, and overlay of the edges of the left image above and the nominally transformed righthand image.

Consider, for example, the effect of applying only the nominal transformation between two different views (Fig. 15). Superimposing the two views demonstrates that the displacement is concentrated mostly in the center area of the face (most likely the area in which we would like to select the sample points). By selecting three corresponding points covering the center area of the face (two extreme eye corners and one mouth corner), the 2D affine transformation (nominal transformation) accounts for most of the displacement in the area of interest at the expense of large displacements at the boundaries (Fig. 15, bottom row on the right). This is expected from the geometric interpretation of affine depth γ_p, as it increases as the object gets farther from the reference plane (Koenderink and Van Doorn, 1991; Shashua, 1991).

Taken together, the use of sign-bits and the nominal transformation suggests that one can compensate for illumination and for relatively small changes in viewing positions from model images taken from the same viewing position. We apply first the nominal transformation to all three model images and obtain three synthesized images. We then apply the photometric alignment from sign-bits to recover the linear coefficients used for compensating for illumination. The three synthesized images are then linearly combined to obtain an illumination-compensated image. The remaining displacement between the synthesized image and the novel image can be recovered by applying the residual motion transformation (along the epipolar direction using the constant brightness equation).

Figure 16 illustrates the alignment steps. The three model images are displayed in the top row and are the same as those used in Section 4.3 for compensating for illumination alone. The novel image (second row, left display) is the same as in Fig. 13, i.e., it is taken from a novel viewing position and novel illumination condition. The image in the center of the second row illustrates the result of attempting to recover the correspondence (using the full correspondence method described in the previous section) between the novel image and one of the model images without first compensating for illumination. The image on the left in the third row is the result of first applying the nominal transformation to the three model images followed by the photometric alignment using the sign-bits (the sample points used by the photometric alignment method are displayed in the image on the right in the second row). The remaining residual displacement between the latter image and the novel image is recovered using the full correspondence method and the result is displayed in the center image in the third row. The similarity between the final synthesized image and the novel image is illustrated by superimposing their step edges (fourth row, right display).

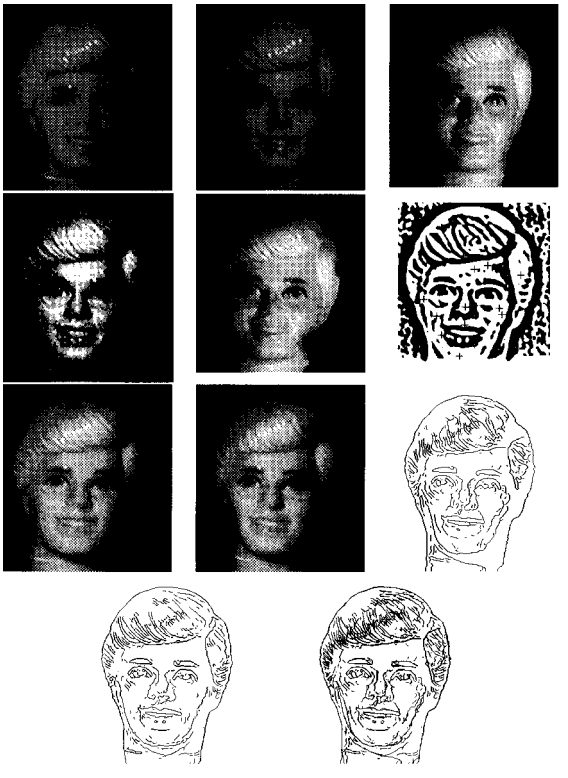

Figure 16. Recognition from a single viewing position (see text for details).

8. Summary and Discussion

In this paper we addressed the connection between recognition of general 3D objects and the ability to create an equivalence class of images of the same object. Recognizing objects eventually reduces to comparing/matching images against each other or against models of objects. This can be viewed as comparing measurements (features, x, y positions of points, and so forth) that ideally must be selected or manipulated such that they remain invariant if coming from images of the same object.

We have distinguished two sources of variability against which invariant measurements are needed. One, is the well known geometric problem of changing viewing positions between the camera and the object. Second, is the photometric problem due to changing illumination conditions in the scene. Our emphasis in this paper was on the latter problem which, unlike the geometric problem of recognition, did not receive much attention in the past. The traditional assumption concerning the photometric problem is that one can recover a reasonably complete array of invariants just from a single image alone, such as the representations produced by edge detectors. It is interesting to note that in the earlier days of recognition, the geometric problem was approached in a similar manner by restricting the class of objects to polyhedra (the so-called "blocks world"). We have argued that complete invariance of edges is achieved when simple block-like objects are considered, whereas for more natural and complex objects, like a face, it may be necessary to explore

model-based approaches, i.e., the photometric invariants are model-dependent.

Motivated by several empirical observations on human vision, and by computational arguments on the use of edge detection, we arrived at two observations: first, there appears to be a need for a model-based approach to the photometric problem. Second, the process responsible for factoring out the illumination during the recognition process appears to require more than contour information, but just slightly more.

We suggested a method, we call photometric alignment, that is based on recording multiple images of the object. We do not use these pre-recorded images to recover intrinsic properties of the object, as in photometric stereo, but rather to directly compensate for the change in illumination conditions for any other novel image of the object. This difference is critical, as we are no longer bounded by assumptions on the light source parameters (e.g., we do not need to recover light source directions) or assumptions on the distribution of surface albedo (e.g., arbitrary distributions of surface albedo are allowed). We have discussed the situations of shadows, specular reflections, and changing spectral compositions of light sources. In the case of shadows, we have seen that the alignment scheme degrades with increasing cast shadow regions in the novel input image. As a result, photometric alignment, when applied to general non-convex surfaces, is most suitable for reconstructing novel images whose illumination conditions are in between those used to create the model images. We have also seen that specular reflections arising from non-homogeneous surfaces can be detected and removed if necessary. Finally, the theory was extended to deal with color images and the case of changing spectral composition of light sources: the of color bands of a single model image of a neutral surface can form a basis set for reconstructing novel images.

We next introduced two new results to explore the possibility of working with reduced representations instead of image grey values—as suggested by empirical evidence from human vision (Section 2). First, step edges and level-crossings of the novel image are theoretically sufficient for the photometric alignment scheme. This result, however, assumes that edges be given at sub-pixel accuracy—a finding that implies difficulties in making use of this result in practice. Second, the sign-bit information can be used instead of edges. Photometric alignment using sign-bits is a region-based process by which points inside the binary regions of the sign-bit image are sampled and each contributes a partial observation. Taken together, the partial observations are sufficient to determine the solution for compensating for illumination. The more points sampled, the more accurate the solution. Experimental results show that a relatively small number of points (10 to 20) are generally sufficient for obtaining solutions that are comparable to those obtained by using the image grey values. This method agrees with the empirical observations that were made in Section 2 regarding the possibility of having a region-based process rather than a contour-based one, the possibility of preferring sign-bits over edges, and the sufficiency of sign-bits for factoring out the illumination. The possibility of using sign-bits instead of edges raises also a potentially practical issue related to changing viewing positions. A region-based computation has the advantage of tolerating a small degree of misalignment between the images due to changing viewing positions. This finding implies that the illumination can be factored out even in the presence of small changes in viewing positions without explicitly addressing the geometric problem of compensating for viewing transformations—a property that was demonstrated experimentally in Section 7.4.

Finally, we have shown how the geometric, photometric, and the correspondence components can be put together to address the case when both changes in illumination and changes in viewing positions occur simultaneously, i.e., recognition of an image of a familiar object taken from a novel viewing position and a novel illumination condition.

8.1. Issues of Future Directions

The ability to interpret Mooney images of faces may suggest that these images are an extreme case of a wider phenomenon. Some see it as a tribute to the human ability to separate shadow borders from object borders (Cavanagh, 1990); here we have noted that the phenomenon may indicate that in some cases illumination is factored out in a model-based manner and that the process responsible apparently requires more than just contour information, but only slightly more. A possible topic of future research in this domain would be to draw a connection, both at the psychophysical and computational levels, between Mooney images and more natural kinds of inputs. For example, images seen in newspapers, images taken under poor lighting, and other low

quality imagery have less shading information to rely on and their edge information may be highly unreliable, yet are interpreted without much difficulty by the human visual system. Another related example, is the image information contained in draftsmen's drawings. Artists rarely use just contours in their drawings and rely on techniques such as "double stroking" to create a sense of relief (surface recedes towards the contours) and highlights to make the surface protrude. These pictorial additions that artists introduce are generally not interpretable at the level of contours alone, yet do not introduce any direct shading information. In other words, it would be interesting (and probably important on practical grounds) to discover a continuous transformation, a spectrum or scale-space of sorts, starting from high-quality grey-level imagery, producing midway low-quality imagery of the type mentioned above, and converging upon Mooney-type imagery.

Another related topic of future interest is the level at which sources of variability are compensated for. In this paper the geometric and photometric sources of variability were factored out based on connections between different images of individual objects. The empirical observations we used to support the argument that illumination should be compensated for in a model-based manner, actually indicate that if indeed such a process exists, it is likely to take place at the level of classifying the image as belonging to a general class of objects, rather than at the level of identifying the individual object. This is simply because the Mooney images are of generally unfamiliar faces, and therefore, the only model-based information available is that we are looking at an image of a face. A similar situation may exist in the geometric domain as well, as it is known that humans can recognize novel views just from a single view of the object.

There are also questions of narrower scope related to the photometric domain that may be of general interest. The question of image representation in this paper was applied only to the novel image. A more general question should apply to the model acquisition stage as well. In other words, what information needs to be extracted from the model images, at the time of model acquisition, in order to later compensate for photometric effects? This question applies to both the psychophysical and computational aspects of the problem. For example, can we learn to generalize to novel images just from observing many Mooney-type images of the object? (changing illumination, viewing positions, threshold, and so forth). A more basic question is whether the Mooney phenomenon is limited exclusively to faces. And if not, what level of familiarity with the object, or class of objects, is necessary in order to generalize to other Mooney-type images of the same object, or class of objects.

At a more technical level, there may be interest in further pursuing the use of sign-bits. The sign-bits were used as a source of partial observations that, taken together, can restrict sufficiently well the space of possible solutions for the photometric alignment scheme. In order to make further use of this idea, and perhaps apply it to other domains, the question of how to select sample points, and the number and distribution of sample points, should be addressed in a more systematic manner.

Acknowledgments

I thank my advisor Shimon Ullman for allowing me to benifit from his knowledge, experience and financial support during my doctoral studies. Thanks to William T. Freeman for producing the contour images in Fig. 4.

References

Aloimonos, J. and Brown, C.M. 1989. On the kinetic depth effect. *Biological Cybernetics*, 60:445–455.

Arnheim, R. 1954. *Art and Visual Perception: A Psychology of the Creative Eye*, University of California Press: Berkeley, CA.

Cavanagh, P. 1990. What's up in top-down processing? In *Proceedings of the XIIIth ECVP*, A. Gorea (Ed.), pp. 1–10.

Coleman, E.N., Jr. and Jain, R. 1982. Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry. *Computer Graphics and Image Processing*, 18:309–328.

Curtis, S.R., Oppenheim, A.V., and Lim, J.S. 1985. Signal reconstruction from Fourier transform sign information. *IEEE Transactions on Acoustic Speech and Signal Processing*, ASSP-33:643–657.

Duda, R.O. and Hart, P.E. 1973. *Pattern Classification and Scene Analysis*. John Wiley: New York.

Farah, M.J. 1990. *Visual Agnosia*. MIT Press: Cambridge, MA.

Faugeras, O.D. 1992. What can be seen in three dimensions with an uncalibrated stereo rig? In *Proceedings of the European Conference on Computer Vision*, Santa Margherita Ligure, Italy, pp. 563–578.

Fischler, M.A. and Bolles, R.C. 1981. Random sample concensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381–395.

Freeman, W.T. and Adelson, E.H. 1991. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13.

Gilchrist, A.L. 1979. The perception of surface blacks and whites. *SIAM J. Comp.*, pp. 112–124.

Hartley, R., Gupta, R., and Chang, T. 1992. Stereo from uncalibrated cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Champaign, IL, pp. 761–764.

Hille, E. 1962. *Analytic Function Theory*. Ginn/Blaisdell: Waltham, MA.

Horn, B.K.P. 1986. *Robot Vision*. MIT Press: Cambridge, Mass.

Horn, B.K.P. and Schunk, B.G. 1981. Determining optical flow. *Artificial Intelligence*, 17:185–203.

Huang, T.S. and Lee, C.H. 1989. Motion and structure from orthographic projections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11:536–540.

Huttenlocher, D.P. and Ullman, S. 1987. Object recognition using alignment. In *Proceedings of the International Conference on Computer Vision*, London, pp. 102–111.

Huttenlocher, D.P. and Ullman, S. 1990. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212.

Kinsbourne, M. and Warrington, E.K. 1962. A disorder of simultaneous form perception. *Brain*, 85:461–486.

Klinker, G.J., Shafer, S.A., and Kanade, T. 1990. A physical approach to color image understanding. *International Journal of Computer Vision*, 4:7–38.

Knill, D.C. and Kersten, D. 1991. Apparent surface curvature affects lightness perception. *Nature*, 351:228–230.

Koenderink, J.J. and Van Doorn, A.J. 1980. Photometric invariants related to solid shape. *Optica Acta*, 27:981–986.

Koenderink, J.J. and Van Doorn, A.J. 1991. Affine structure from motion. *Journal of the Optical Society of America*, 8:377–385.

Land, E.H. and McCann, J.J. 1971. Lightness and retinex theory. *Journal of the Optical Society of America*, 61:1–11.

Logan, B. 1977. Information in the zero-crossings of band pass signals. *Bell Syst. Tech. J.*, 56:510.

Lowe, D.G. 1985. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishing: Hingham, MA.

Maloney, L.T. and Wandell, B. 1986. A computational model of color constancy. *Journal of the Optical Society of America*, 1:29–33.

Marr, D. 1979. Early processing of visual information. *Proceedings of the Royal Society of London B*, 175:483–534.

Marr, D. 1982. *Vision*. W.H. Freeman and Company: San Francisco, CA.

Marr, D. and Hildreth, E.C. 1980. Theory of edge detection. *Proceedings of the Royal Society of London B*, 207:187–217.

Mooney, C.M. 1960. Recognition of ambiguous and unambiguous visual configurations with short and longer expsures. *Brit. J. Psychol.*, 51:119–125.

Morrone, M.C. and Burr, D.C. 1988. Feature detection in human vision: A phase-dependent energy model. *Proceedings of the Royal Society of London B*, 235:221–245.

Moses,Y. 1993. Face recognition: Generalization to novel images. Ph.D. Thesis, Applied Math. and Computer Science, The Weizmann Institute to Science, Israel.

Mundy, J. and Zisserman, A. 1992. *Geometric Invariances in Computer Vision*, MIT Press: Cambridge.

Mundy, J.L., Zisserman, A., and Forsyth, D. 1994. *Applications of Invariance in Computer Vision*, Springer-Verlag: LNCS 825.

Nakayama, K., Shimojo, S., and Silverman, G.H. 1989. Stereoscopic depth: Its relation to image segmentation, grouping, and the recognition of occluded objects, *Perception*, 18:55–68.

Nishihara, H.K. 1984. Practical real-time imaging stereo matcher. *Optical Engineering*, 23(5):536–545.

Pearson, D.E., Hanna, E., and Martinez, K. 1986. Computer-generated cartoons. In *Images and Understanding*, H. Barlow, C. Blakemore, and M. Weston-Smith (Eds.), Cambridge University Press: New York, NY, 1990. Collection based on Rank Prize Funds' International Symposium, Royal Society.

Perona, P. and Malik, J. 1990. Detecting an localizing edges composed of steps, peaks and roofs. In *Proceedings of the International Conference on Computer Vision*, Osaka, Japan.

Poggio, T. 1990. 3D object recognition: On a result of Basri and Ullman. Technical Report IRST 9005-03.

Potter. M.C. 1975. Meaning in visual search. *Science*, 187:965–966.

Rosenblatt, F. 1962. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books: Washington, DC.

Saff, E.B. and Snider, A.D. 1976. *Fundamentals of Complex Analysis*. Prentice-Hall: New-Jersey.

Sanz, J. and Huang, T. 1989. Image representations by sign information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11:729–738.

Shafer, S.A. 1985. Using color to separate reflection components. *COLOR Research and Applications*, 10:210–218.

Shashua, A. 1991. Correspondence and affine shape from two orthographic views: Motion and recognition. A.I. Memo No. 1327, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Shashua, A. 1992. Geometry and photometry in 3D visual recognition. Ph.D. Thesis, M.I.T. Artificial Intelligence Laboratory, AI-TR-1401.

Shashua, A. 1994. Projective structure from uncalibrated images: Structure from motion and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):778–790.

Shashua, A. 1995. Algebraic functions for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):779–789.

Shashua, A. and Navab, N. 1996. Relative affine structure: Canonical model for 3D from 2D geometry and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9).

Shashua, A. and Toelg, S. 1994. The quadric references surface: Applications in registering views of complex 3D objects. In *Proceedings of the European Conference on Computer Vision*, Stockholm, Sweden.

Shashua, A. and Ullman, S. 1988. Structural saliency: The detection of globally salient structures using a locally connected network. In *Proceedings of the International Conference on Computer Vision*, Tampa, FL, pp. 321–327.

Shashua, A. and Ullman, S. 1991. Grouping contours iterated pairing network. In *Advances in Neural Information Processing Systems 3*, R.P. Lippmann, J.E. Moody, and D.S. Touretzky (Eds.), Morgan Kaufmann Publishers: San Mateo, CA, pp. 335–341. *Proceedings of the Third Annual Conference NIPS*, Dec. 1990, Denver, CO.

Ullman. S. 1986. Aligning pictorial descriptions: An approach to object recognition. *Cognition*, 32:193–254, 1989. Also in MIT AI Memo 931.

Ullman S. and Basri. R. 1989. Recognition by linear combination of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13:992–1006, 1991. Also in M.I.T. AI Memo 1052.

R.J. Woodham. 1980. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19:139–144.