In CMU Computer Science: a 25th Anniversary Commemorative, ACM Press, 1991.

14

Computer Vision as a Physical Science

Takeo Kanade

14.1 Introduction

Vision is one of the most important perceptual capabilities that any autonomous intelligent system, either natural or artificial, *can* possess in order to operate in the real world. Computer vision encompasses the development of both the computational theories and the technological means to realize artificial vision systems with performance equal to or greater than that of humans.

The goal of computer vision turns Out to be extremely difficult. Some of the difficulties are technological, such as the requirements for huge amounts of processing power, memory, and communication bandwidth. Other difficulties are more fundamental. A large number of factors, such as object shape, illumination, surface properties, sensor characteristics, and more, all contribute to determining the color and intensity of image pixels; the effects of any single factor are confounded by the effects of other factors. Consequently, many early vision problems of recovering scene properties (such as shape) from images are underconstrained. or ill posed, meaning that the images alone do not contain enough information to uniquely solve them. Therefore correctly interpreting images and constructing descriptions from them requires additional constraints and knowledge.

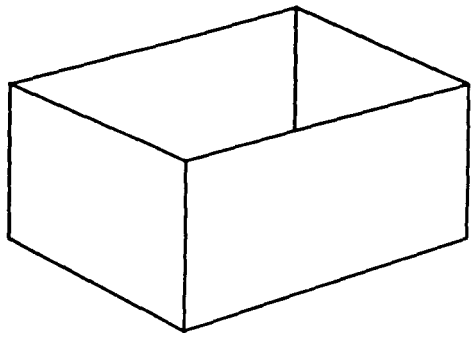
Yet, we humans seem to do very well at interpreting images. Given a two-

dimensional (2D) image we easily determine the correct interpretation of the three-dimensional (3D) relationships among objects in the *scene*. We identify objects consistently over a wide range of viewpoints and lighting conditions. We determine surface properties such as roughness and reflectance from images. We effectively utilize vision to map the world around us, enabling us to move without collisions or to grasp objects accurately. Human vision is an existence proof of a most powerful vision machine, which *can* deal very robustly with many difficult vision problems, such as distortions by projection, motion, stereo, texture, shading, and color. In doing so, humans do not seem to use much knowledge about the physical processes underlying those problems. In fact, most people know very little about optics, geometry, and physics. Moreover, when given images from exotic sensors, such as synthetic aperture radars (SAR). scanning electron microscopes (SEM), and forward-looking infrared (FLIR) sensors, humans *can* often interpret them correctly without asking much about how the images were created.

Historically, the fact that human vision provides a most compelling reference model and yet does not seem to rely on the knowledge of physical aspects of vision led many vision researchers to rush out and attempt to invent vision "algorithms" or build vision "systems" without first determining the information that images actually contain. Attention naturally focused on phenomenological performance, since it appeared that the underlying physical phenomena were too complicated to model, that images were too noisy for reliable algorithmic feature extraction, and that humans seemed to resolve such difficulties by using empirical domain-specific knowledge. Consequently, this approach was inevitably heuristic, since the major source of ideas was introspection or analogy from mechanisms that natural systems might use. The results from this approach were hard to characterize and to generalize. Basic vision problems were neither identified nor addressed.

However, attention has recently turned to putting the geometrical, physical, and optical processes underlying vision into a quantitative, computational framework. Now the emphasis is on developing physical models for computer vision. Such modeling reveals the structure of visual information: the exact information that is contained in an image, the limits of processing algorithms, and the heuristic knowledge required to resolve any remaining ambiguity. Thus algorithms derived from physical modeling are far more powerful and quantitative, and their performance far more predictable and generalizable than previous ad hoc methods based solely on heuristics. In fact. one of the most exciting discoveries in recent computer vision research is that natural generic constraints are often sufficientto solve many fundamental vision problems, some of which had been thought impossible to solve without applying heuristics.

In the last 10 years, the vision group of Carnegie Mellon University (CM
has been spearheading the development of a systematic theory for vision
on physical knowledge. This theory. which I refer to as physically based



A "Box" line **drawing**

EXHIBIT 14.1
A "box' line drawing.

emphasizes the use of knowledge about geometry, physics, optics, and statistics to model and solve basic vision problems. It is appropriate on the occasion of the CMU Computer Science 25th Anniversary to highlight our contributions to this new approach to vision. This chapter will illustrate physically based vision by using examples that my colleagues and I have developed here at CMU. These examples are drawn from three areas in computer vision: the determination of 3D shape from images; the analysis of object color and surface reflection, and uncertainty in visual measurements. In each area, I will use a (seemingly) simple problem as the background, then present our solution, and then give a broader perspective in that area

14.2 Geometry and Shape Constraints

14.2.1 A Problem: Interpreting Line Drawings

As the first example, consider the simple line drawing shown in Exhibit 14.1. What shape does this represent? Most people would be quick to say that this is a line drawing of a box with no lid. When asked about the reason for that interpretation, they would say, "I learned it over the years," or 'That shape is

the most familiar to me." Though not incorrect, these answers simply duck the questions about computational aspects of vision.

The line drawing is 2D, and the interpretation of it is a 3D shape. In general, many different shapes *can* give rise to the same line drawing. Therefore the process of interpretation must resolve ambiguity. To reach a single interpretation, some constraints about possible interpretations must have been used. Moreover, these constraints must be very strong: not only do people tend to agree on a single interpretation, but some people find *the* possibility of multiple interpretations difficult to accept. The line drawing does not have any shading or color, so the constraints must be geometrical in nature. The natural question to ask is, "How far does geometry constrain the interpretation of the line drawing?"

14.29 The Origami World

The study of computer interpretation of line drawings as three-dimensional scenes has captured interest in computer vision from the beginning. Guzman [1] wrote a program to segment line drawings into objects based on a collection of heuristic rules on the "strengths" of links between regions. Huffman [3] discovered a mathematical way to capture the geometrical constraints of a solid "trihedral" world by using labels *that* represent physical meanings of lines. Waltz [32] extended the idea to include shadows as well as devising an efficient procedure for labeling. However, the problem of multiplicity of interpretations was not addressed. Moreover, the Huffman-Waltz labeling could not handle the simple line drawing of Exhibit 14.1: it is classified as "impossible." I developed a theory of the Origami World [4] to begin to answer these questions.

Imagine a world that consists entirely of planar surfaces, which may be folded, cut, or glued together only along straight lines. This world is named the "Origami World." In the Origami World. we can develop a mathematical algorithm, which, given a line drawing like Exhibit 14.1, specifies all the shapes that can generate the given picture.

In the Origami World, Exhibit 14.1 could be any one of eight different shapes. Two of them *are* shown in Exhibit 14.2 by using special symbols to represent shapes. The interpretation on the left represents a "normal" box like the one we tend to consider. The interpretation on the right, however, represents another shape, which does not look like a "normal" box, but *can* actually generate the same picture. Moreover, the labelings shown in Exhibit 14.2 actually specify only the qualitative nature of the shape. For example, in the "normal" box interpretation, humans think of only a rectangular box where the front walls of the box meet at a right angle. However, any angle between 0° and 90° is in fact possible and the resultant shape projects onto the same line drawing, if other parts of the shape vary accordingly. Likewise, each interpretation in Exhibit 14.2 actually represents a continuous family of possible 3D shapes that

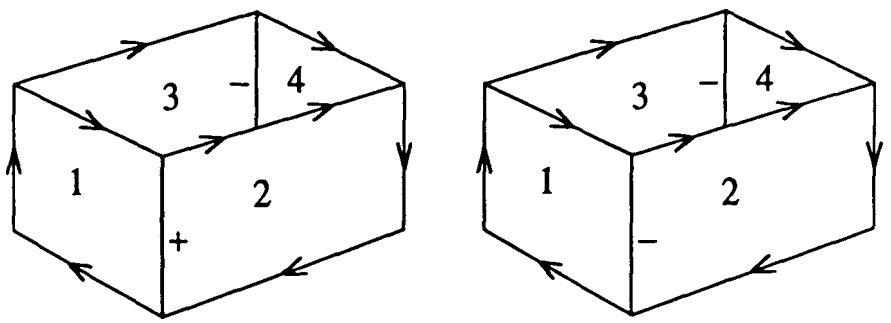


EXHIBIT 14.2
"Box" interpretations. Two are shown here by assigning Huffman's labels: +, −, and ↑ to each line. The labels signify the physical meaning of the lines. The labels + and − stand for a convex edge, that is, the two surfaces meet there and form convexity or concavity. respectively, when seen from the current viewing direction. The label ↑ stands for an occluding edge. That means that the region to its right, when standing in the direction of the arrow, occludes the region to its left. The interpretation on the left corresponds to a "normal" box, where surfaces 1 and 2 form convexity and occlude surfaces 3 and 4 which form concavity. The interpretation on the right, however, represents a "squashed" shape, since surfaces 1 and 2 also form a concavity.

can generate the same line drawing. Metaphorically speaking, there *are* 8 x ∞ interpretations of Exhibit **14.1** in the Origami World. It should be remembered that the real world is larger than the Origami World, and thus there are even more possibilities since the real world is not limited to planes and straight edges.

14.2.3 Principle of Nonaccidental Regularities

Why, then, do we tend to consider only a single interpretation. a so-called rectangular box shape? To say "the shape is *more* familiar" does not really answer the question, since most of us, in fact, *cannot* think of multiple interpretations. We do not select a particular interpretation after we think of all the possibilities. Rather we think of *only* the rectangular shape. Thus geometrically *speaking,* additional shape constraints must be used in a relatively early stage in order lo reach the particular interpretation. One interesting class of constraints can be obtained from the principle of nonaccidental regularities [5], which states, "Regularities observable in the picture *are* not from accidental alignments, but are projections of real regularities." Examples of the principle include the following.

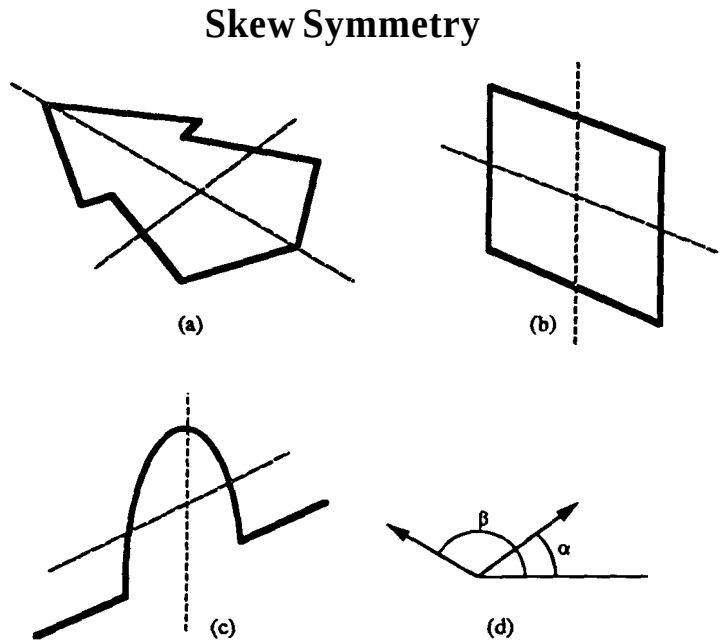


EXHIBIT 14.3
Skew symmetry.

- Parallelism: parallel lines in an image are to be interpreted as parallel lines in 3D.
- *Texture gradient*: a gradient in the spacing of textured elements is interpreted as regularly spaced elements in the 3D world with a surface slant relative to the viewer.
- ***Skew symmetry:*** skew symmetry in an image is interpreted as real symmetry viewed from some unknown view direction.

Skew symmetry was a new concept that I introduced [5]. As illustrated in Exhibit 14.3, skew symmetry is an image feature in which a reflective property is observed with respect to skewed axes, rather than perpendicular axes. Relating a skew symmetry in an image to a real symmetry in space creates strong constraints on surface orientations.

An important point about nonaccidental regularities is that 3D regularities in the scene always result in corresponding 2D regularities in the image, but the inverse is not always true. For example, parallel lines in 2D could be the result of

a particular alignment of nonparallel 3D lines. The probability of such an alignment is vanishingly small, however. The principle of nonaccidental regularities formalizes the fact that the preferred interpretation of 2D regularities is in terms of 3D regularities. Once we assert the principle of nonaccidental regularities, we *can* use a mathematical technique, such as the gradient space representation [6], to map the image properties into the constraints that the interpreted shape must satisfy. Those constraints *can* be used to narrow and screen the possible interpretations while creating partial interpretations. Coming back to the box example, we *can* actually prove that the so-called rectangular natural box is the only interpretation that can satisfy the nonaccidental regularities principle.

14.2.4 Perspective: Geometric Constraints

One of the contributions of the Origami World is that it demonstrated a simple fact in vision: there are a multiplicity of possible image interpretations. and *if* we want to reach a unique interpretation, we must use constraints or heuristics. Since humans usually think of only a single interpretation, many vision researchers accepted, probably too hastily, the requirement that a computer vision program *must* also generate only a single interpretation. Early researchers attempted to meet this requirement by incorporating heuristics, often implicitly, without understanding their effects, limitations, or implications. In contrast, in the Origami World, interpretation was constrained by the principle of nonaccidental regularities, which enumerated a collection of rules relating image and world features, and permitted an exact specification of the set of possible interpretations.

The individual rules of nonaccidental regularities may have been conceived from observations of human perception, and they are heuristic in the sense that they do not always hold. However, the principle of nonaccidental regularities was applied in ways that are purely geometrical and rigorous. The implications were clearly defined and therefore it was possible to predict the consequences when rules did not apply. In this sense they are not ad hoc. This is in direct contrast to the heuristic methods, ranging from Guzman's line-drawing interpretation method [1] of the early 1970s to the use of global minimization of a certain energy-related term to resolve ambiguities in matching, smoothing, or interpreting patterns. In these cases, the implications are neither clearly defined nor predictable in terms of physical reality.

A series of works appeared in the last decade which formalized many of the computational constraints which relate properties in the image domain to 3D shape constraints. The contributions of our CMU vision group include a theory for affine-transformable patterns by Kanade and Kender [6], Kender's theory of shape from texture [8], Shafer's theory for recovering shape from occluding contours of generalized cylinders [27], and, more recently. Krumm and Shafer's analysis of image spectrograms [12].

14.3 Color and Reflectance

14.3.1 A Problem: Highlights in Color Photographs

Examine the color image of Color Plate 1. In addition to cylindrical and toroidal shapes of objects, we can readily recognize that the object surfaces are plastic and glossy in appearance. Also, we can conclude that the bright white regions are due to highlights. Interestingly, we do not interpret them as white paint on the surfaces.

Shape, surface glossiness, and specularity are scene properties that we seem to be able to deduce from the color image, although there is no apparent direct one-to-one mapping between observable features in the image and those scene properties. From introspection we may develop a heuristic rule for the task of extracting highlights from the image, such as

If intensity > 100, *then* highlight.

This rule may work most of the time. It is clear, however, that this rule does not capture the essence of highlights. Thus it will fail, but we don't know when or exactly why. Highlights in a color photograph must be a result of some physical process that involves shape, surface properties, and illumination. Isn't there a more systematic way, based on physical knowledge rather than phenomenological descriptions, to detect highlights, and even to recover some of the properties of the object and the illumination from the image?

14.3.2 Dichromatic Reflection Model

Shafer, Klinker, and myself have worked on this color understanding problem since 1984. Our approach was to call upon a physical model of color reflection. The model we used is called the dichromatic reflection model [28] for opaque dielectric materials, such as plastics. Exhibit 14.4 sketches the primary reflection processes. When light from the illumination source hits the surface, it smkes the interface with the transparent medium. Some of the light is reflected immediately according to Fresnel's laws. This light, which we call *surface reflection*, has a color that is typically about the same as the illuminant. Surface reflection accounts for the glossiness. Surface reflection is highly directional—if the surface is smooth, the surface reflection will be very specular, creating highlights; if the surface is rough. it will be somewhat diffused.

The light that is not reflected at the interface penetrates into the bulk of the material, and there it begins to scatter off the pigment or other colorant particles in the material. Eventually, some portion of it is reflected back across the interface into the air; we call this body *reflection*. Light from body reflection

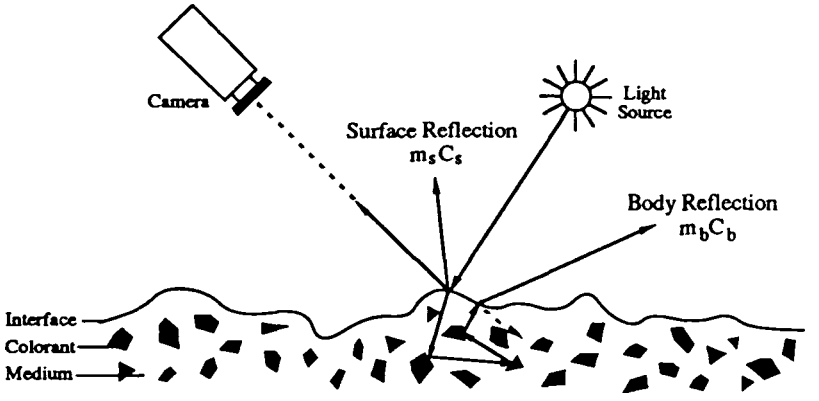


EXHIBIT 14.4
Dichromatic model of color reflection.

has a color that is determined by the object colorant as well as the illuminant. A typical and appropriate model for the strength of body reflection is the Lambertian model, which states that the amount of reflection is uniform in direction and is determined by the product of the reflectivity of the material and the cosine of the incident angle. Thus, surfaces facing more toward the light source show brighter color, while those facing more away from it show darker color. Body reflection is responsible for "object color," which is the characteristic color of a specific object, and its shading provides an important clue to the perception of the object shape.

In summary, the dichromatic reflection model states that the observed color at each point in the image consists of two colors, the surface reflection color and the body reflection color. Hence, the name "dichromatic reflection model." Under the same illumination, the two component colors themselves do not vary across surfaces with the same color, but the magnitudes (relative intensities) of these color components vary from point to point due to variation in the geometric relationships between the surface and the light source. Thus if we represent a color by the 3D color vector $C = (R, G, B)$, the color at (x, y) in an image is given by

$$\mathbf{C}(x, y) = m_s(x, y)\mathbf{C}_s + m_b(x, y)\mathbf{C}_b$$

where:

$\mathbf{C}_s = (R_s, G_s, B_s)$: color of surface reflection

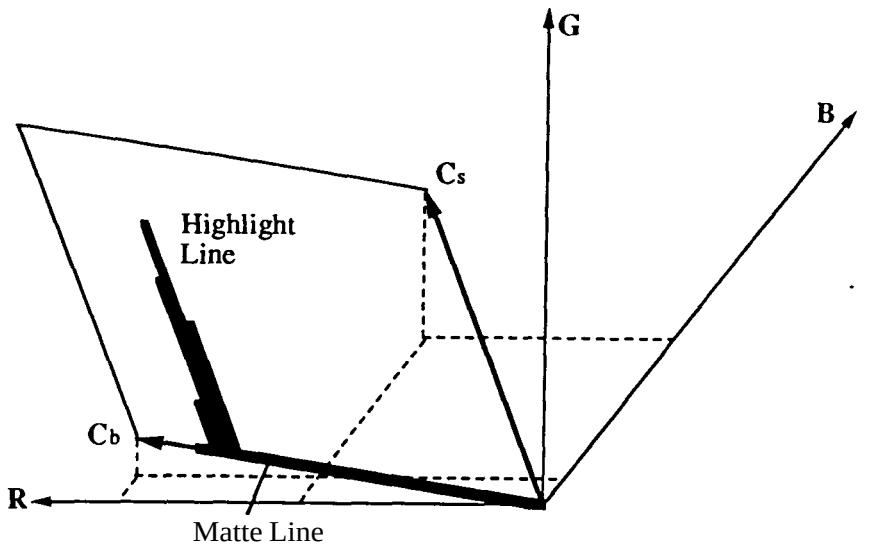


EXHIBIT 14.5
Color distribution is constrained on a dichromatic plane.

$\mathbf{C}_b = (R_b, G_b, B_b)$: color of body reflection
$m_s(x, y)$ and $m_b(x, y)$: scalar magnitudes of surface and body reflections, respectively.

An interesting interpretation of this model arises if we examine the histogram of image colors from the points belonging to a single surface. According to the model, the observed color vector $\mathbf{C}(x, y)$ at a pixel is a linear combination of two vectors $\mathbf{C}_s$ and $\mathbf{C}_b$. This means that even though the red plastic doughnut in Color Plate 1 includes various colors in it-bright red, dark red, and even white—they cannot be distributed arbitrarily in the color space. They must be on the plane, called the dichromatic plane, spanned by the two vectors as shown in Exhibit 14.5. Moreover, for most points. there is very little surface reflection; thus $m_s(x, y)$ is nearly zero and the color simply lies somewhere along the vector $\mathbf{C}_b$. We call this a *matte line.* All the points with significant amounts of highlight come from a small area on the object surface and thus have nearly the same amount of body reflection $m_b(x, y)$. Thus they form a sort of spike, called a *highlight line,* in the color space whose direction is parallel to the illuminant color $\mathbf{C}_s$.

This observation has been verified experimentally, both by ourselves [10] and other researchers [13,31]. In fact, Tominaga [31] has found that the model holds for a wide range of material surfaces. In the upper left image of Color Plate

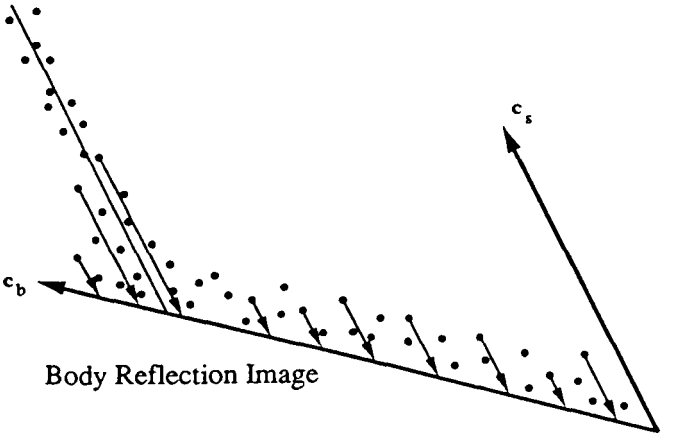


EXHIBIT 14.6
Body reflection **image.**

2, a plastic orange cup is illuminated by a white light. The upper right image of the plate shows the histogram of the color image. As expected, we see **an** L-shaped distribution in a color space. We observe a matte line in the direction of the body reflection, C_b, and a highlight line in the direction of the surface reflection, C,. The bend at the end of the highlight line is due to saturation of the camera. The highlight is **so** intense that certain color components (in this case the red value) have reached their maximum value, and can no longer change, thus resulting in a bend, which we call the ***saturation line.*** By examining the dismbution on the dichromatic plane, we can identify the matte line, the highlight line, and the saturation line.

14.3.3 Separation of Highlights and Image Segmentation

We *can* now write **a** program [10] that analyzes the color distribution of a given picture and identifies the matte and highlight lines. thereby calculating the vectors C, and C_b. Then, refemng to Exhibit **14.6,** imagine that we project all the colors in the dichromatic plane onto the matte line in the direction of the surface reflection, **i.e.,** along the vector C,. In other words, we force the value of m_s to **be** zero and calculate the color consisting of only the body reflection. If we generate **a** picture from this projected distribution, we should see a picture of the scene with no surface reflection. Since the surface reflation accounts for

glossiness and highlight, the resulting picture should lose all the glossiness and highlights and includes only shaded matte color. The lower right image in Color Plate 2 is such a result for the upper left image. It should be noticed that not only have we removed the highlight, but we have recovered the color behind it. This is the picture we would see if the object was not made of plastic, but of a material with a matte surface.

Similarly, if we project all of the colors on the dichromatic plane, along C_b, onto the vector $\mathbf{C}_s$, then we force m_b to be zero. This means that we have colors with only surface reflection and no body reflection. If we generate a picture from this distribution, it will show only highlights; this is shown in the lower left image in Color Plate 2. It should also be noted that highlights are not binary phenomena; they have gray scale.

The algorithm just presented cannot be applied directly to our original problem (see Color Plate 1), since the image includes multiple objects. The color histogram of the whole image is shown in Color Plate 3 and clearly shows that it consists of many L-shaped histograms, each of which must be treated individually. If we know that a particular region of the image comes from an object of a single color, the distribution of the color within that particular region will follow our constraints. However, since the image includes multiple colors, we have to find out which region corresponds to each color. In order to distinguish each region, we have to know its true color, since the apparent color in the image can vary significantly, even for the same object This is exactly the same circular problem that computer vision researchers previously encountered in segmenting color images into objects, and without a systematic model they had to rely on the assumption of uniform color [24].

What we need to break the cycle is a way to group image points that accounts for the color variation accountable by the physical model. In fact, all we really need is a way to examine a small neighborhood of the image and make a good guess about the reflection color vectors. Once we make such a hypothesis about the model for each neighborhood, we can measure the extent of the neighborhoods whose color distribution can be explained by the model vectors, and then group together those neighborhoods. We have developed a method for color image segmentation by devising techniques for creating and testing such local hypotheses [9,11]. Color Plate 4 shows the result of segmenting Color Plate 1. Notice that the segmentation is not affected by highlights or shadings which have often fooled traditional image segmentation algorithms based on apparent colors.

Once we have segmentation, we can apply the previous analysis to each region and separate the body and surface reflections. In fact, since our segmentation method hypothesizes reflection color vectors, the projection into reflection components is a simple by-product of segmenting the image. Color Plate 5(a) shows the body reflection image of the whole scene, and conversely, Color Plate 5(b) shows the surface reflection image.

14.3.4 Perspective: Optical Constraints

Given the original input image, we have succeeded in automatically segmenting highlights and in calculating the apparent, shaded color of the object. We did not rely on any traditional heuristics based on clustering techniques or a phenomenological theory of color perception. The physical reflection model of color provided constraints that we exploited for analysis. In the past, color segmentation and edge detection almost always been based on grouping of points with uniform or nearly uniform color [24]; but these techniques utterly fail when presented with an image of an object that has a bright highlight of a color different from the object color.

It should be noted that we did not assume any prior knowledge about the real colors and shapes of objects nor the real color and direction of the light source in separating body and surface reflections. Actually, once we have the separation, there is a possibility for recovery of these four unknowns from the given image. First, since the body reflection image admits a Lambertian model, we can apply the shape-from-shading method for shape recovery. Second, once the shape (and thus surface orientations) is known, the locations of the peaks of surface reflection provide constraints on the direction of illumination due to the mirror-like reflection geometry. Third, the color of surface reflection roughly corresponds to that of the illumination. Finally, combining the knowledge of illumination color and the body reflection image will enable us to recover the real object color.

In the early 1970s, Horn of Massachusetts Institute of Technology pioneered the use of a reflection model in computer vision in his work on image intensity understanding [2]. Various methods, notably shape from shading and photometric stereo. were developed thereafter. They have suffered, however, from the use of models of reflection that were too idealized (such as a pure Lambertian model) and the lack of appropriate models to account for interreflections. It has been recognized that formulation of more sophisticated and realistic models of reflection to deal with a broader class of surfaces and to cope with interreflection is necessary to make the approach more realistic and powerful.

Nayar, Lkeuchi, and Kanade proposed a unified reflectance model composed of the diffuse lobe, the specular lobe, and the specular spike [21]. It is capable of describing the reflection from surfaces that may vary from very smooth to very rough. Another significant advancement deals with interreflection due to concave surfaces or concavities formed by multiple objects in the scene. The interreflection causes almost all of the existing shape-from methods based on image intensity to produce erroneous results. Interreflection is a very difficult problem for which very little research has been done [14]. Nayar, Ikeuchi, and Kanade developed a theory of shape from interreflections and demonstrated recovery of the shape of an object even under interreflection with unknown surface reflectances [20]. Also, Novak and Shafer have been develop-

ing a model for color interreflection as an extension of the dichromatic reflection model [26].

Appropriate physical modeling can result in practical, useful devices with robust capabilities. Based on the unified reflectance model, we have built a new device, called a photometric sampler, for surface inspection [19]. It uses extended light sources and can extract reliably both shape and reflectance properties of hybrid surfaces. Also, Nayar and Nakagawa developed a practical inspection device for such surfaces as a tungsten paste filling in a via-hole on a ceramic substrate that has a size of about 100 microns and includes specular reflection and 3D texture [22]. Based on an appropriate model of reflection of rough surfaces and focusing, the device can measure the shape with an accuracy of several microns.

Also, in developing physically based vision. we came to realize the need for a controlled environment where we can take images with accurate knowledge of ground truth and where we can control lighting and camera parameters. Such an environment is critical in order to accurately test and evaluate vision theories and methods. Traditionally, in the computer vision community, the test images were taken without much control, and therefore it was often unclear whether a theory being tested was incorrect or the data were inappropriate for testing the theory. We have built a unique facility called the Calibrated Imaging Laboratory (CIL) [25]. The CIL consists of many television cameras including a very high precision camera, controllable lighting, a high-precision 6-degrees-of-freedom computer-controllable jig to mount and move cameras, filters, test objects, and associated electronics. The CIL has made a significant impact on turning computer vision into a quantitative scientific discipline, and has introduced a new area of research in how to obtain high-quality images for computer vision by active control of the camera and lens [23].

14.4 Shape, Motion, and Uncertainty

14.4.1 A Problem: Baseline and Uncertainty in Stereo

Stereopsis is one of the fundamental ways to measure depth. Exhibit 14.7(a) illustrates the geometry for a binocular stereo system with the left camera L and the right camera R separated by a baseline B. If an image feature point, such as an edge, located at x_L in the left image and an image point at x_R in the right image are projections of the same physical point P in space, then by triangulation, we can measure the distance to P. This is the principle of depth measurement by stereo vision.

In practice. the points x_L and x_R of features in the images can be located only within a certain accuracy. This is due to both image noise and limited sensor

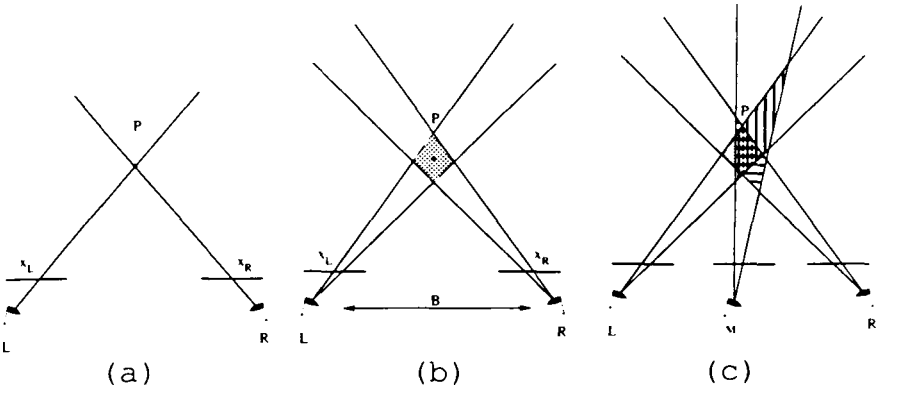


EXHIBIT 14.7
Stereo geometry.

resolution. Uncertainty in the image-position measurements leads to uncertainty in the final calculation of the scene point, as indicated by the diamond-shaped region in the Exhibit 14.7(b). The uncertainty in depth can be reduced by increasing the baseline (i.e., by spreading the cameras farther apart). since then the triangle becomes shallower, making the position of the vertex less sensitive to the orientation of the sides. This is exactly the constraint that civil engineers apply in surveying.

Finding pairs of corresponding points, x_L and x_R, that come from the same physical point, usually called the correspondence problem, is actually the hardest part of the stereo problem. Making the baseline longer unfortunately makes the correspondence problem more difficult. The longer the baseline is, the more different the right and left images are from each other. The same point in space may appear differently due to a different viewpoint and foreshortening. or may even appear in only one image due to occlusion. Here we have a fundamental dilemma in stereo vision: as we make the baseline longer, the depth measurement becomes more accurate, but at the same time the matching becomes more difficult. and vice versa. How can we solve this dilemma?

14.4.2 Managing Uncertainty for Incremental Stereo

Matthies, Szeliski. and Kanade have analyzed the structure of uncertainty in stereo measurements [17]. Imagine that we place one more camera *M* in the

middle of the baseline as shown in Exhibit 14.7 (c). This creates two more stereo problems: one between cameras L and M and the other between cameras M and R. With shorter baselines, the correspondence problems for the two intermediate stereos are less severe than the original stereo between L and R, even though the depth measurements by them would not be as good. By solving the two new matching problems successfully, we have solved the original matching problem, because the middle point corresponds to both the left and right points. This idea is called trinocular stem, and has been studied by several researchers to exploit the additional constraints that the middle camera provides.

However, an interesting question that had not been asked before concerns the uncertainty of measurements. Do the two new measurements due to the middle camera M help reduce the uncertainty in the depth measurement of P? Matthies and Shafer [18] gave a way to relate the uncertainty of image measurements with the uncertainty of depth measurement and model it by a covariance. By using that formulation and the theory of optimal estimation, it was shown that the answer to the question is "yes" despite the fact that the two additional measurements are expected to be more uncertain than the original one. Though straightforward, this conclusion is very significant.since bringing in the third camera helps not only to simplify the matching problem but also reduce the uncertainty. We can add more camera positions between L, M, and R for further improvement. and so on.

An analysis proves that if N cameras are placed between L and R, then the uncertainty of the depth measurement decreases at the rate of N cubed. That is

$$\sigma^2(N) \sim \frac{1}{N^3}\sigma_e^2$$

where $\sigma^2(N)$ is the uncertainty of the final depth measurement with N cameras, and σ_e^2 is the uncertainty of the image-position measurement. Experiments using real images demonstrated that the uncertainty decreases as expected, as shown in Exhibit 14.8 which plots $\sigma(N)$ versus the number of cameras.

The above derivation may have given the impression that the solution requires processing of all N images at the same time. Rather. the solution can be implemented sequentially by using pairs of images from neighboring cameras at a time. Thus we can devise an incremental stereo system that produces and refines a depth map as a single camera is moved sideways. For this purpose we reformulated the system equations in terms of the current frame by using the feature position x_i in the i-th image and the inverse of the depth $\frac{1}{z}$ to be the state variables of a dynamic system. This reformulation allowed us to view the incremental stereo as an instance of dynamic system estimation and to apply the Kalman-filtering technique.

The diagram of Exhibit 14.9 explains the method. A camera is moved from left to right by a small amount at each step. The sequence of images is processed

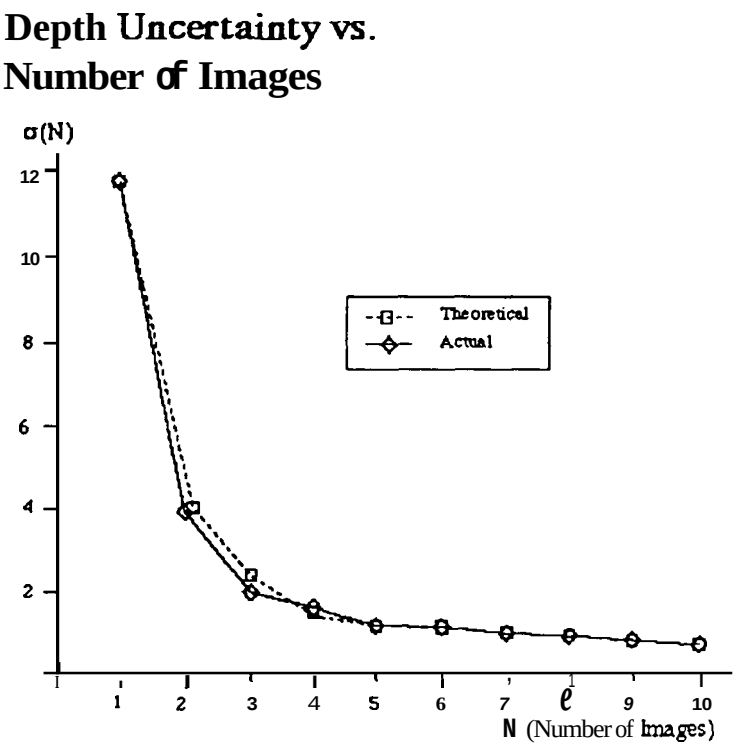


EXHIBIT 14.8
Uncertainty versus number of images.

by the method shown in the diagram. The first pair of neighboring images are processed as a stereo pair. The resulting inverse-depth map, though very noisy, is stored. The next pair of images are then processed similarly to produce the second depth measurement. In the meantime, the stored depth map is transformed into the depth map in the coordinate system of the next camera position. The two depth maps are integrated by using the Kalman filtering technique, and stored. The process repeats until the camera reaches the end.

An actual experiment was done using a scale model of a city in the CIL. Ten images were taken in which consecutive images were taken only 0.05 in. (1.27 mm) apart. Thus the total baseline was only 0.5 in. (1.27 cm). while the distance to the scene was 20 to 40 in. (50 to 100 cm). Exhibit 14.10 shows the first image of the sequence. The final result was a depth map of the scene. Exhibit 14.11

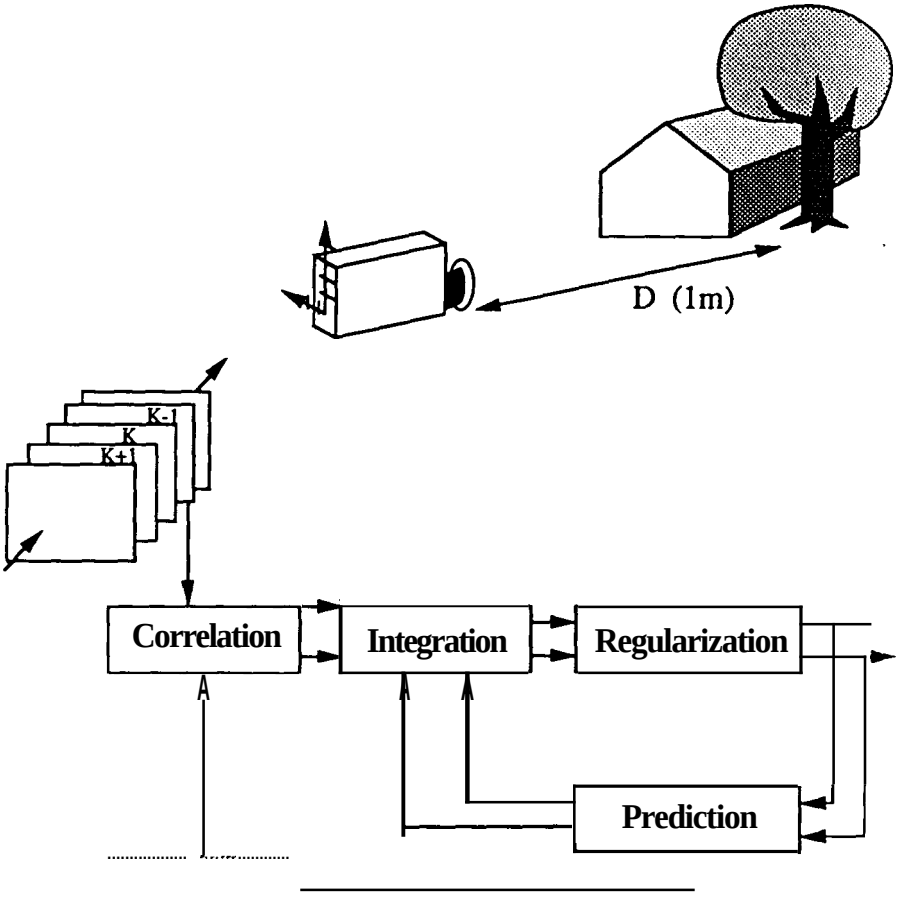


EXHIBIT 14.9
Depth map recovery by a Kalman filter method.

shows the depth map presented as a grey-level image, *in* which closer points are encoded brighter. A perspective view of the reconstructed scene, made by "painting" the original intensity image on the depth map is presented in Exhibit 14.12. We can see that the structure of the scene, including buildings, streets, *cars,* trees, and a distant bridge, is well recovered.

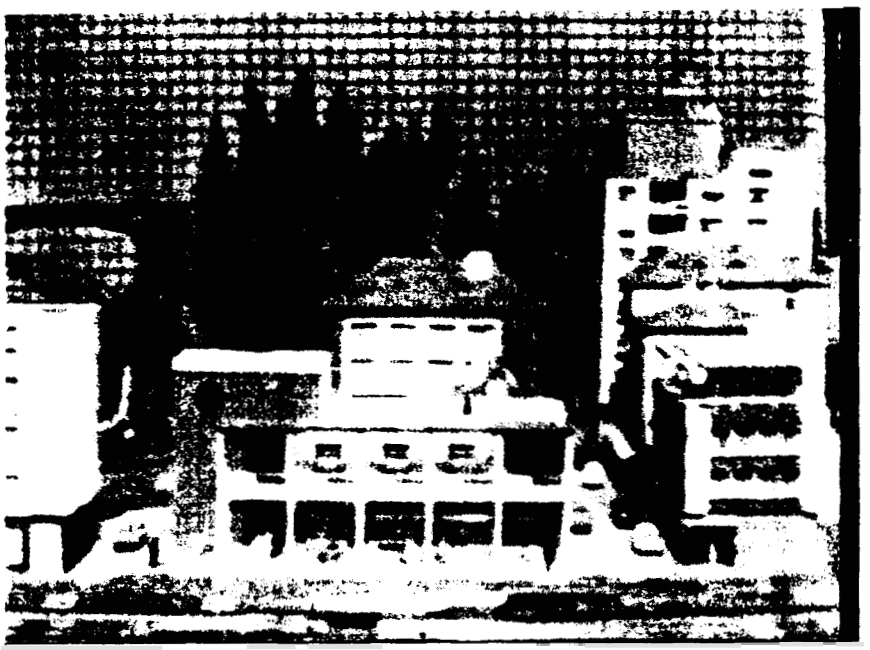


EXHIBIT 14.10
The first image of the sequence of ten images of a scale model of a city.

14.4.3 Perspective: Statistical Constraints

It should be noted that the above results were obtained for a stereo with an extremely narrow baseline: 1.27 mm for the neighboring pair and 1.27 cm for the farthest pair, and the triangle with a 1:100 ratio of baseline to scene depth. As a result, each pair of stereo images are so close that the matching or correspondence problem has become almost trivial. Therefore although many images must be processed, the total computation has not increased by much. In fact, there is a chance that it is reduced because the computation is now very local and uniform. This was made possible because we analyzed and modeled the structure of uncertainty in stereo, and developed the algorithm based on that model.

In the past, literature on stereo exclusively dealt with correspondence problems. Our work on incremental stem has shown that another important problem in stereo is management of measurement uncertainty. Uncertainty becomes particularly important when the stem is used with a mobile robot for both re-

EXHIBIT 14.11
The computed depth map.

covering the depth map and locating itself in the environment; in this case, the uncertainty of depth measurements and the uncertainty of robot motion interact and create a cycle of uncertainty [16]. Mauhies and Shafer demonstrated that appropriate modeling of depth uncertainty can greatly improve the accuracy in recovering the robot motion [18]. Mauhies [15] developed dynamic stereo vision for using stereo vision both to estimate the 3D structure of the scene. and to *estimate* the motion of the robot as it travels through an unknown environment. The key idea is to monitor the uncertainty of the depth map and to use appropriate stereo systems, either narrow baseline or wide baseline, depending on the situation. **Szeliski [29]** developed a framework for Bayesian modeling of uncertainty in low-level vision. His framework allows us to define and compute a prior model of the scene, a sensor model and a posterior model.

A new stereo algorithm with an adaptive window developed by Kanade and Okutomi [7] relates the disparity uncertainty with the matching uncertainty. They showed that by modeling the disparity (inverse depth) variation within a matching window, the uncertainty of matching can be evaluated. Therefore it

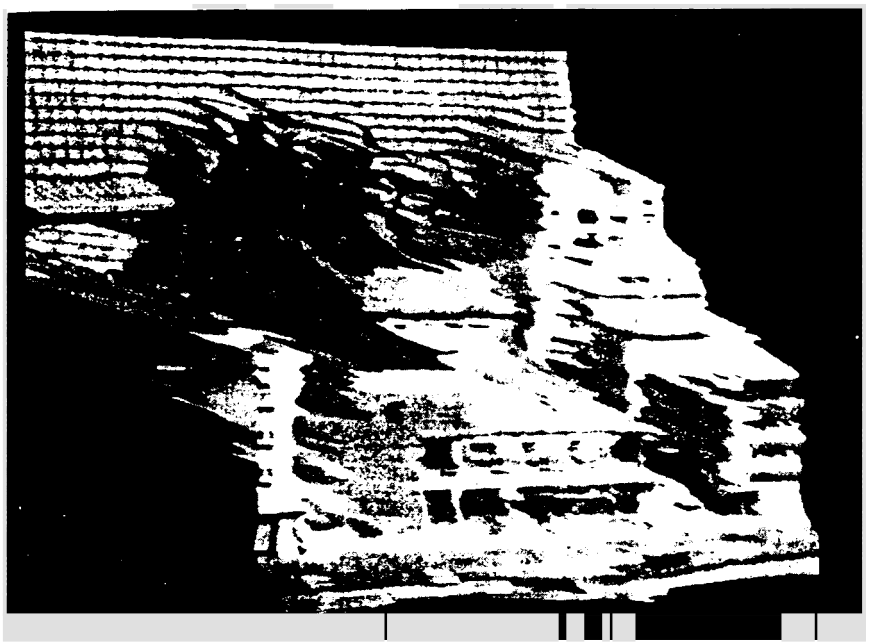


EXHIBIT 14.12
Perspective views of the recovered depth map.

allows selection of the window size that results in the disparity estimation with the least uncertainty.

Tomasi and Kanade [30] showed that, in the problem of recovering motion and structure from a sequence of images taken from a long distance, it is far more advantageous to recover the relative shape of an object directly, rather than going through absolute depth. Their theory of shape-from-motion without depth captures the effect of noise as constraints on the approximate rank of a matrix. Based on the theory, they have demonstrated very accurate recovery of motion (less than 0.02% error) and shape (less than 0.5% error) from a sequence of images taken of an object of size 4 cm from a distance of approximately 3.5 meters.

14.5 Conclusion

The physically based theory of vision presented here has focused on determining the information that is contained in images and developing the constraints that are applicable to extract the information. We have emphasized the difference between heuristics that seem to work some of the time and constraints that are correct for a well-defined range of situations. These constraints can be derived from models of the geometry, physics, optics, and statistics of vision. Vision based on physical models results in the formulation of problems in such a way that extensions and modifications can be clearly stated and researched. Moreover, the limitations of such models can be deduced.

It may have been noticed that the presentation above included little or no discussion about "algorithms" that perform the task of extracting information or on "systems" or "computational mechanisms" that implement the algorithms. In physically based vision, the emphasis is on the formulation of the physical models and derivation of constraints; algorithms and implementations can be developed based on the models. This approach is almost parallel to Marr's three levels in understanding vision [14]: computational theory; representation and algorithms; and implementation. Our physically based theory is most akin to Marr's level of computational theory, in which the performance of the device is characterized as a mapping from one kind of information to another, the abstract properties of this mapping are defined precisely, and its appropriateness and adequacy for the task at hand are demonstrated. Our theory certainly deals with these issues rather than how to implement the theory efficiently on specific hardware.

A subtle, yet important, difference between our work and that of the Marr school is that our theory focuses on formulating the structure in which the physical processes involved in creating or acquiring images encode information, and in extracting the constraints that can be exploited in decoding images based on the physical models. Marr was largely motivated by human visual perception, so the "what," "why," and "how" of computer vision were often justified relative to human perception. We do not need to limit computer vision tasks to those for which human visual perception systems have counterparts. Currently, humans outperform machines by far in most tasks. Thus computer vision can learn much from human visual systems. However, this should not mean that copying or mimicking a human, in either defining goals (phenomenological performance), devising solutions (introspections), or implementing algorithms (neurons), is the best way to lead to the ultimate computer vision systems. In fact, machines may well exceed humans in the future (and in some cases do so already).

In the past, a vision problem was often stated as, "Given an image, devise an interpretation algorithm for ...." Vision researchers then rushed out to write

a program for dealing with the given image. However, according to Marr [14], vision is the process of discovering from images (data) what is present in the (physical) world and where it is. If so, as with any physical science, a clear technical understanding of the physical nature of the data (i.e., images) is required for formulating a solution. The physical processes underlying vision take place before the interpretation of images starts. Hence, computer vision must be a physical science. At least half of it.

14.6 Acknowledgments

I would like to thank all the members of the CMU vision group for their contribution—not just those whose work I referred to in this chapter, but all the others that have contributed to transforming computer vision to a physical science. Steven Shafer, who is one of the strongest driving forces of the effort, provided critical comments on the chapter, and Keith Gremban spent long hours in reading and editing several versions of this chapter during its evolution. I thank them for their comments and inputs, which have greatly improved this chapter.

References

[1] A. Guzman. Computer recognition of three dimensional objects in a visual scene. Technical Report MAC-TR-59. Massachusetts Institute of Technology. Cambridge. MA. 1968.

[2] B. K. P. Horn. Understanding image intensities. *Artificial Intelligence,* 8(2):201-231, 1977.

[3] D. A. Huffman. Impossible objects as nonsense sentences. In B. Meltzer and D. Michie, eds., *Machine Intelligence 6,* chapter 19, pp. 295-323. American Elsevier Publishing, New York 1971.

[4] T. Kanade. A theory of origami world. *Artificial Intelligence.* 13:279-311, 1980.

[5] T. Kanade. Recovery of the 3D shape of an object from a single view. *Artificial Intelligence.* 17:409-460, 1981.

[6] T. Kanade and J. R. Kender. Mapping image properties into shape constraints: Skewed symmetry, affine-transformable patterns. and the shape-from-texture paradigm. In Human and *Machine Vision.* Academic Press, 1983.

[7] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. Technical Report CMU-CS-90-120, Carnegie Mellon University School of Computer Science, 1990.

[8] J. R. Kender. *Shape from Texture.* PhD thesis, Department of Computer Science, Carnegie-Mellon University, 1980.

[9] G. J. Klinker. *A Physical Approach to Color Image Understanding.* PhD thesis, Carnegie Mellon University, Computer Science Department, 1988.

[10] G. J. Klinker. S. A. Shafer. and T. Kanade. The measurement of highlights in color images. ***International Journal of Computer Vision,*** 2(1):7-32, 1988.

[11] G.J. Klinker. S.A. Shafer. and T. Kanade. A physical approach to color image understanding. *International Journal of Computer* ***Vision,*** 4(1):7-38, 1990.

[12] J. Krumm and S. Shafer. Local spatial frequency analysis for computer vision. Technical Report CMU-RI-TR-90-11. Camegie Mellon University The Robotics Institute. 1990.

[13] H. C. Lee. Method for computing the scene-illuminant chromaticity from specular highlights. ***Journal of the*** *Optical* ***Society of America.*** 3(10):1694-1699, 1986.

[14] D. Marr. ***Vision.*** Freeman. 1981.

[15] L. Matthies. Dynamic stereo vision. PhD thesis, Camegie Mellon University. Computer Science Department, 1989.

[16] L. Matthies and T. Kanade. *The* cycle of uncertainty and constraints in robot perception. In B. Bolles and B. Roth. ***eds.. Robotics Research,*** volume 4, pp. 327-336. Cambridge, MA: MIT Press, 1988.

[17] L. Matthies. T. Kanade. and R. Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. ***International*** *Journal* of ***Computer*** *Vision.* 3:209-236, 1989.

[18] L. Matthies and S. Shafer. Error modeling in stereo navigation. IEEE ***Journal Robotics and Automation.*** pp. 239-248, December 1987.

[19] S. Nayar, K. Ikeuchi. and T. Kanade. Shape and reflectance from an image sequence generated using extended sources. In ***Proceedings*** of 1989 ***IEEE International Conference on Roboticss and Automation,*** pp. 28-35. New York: Institute of Electrical and Electronics Engineers, 1989.

[20] S. Nayar. K. Ikeuchi, and T. Kanade. Shape from interreflections. Technical Report CMU-RI-TR-90-14. Carnegie Mellon University, Robotics Institute, 1990.

[21] S. Nayar. K. Ikeuchi, and T. Kanade. Surface reflection: Physical and geometrical perspectives. In ***Proceedings of Image Understanding Workshop,*** Morgan Kaufman, 1990.

[22] S. Nayar and Y. Nakagawa. Shape from focus: An effective approach for rough surfaces. In ***Proceedings*** of ***I990 IEEE International Conference on Roboticss and Automation.*** New York Institute of Electrical and Electronics Engineers, 1990.

[23] C. Novak. S. Shafer. and R. Willson. Obtaining accurate color images for machine vision research. In ***Proceedings SPIE Conference on Perceiving.*** *Measuring,* ***and*** *Using* ***Color.*** Santa Clara CA, 1990. SPIE.

[24] R. Ohlander. K. Price. and D. R. Reddy. Picture segmentation using a recursive region splitting method. ***Computer Graphics and Image Processing,*** 8:313-333, 1978.

[25] S. Shafer. The Calibrated Imaging Lab under construction at CMU. In ***Proceedings DARPA Image Understanding Workshop,*** p. 509. SAIC. 1985.

[26] S. Shafer. T. Kanade. G. Klinker, and C. Novak. Physics-based models for early vision by machine. In ***Proceedings SPIE Conference on Perceiving, Measuring, adn Using Color,*** Santa Clara. CA, 1990. SPIE.

[27] S. A. Shafer. ***Shadow Geometry and Occluding Contours of Generalized Cylinders.*** PhD thesis, Department of Computer Science, Carnegie-Mellon University, 1983.

[28] **SA.** Shafer. Using color to separate reflection components. Color ***Research and Application***, 10(4):210-218, 1985.

[29] **R.** Szeliski. ***Bayesian Modeling of Uncertainty in Low-Level Vision.*** PhD thesis. Camegie Mellon University. Computer Science Department. 1988.

[30] C. Tomasi and T. Kanade. Shape and motion from image streams: a factorization method. Technical Reports CMU-CS-90-166 and CMU-CS-91-105, Carnegie Mellon University, School of Computer Science. 1990, 1991.

[31] **S.** Tominaga and B.A. Wandell. Standard surface reflectance model and illuminant estimation. ***Journal of the Optical Society of America,*** 6(4):576-584, 1989.

[32] D. Waltz. Generating semantic descriptions **from** drawings of scenes with shadows. In P. H. Winston, ed.. ***The psychology of Computer Vision,*** pp. 19-92. McGraw-Hill, New **York.** 1975.