Perceptual Optics, Visual Radiometry, The Imaging Equation, and Their Role in Visual Reconstruction*

Stefano Soatto
University of California, Los Angeles
soatto@ucla.edu

Technical Report UCLA CSD-TR030055
December 8, 2003

Abstract

These notes describe a simple image formation model suited for visual inference (reconstruction, reprojection and recognition). Image formation models commonly used in computer graphics are not suitable for analysis, since they contain parameters that are not identifiable. In this sense, physical optics is not suitable for visual inference and reconstruction. Therefore, we seek the simplest possible model that is general enough to capture the phenomenology of natural images, and at the same time is minimal in the sense of containing all and only the parameters that can be identified.

1 When do two images portray the same scene?

Many tasks in vision can be traced back to the question: "when do two (or more) images show (portions of) the same scene?" Figure 1 illustrates the problem. In order to answer we need to agree on what "images" are and what "scene" (or "object") means. More importantly, we need to agree on how the two are related.

1.1 What is the "image" ...

An "image" is just an array of positive numbers that measure the intensity (irradiance) of light (electromagnetic radiation) incident a number of small regions ("pixels") located on a surface. We will deal with gray-scale images on flat, regular arrays, but one can easily extend the reasoning to color or multi-spectral images on curved surface, for instance omni-directional mirrors. In formulas, a digital image is a function $I : [0, N_x - 1] \times [0, N_y - 1] \to [0, N_g - 1];\ (x, y) \mapsto I(x, y)$ for some number of horizontal and vertical pixels N_x, N_y and grey levels N_g. For simplicity, we will neglect quantization in both pixels and gray levels, and assume that the image is given on a continuum $\Omega \subset \mathbb{R}^2$, with values in the positive reals:

$$I : \Omega \subset \mathbb{R}^2 \to \mathbb{R}_+;\ \mathbf{x} \mapsto I(\mathbf{x}) \tag{1}$$

where $\mathbf{x} \doteq [x,\ y]^T \in \mathbb{R}^2$. When we consider more than one image, we index them with t, which may or may not indicate time: $I(\mathbf{x}, t)$. This abstraction in representing images is all we need for the purpose of these notes.

1.2 What is the "scene"...

A simple description of the "scene", or the "object", is less straightforward. This is a *modeling* task, for which there is no right or wrong answer, and finding a right model is as much of an art as it is a science; one has to exercise discretion to strike a compromise between simplicity and realism. We consider the scene

*The author wishes to thank Hailin Jin, Andrea Vedaldi and Ying-Nian Wu for numerous discussions on the topic of these notes. Research related to these notes was sponsored by AFOSR, ONR and NSF.

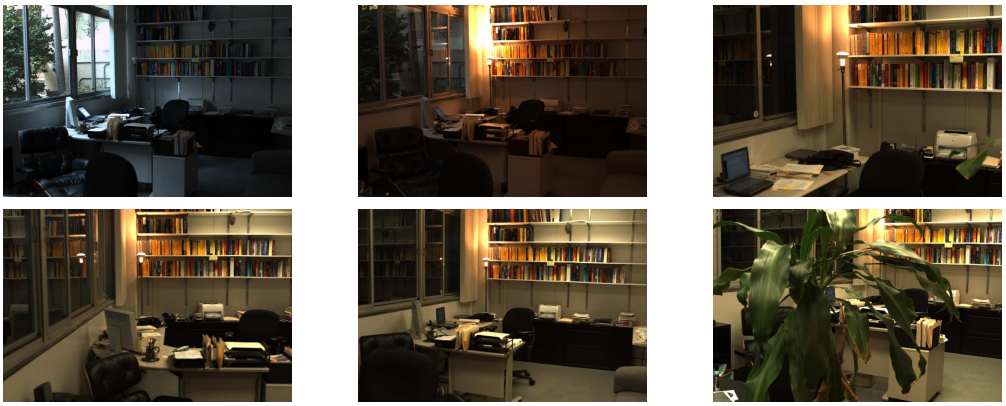

Figure 1: *Examples of variability among different images of the same scene (top-left): illumination (top-center), viewpoint (top-right, bottom-left), removal/replacement of parts (bottom-center), partial occlusion (bottom-right).*

as a collection of "objects" that are volumes bounded by closed, piecewise smooth surfaces embedded in $\mathbb{R}^3$. We call the generic surface S_i, with $i = 1, \dots, N_o$, the number of objects. Each surface is described relative to a (Euclidean) reference frame, which we call $g_i \in SE(3)$. The two entities

$$S_i \subset \mathbb{R}^3; \; g_i \in SE(3) \; \forall \; i = 1, \dots, N_o \tag{2}$$

describe the **geometry** of the scene, and in particular we call g_i the *pose* relative to a fixed ("inertial") reference frame[1] and S_i the *shape* of objects, although a more proper definition of shape would be the quotient S_i/g_i [13]. This is, however, inconsequential as far as our discussion is concerned.

Objects interact with light in ways that depend upon their *material*. Describing the interaction of light with matter is a nightmare if one seeks physical realism: one would have to start from Maxwell's equations and describe the scattering properties of the volume contained in each object. That is well beyond our scope. Besides, we do not seek physical realism, but only to capture the phenomenology of the material to the extent in which it affects the answer to our questions. We will therefore start from a much simpler model, one that is popular in computer graphics, because it can describe with sufficient accuracy a sufficient number of real-world objects: each point p on an object S_i has associated with it a function $\beta_i : \mathbb{H}^2 \times \mathbb{H}^2 \to \mathbb{R}_+; (v, l) \mapsto \beta_i(v, l)$ that determines the portion of energy[2] coming from a direction l that is reflected in the direction v, each represented as a point on the half-sphere $\mathbb{H}^2$ centered at the point p. This is called the *bi-directional reflectance distribution function* (BRDF) and measured in [1/sterad]. This model neglects diffraction, absorption, subsurface scattering etc.; the BRDF only describes the reflective properties of materials (*reflectance*).

Remark 1 (The local frame) *To make the notation more accurate, we define a Euclidean reference frame, called the* local frame*, centered at the point p with the third axis along the normal to the surface, $e_3 = \nu_p \in T_pS_i$ and first two axes aligned with the directions of principal curvature,*[3] *$e_1 = u_p$, $e_2 = v_p$ such that* $\text{span}(u_p, v_p) = T_pS_i$ *(see Figure 2). We call such a local reference frame g_p. The conditions above yield*

$$g_p = \begin{bmatrix} \begin{bmatrix} u_p & v_p & \nu_p \end{bmatrix} & p \\ 0 & 1 \end{bmatrix} \tag{3}$$

[1] If a point p is represented in coordinates via $\mathbf{X} \in \mathbb{R}^3$, then the transformed point gp is represented in coordinates via $R\mathbf{X} + T$, where $R \in SO(3)$ is a rotation matrix and $T \in \mathbb{R}^3$ is a translation vector. The action of $SE(3)$ on a vector is denoted by g_*v, so that if the vector v has coordinates $V \in \mathbb{R}^3$, then g_*v has coordinates RV. See [14], chapter 2 and appendix A, for more details.

[2] The term "energy" is used colloquially here to indicate radiance, irradiance, radiant density, power etc.

[3] Principal curvature directions are the eigenvectors of the curvature tensor, u, v, with principal curvatures κ_u, κ_v as their corresponding eigenvalues (see for instance [3], page 144.). At some points, the principal directions may not be well-defined (e.g. at a point on a plane), in which case u, v will be defined up to a rotation about ν_p.

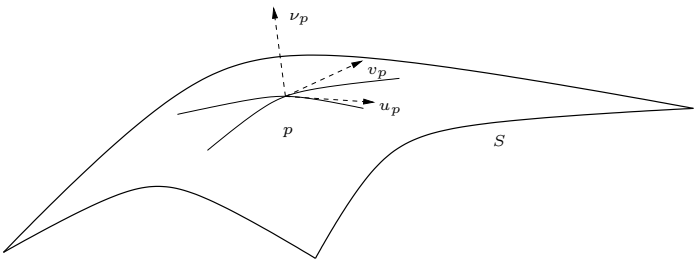

Figure 2: Local reference frame at the point p.

where u_p, v_p and ν_p are unit vectors. Therefore, a point q in the inertial reference frame will transform to $g_p q$ in the local frame at p. Similarly, a vector v in the inertial frame will transform to $g_{p_*} v$ in the local frame where, according to footnote 1,

$$g_{p_*} = \begin{bmatrix} \begin{bmatrix} u_p & v_p & \nu_p \end{bmatrix} & 0 \\ 0 & 0 \end{bmatrix}. \tag{4}$$

The total energy radiated by the point p in a direction v is obtained by integrating, of all the energy coming from the *light source*, the portion that is reflected towards v, according to the BRDF. The light source is the collection of objects that can radiate energy. In principle, every object in the scene can radiate energy (either by reflection or by direct radiation), so the light source is just the scene itself, $L = \cup_{i=1}^{N_o} S_i$, and the energy distribution can be described by a distribution of directional measures on L, which we call $dE \in \mathcal{L}_{\text{loc}}(L \times \mathbb{H}^2)$, the set of locally integrable distributions on L and the set of directions. These include ordinary functions as well as ideal delta measures. The distribution dE depends on the properties of the light source, which is described by a function $R_L : L \times \mathbb{H}^2 \to \mathbb{R}$ of the point q on the light source and a direction (see next subsection for the relation between dE and R_L). The collection

$$\beta_i(\cdot,\cdot) : \mathbb{H}^2 \times \mathbb{H}^2 \to \mathbb{R}_+,\ i = 1, \dots, N_o;\ L \text{ and } dE : L \times \mathbb{H}^2 \to \mathbb{R}_+ \tag{5}$$

describes the **photometry** of the scene (reflectance and illumination). Note that β_i depends on the point p on the surface, and we are imposing no restrictions on such a dependency. For instance, we do *not* assume that β_i is constant with respect to p (homogeneous material). When emphasizing such a dependency we write $\beta(v, l; p)$.

In addition, reflectance (BRDF) and geometry (shape and pose) are properties of each object that can change over time. So, in principle, we would want to allow $\beta_i,\ S_i,\ g_i$ to be functions of time. In practice, we will assume that the material of each object does not change, but only its shape, pose and of course illumination. Therefore, we will use

$$S_i = S_i(t);\ g_i = g_i(t), \quad t \in [0, T] \tag{6}$$

to describe the **dynamics** of the scene. The index t can be thought of as *time*, in case a sequence of measurements is taken at adjacent instants or continuously in time, or it can be thought of as an *index* if disparate measurements are taken under varying conditions (shape and pose). Note that, as we mentioned, the light source (L, dE) can also change over time. When emphasizing such a dependency we write $L(t)$ and $dE(q, l; t)$.

Example 1 *The simples surface S_i one can conceive of is a plane: $S_i = \{p \in \mathbb{R}^3 \mid \langle \nu_i, p \rangle = d_i\}$ where ν_i is the unit normal to the plane, and d_i is its distance to the origin. For a plane not intersecting the origin, $1/d$ can be lumped into ν, and therefore three numbers are sufficient to completely describe the surface in the inertial reference frame. In that case we simply have S_i a constant, and $g_i = e$, the identity. A simple light source is an ideal point source, which can be modeled as $L \in \mathbb{R}^3$ with infinite power density $dE = E_l \delta(q - L)$. Another common model is a constant ambient illumination, which can be modeled as a sphere $L = \mathbb{S}^2$ with $dE = E_0 dL$. We will examples of various models for the BRDF later.*

Figure 3: *A complex shape (woven thread) with simple reflectance (homogeneous albedo), or a simple shape (a smooth surface) with complex reflectance (texture)?*

Remark 2 (Choosing a level of granularity in the representation) *Note that by assuming that the world is made of surfaces we are already imposing significant restrictions, and we are implicitly choosing a level of description for our representation. Consider for instance the fabric shown in Figure 3. There is no surface there. The fabric is made of thin one-dimensional threads, just woven tightly enough to give the impression of spatial continuity. Therefore, we choose to represent them as a smooth surface. Of course, the variation in the appearance due to the fine-scale structure of the threads has to be captured somehow, and we delegate this task to the reflectance model. Naturally, one could even describe each individual thread as a cylindrical surface modeled as an object S_i, but this is well beyond the detail that we want to capture. Figure 3 highlights the modeling tradeoff between shape and reflectance: one could model the fabric as a very complex object (woven thread) made of homogeneous material (wool), or as a very simple object (a smooth surface) made of textured material. This is a modeling choice, and there is no right or wrong answer.*

Remark 3 (Tradeoff between shape and motion) *We note that, instead of allowing the surface S_i to deform arbitrarily in time via $S_i(t)$,* and *moving rigidly in space via $g_i(t) \in SE(3)$, we can lump the motion and deformation into $g_i(t)$ by allowing it to belong to a more general class of deformations G, for instance diffeomorphisms, and let S_i be constant. Alternatively, we can lump the deformation $g_i(t)$ into S_i and just describe the surface in the inertial reference frame via $S_i(t)$. This can be done with no loss of generality, and it reflects a fundamental tradeoff modeling the interplay between shape and motion [20].*

Now, if we agree that a scene can be described by its *geometry, photometry* and *dynamics*, we must decide how these relate to the measured images.

1.3 And how are the two related?

Given a description of the geometry, photometry and dynamics of a scene, a model of the image is obtained through a description of the *imaging device*. An imaging device is a series of elements designed to direct light propagation. This is typically modeled through diffraction, reflection, and refraction. We will ignore the first two propagation effects, and only consider the effects of refraction. For simplicity, we can also assume that the set of objects that act as light sources and those that act as light sinks are disjoint, so that $S_i \cap L = \emptyset$, i.e. we ignore inter-reflections. In that case, we can just lump all the objects into one, which we call the scene $S \doteq \cup_{i=1}^{N_o} S_i$ with its corresponding BRDF, $\beta = \cup_{i=1}^{N_o} \beta_i$. Note that S needs not be simply connected.

Now, using the notation introduced in the previous subsection, we want to determine the energy that impinges on a given pixel as a function of the shape of the scene S, its BRDF β, the light source L and its energy distribution dE, and the position and orientation of the camera. We do so in two steps. First we compute the power radiated from a given neighborhood of the light source L to a given neighborhood on the surface S. Then we compute the portion of such power that is measured at a given pixel $\mathbf{x}$.

For simplicity, given the tradeoff between shape and motion discussed in remark 3, we describe the (possibly time-varying) shape of the scene in the inertial frame and drop the explicit description of its pose. In fact, to further simplify the notation, we can choose the inertial frame to coincide with the position and orientation of the viewer at time $t = 0$, so that if $I(\mathbf{x}_0, 0)$ is the first image, then the scene can be described as a surface parameterized by $\mathbf{x}_0$: $S(\mathbf{x}_0, t)$. We then describe the position and orientation of the camera at time t relative to the camera at time 0 using a moving Euclidean reference frame[4] $g(t) \in SE(3)$.

Vanilla radiometry

This section can be skipped at a first reading. However, we recommend eventually going through it in order to appreciate how the various quantities come into existence.

We describe the light source using its *radiance*, $R_L(q, l)$, which indicates the power density per unit area and unit solid angle emitted at a point $q \in L$ in a given direction $l \in \mathbb{H}^2$, and is measured in [W/sterad/m^2]. This is a property of the light source. When we consider the particular direction l from a point $q \in L$ on the light source towards a point $p \in S$ on the scene, this is given by $g_{q*}(p - q) = g_q p - 0 = g_q p$. Therefore, given a solid angle $d\Omega_L$ and an area element dL on the light source, the power per solid angle and unit foreshortened[5] area radiated from a point q towards p is given by

$$R_L(q, g_q p) d\Omega_L \langle \nu_q, g_q p \rangle dL \tag{7}$$

where $g_q p \in \mathbb{H}^2$ is intended as a unit vector. Now, how big a patch dL of the light we see standing at a point p on the scene depends on the solid angle $d\Omega_S$ we are looking through. Following figure 4 we have that

$$dL = d\Omega_S \|p - q\|^2 / \langle \nu_q, l_{qp} \rangle \tag{8}$$

where we have defined $l_{qp} \doteq q - p/\|q - p\|$ and the inner product at the denominator is called *foreshortening*. Similarly, the solid angle $d\Omega_L$ shines a patch of the surface dS. The two are related by

$$d\Omega_L = \frac{dS}{\|p - q\|^2} \langle \nu_q, l_{pq} \rangle \tag{9}$$

where $l_{pq} = -l_{qp} = p - q/\|p - q\|$. Substituting the expressions of $d\Omega_L$ and dL in the previous two equations into (7), one obtains the infinitesimal power received at the point p.

Now, we want to write the portion of power exiting the surface at p in the direction of a pixel $\mathbf{x}$ through an area element dS. First, we need to write the direction of $\mathbf{x}$ in the local reference frame at p. We assume that $\mathbf{x}$ is a unit vector, obtained for instance via central perspective projection

$$\pi : \mathbb{R}^3 \longrightarrow \mathbb{S}^2; \; p \mapsto \pi(p) \doteq \mathbf{x}. \tag{10}$$

However, the point p is written in the inertial frame, while $\mathbf{x}$ is written in the frame of the camera at time t. We need to first transform $\mathbf{x}$ to the inertial frame, via $g_*(t)^{-1}\mathbf{x}$, and then express this in the local frame at p, which yields $g_{p_*}^{-1} g_*(t)^{-1}\mathbf{x}$. We call the normalized version of this vector $l_{p\mathbf{x}}(t)$. Then, we need to integrate the infinitesimal power radiated from all points on the light source through their solid angle $d\Omega_L$ against the BRDF[6], which specifies what portion of the incoming power is reflected towards $\mathbf{x}$. This yields the infinitesimal energy that p radiates in the direction of $\mathbf{x}$ through an area element dS:

$$R_S(p, \mathbf{x}) dS(p) = \int_L \beta(l_{p\mathbf{x}}(t), g_p q) R_L(q, g_q p) d\Omega_L(q) \langle \nu_q, g_q p \rangle dL(q) \tag{11}$$

[4]This can be confusing at first: $g(t)$ from now on indicates the pose of the camera, and has nothing to do with the scene. Earlier we used $g_i(t)$ to describe the pose of various surfaces within the scene, but since we now lump all g_i into S we no longer have an explicit description of the pose of objects in the scene, which are always referred to the inertial reference frame.

[5]If the area element on the light source is dL, the portion of the area seen from p is given by $\langle \nu_q, g_q p \rangle dL$; this is called the *foreshortened area*.

[6]The following equation, which specifies that the scene radiance is a linear transformation of the scene radiance via the BRDF is merely a model, and not something that can be proven. Indeed this equation is often used to define the BRDF.

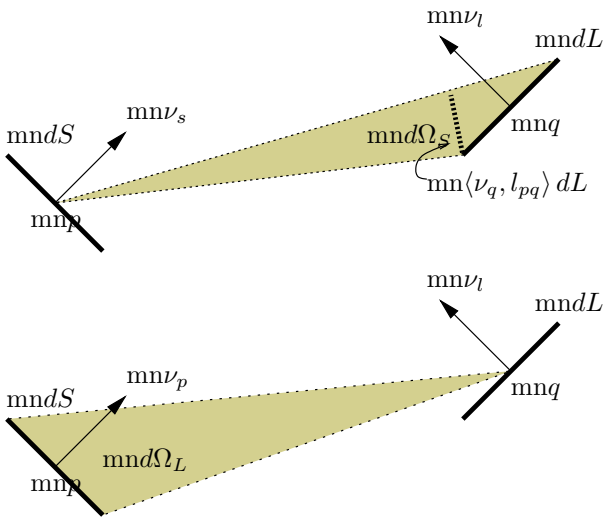

Figure 4: Energy balance: a light source patch dL radiates energy towards a surface patch dS. Therefore, the power injected in the solid angle $d\Omega_L$ by dL equals the power received by dS in the solid angle $d\Omega_S$. Equation (9) expresses this balance in symbols.

where the arguments in the infinitesimal forms $dS, dL, d\Omega_L$ indicate their dependency. Now, we can substitute[7] the expression of $d\Omega_L$ from (9) and simplify the area element dS, to obtain the *radiance* of the surface at p

$$R_S(p, \mathbf{x}) = \int_L \beta(l_{p\mathbf{x}}(t), g_p q) R_L(q, g_q p) \frac{\langle \nu_q, g_q p \rangle}{\|p - q\|^2} \langle \nu_p, l_{pq} \rangle dL(q) \tag{12}$$

Since the norm the norm $\|p - q\|$ is invariant to Euclidean transformations, we can write it as $\|g_q p\|$. Now, if the size of the scene is small compared to its distance to the light, this term is almost constant, and therefore the measure

$$dE(q, g_q p) \doteq R_L(q, g_q p) \frac{\langle \nu_q, g_q p \rangle}{\|g_q p\|^2} dL(q) \tag{13}$$

can be thought of as a property of the light source. Since we cannot untangle the contribution of R_L from that of dL, we just choose dE to describe the power distribution radiated by the light source. Therefore, we have

$$R_S(p, \mathbf{x}) = \int_L \beta(l_{p\mathbf{x}}(t), g_p q) \langle \nu_p, l_{pq} \rangle dE(q, g_q p). \tag{14}$$

This is the portion of power per unit area and unit solid angle radiated from a point p on a reflective surface towards a point $\mathbf{x}$ on the image at time t. The next step consists of quantifying what portion of this energy gets absorbed by the pixel at location $\mathbf{x}$. This follows a similar calculation, which we do not report here, and instead refer the reader to [8] (page 208). There, it is argued that the irradiance at the pixel $\mathbf{x}$ is equal to the radiance at the corresponding point p on the scene, up to an approximately constant factor, which we lump into R_S. The point p and its projection $\mathbf{x}$ onto the image plane at time t are related by the equations

$$\mathbf{x} = \pi(g(t)p) \quad p = g(t)^{-1} \pi_S^{-1}(\mathbf{x}) \tag{15}$$

where $\pi_S^{-1} : \mathbb{S}^2 \to \mathbb{R}^3$ denotes the inverse projection, which consists in scaling $\mathbf{x}$ by its depth $Z(\mathbf{x})$ in the current reference frame, which naturally depends on S. Therefore, the equation below, known as the

[7] Most often in radiometry one performs the integral above with respect to the solid angle $d\Omega_S$, rather than with respect to the light source. For those that want to compare the expression of the radiance R_S with that derived in radiometry, it is sufficient to substitute the expressions of dL and $d\Omega_L$ above, to obtain $R_S(p, \mathbf{x}) = \int_{\mathbb{H}^2} \beta(l_{p\mathbf{x}}(t), g_p q) R_L(q, g_q p) \langle \nu_p, g_q p \rangle d\Omega_S(p)$. In our context, however, we are interested in separating the contribution of the light and the scene, and therefore performing the integral on L is more appropriate.

irradiance equation, takes the form

$$I(\mathbf{x},t) = R_S(p, \pi(g(t)p)) = R_S(g(t)^{-1}\pi_S^{-1}(\mathbf{x}), \mathbf{x}). \tag{16}$$

After we substitute the expression of the radiance (14), we have the *imaging equation*, which we summarize in the next subsection for those who decided to skip this one.

1.4 The imaging equation

Summarizing the derivation in the previous subsection, we see that the intensity (irradiance) measured at a pixel $\mathbf{x}$ on the image indexed by t is given by

$$\boxed{\begin{cases} I(\mathbf{x},t) = \int_L \beta(l_{p\mathbf{x}}(t), g_p q)\langle \nu_p, l_{pq}\rangle dE(q, g_q p); \\ \mathbf{x} = \pi(g(t)p);\ p \in S \end{cases}} \tag{17}$$

where the symbols above are defined as follows:

Notation: In the equation above, we have defined $l_{p\mathbf{x}} \doteq g_{p_*}^{-1} g_*(t)^{-1}\mathbf{x}$, g_p and g_{p_*} are defined by equation (3) and (4) respectively, $l_{pq} \doteq p - q/\|p-q\|$ and $g_p q$ indicates the (normalized) direction from p to q, and similarly for $g_q p$;

Light source: $L \subset \mathbb{R}^3$ is the (possibly time-varying) collection of light sources emitting energy with a distribution $dE : L \times \mathbb{H}^2 \to \mathbb{R}_+$ at every point $q \in L$ towards the direction of a point p on the

Scene: a collection of (possibly time-varying) piecewise smooth surfaces $S \subset \mathbb{R}^3$; $\beta : \mathbb{H}^2 \times \mathbb{H}^2 \times S \to \mathbb{R}$ is the bidirectional reflectance distribution function (BRDF) that depends on the incident direction, the reflected direction and the point $p \in S$ on the scene S and is a property of the material.

Motion: relative motion between the scene and the camera is described by the motion of the camera $g(t) \in SE(3)$ and possibly the action of a more complex group G, or simply by allowing the surface $S(t)$ to change over time.

Projection: $\pi : \mathbb{R}^3 \mapsto \mathbb{S}^2$ denotes ideal (pinhole) perspective projection, modeled here as projection onto the unit sphere, although the same model applies if $\pi : \mathbb{R}^3 \to \mathbb{P}^2$, in which case $l_{p\mathbf{x}}$ has to be normalized to unit length.

Visibility and cast shadows: One should also add to the equation two characteristic function terms: $\chi_v(\mathbf{x},t)$ outside the integral, which models the visibility of the scene from the pixel $\mathbf{x}$, and $\chi_s(p,q)$ inside the integral to model the visibility of the light source from a scene point (cast shadows). We are omitting these terms here for simplicity. However, in some cases that we discuss in the next section, discontinuities are the only source of visual information.

Remark 4 (A philosophical aside on scene modeling) *One could argue that the* real *world cannot be captured by simple mathematical models of the type just described, and even classical physics is largely inadequate for the task. However, we are not looking for an* absolute *model. Instead, we are looking to describe the scene at the level of granularity that is suitable for us to be able to perform inference and accomplish certain spatial tasks. So, what is the "right" granularity? For us a suitable model of the scene is one that can be validated with other existing sensing modalities, for instance touch. This is well illustrated by the fabric of Figure 3, where at the level of granularity required the scene can be safely described as a smooth surface.*

Notice that this is similar to what other researchers have suggested by describing the scene as a functional *that cannot be directly measured. However, such a functional can be* evaluated *with various test-functions. Physical instruments provide a set of test functions, and imaging device provide yet another set of test functions. The goal of the imaging model, therefore, can be thought of as relating the value of the scene functional obtained by probing with physical instruments to the value obtained by probing with images.*

2 Special cases of the imaging equation and their role in visual reconstruction

The imaging equation is relevant because most of computer vision is about inverting it; that is, inferring properties of the scene (shape, material, motion) regardless of pose, illumination and other nuisances (the visual reconstruction problem). However, it its general formulation above, the imaging equation cannot be inverted. Therefore, it is common to make assumptions on some of the unknowns to recover the others. In this section we aim at enumerating a collection of special cases that compounded characterize most of what can be done in visual inference. We start with models of reflection.

2.1 Empirical reflectance models

Most common materials can be described by a BRDF. Exceptions include translucent materials (e.g. skin), anisotropic material (e.g. brushed aluminum), micro-structured material (e.g. hair) etc. However, since our goal is not realism in a physical simulation, we are content with some common BRDF that are well established in computer graphics: Phong (corrected) [17], Ward [22] and Torrance-Sparrow (simplified) [21].

Phong (corrected) $\beta(v,l) = \rho_d(p) + \rho_s(p)\cos^c\delta/\cos\theta_i\cos\theta_o$.
Here $\cos\delta = \langle g(t)^{-1}\mathbf{x} + q/\|q\|, \nu_p\rangle$ where each term in the inner product is normalized, and $\theta_i \doteq \arccos\langle l, \nu_p\rangle$, and $\arccos(\theta_o) \doteq \langle v, \nu_p\rangle$; $c \in \mathbb{R}$ is a coefficient that depends on the material.

Ward $\beta(v,l) = \rho_d(p) + \rho_s(p)\frac{\exp(-\tan^2(\delta)/\alpha^2)}{\sqrt{\cos\theta_i\cos\theta_o}}$.
Here $\alpha \in \mathbb{R}$ is a coefficient that depends on the material and is determined empirically.

Torrance-Sparrow (simplified) $\beta(v,l) = \rho_d(p) + \rho_s(p)\frac{\exp(-\delta^2/\alpha^2)}{\cos\theta_i\cos\theta_o}$.

Separable radiance As Nayar and coworkers point out [], the radiance for the latter model can be written as the sum of products, where the first factor depends solely on material (diffuse and specular albedo), whereas the second factor compounds shape, pose and illumination.

In all these cases, $\rho_d(p)$ is an unknown function called *(diffuse) albedo*, and $\rho_s(p)$ is an unknown function called *specular albedo*. Diffuse albedo is often called just albedo, or, improperly, *texture*.

Note that the first term (diffuse reflectance) is the same in all three models. The second term (specular reflectance) is different. Surfaces whose reflectance is captured by the first term are called Lambertian, and are by far the most studied in computer vision. The following list reflects the organization of the following subsections, and illustrates a taxonomy of illumination and reflectance models.

- diffuse reflection (Lambert)
  - constant illumination
    - ambient light
      - constant albedo: silhouettes
      - smooth albedo: stereoscopic segmentation
      - piecewise constant/smooth albedo: region-based segmentation on surfaces
      - nowhere constant albedo: multi-view stereo
    - point light
      - constant albedo: stereoscopic shading
      - general albedo: multi-view stereo
    - general light: multi-view stereo and the correspondence problem
  - constant viewpoint: photometric stereo
- diffuse/specular reflection (P/W/T-S)
  - constant illumination

    - ∗ ambient light: back to Lambert (almost)
    - ∗ point lights: inverse global illumination
    - ∗ general lights: multi-view stereo beyond Lambert
  - constant viewpoint: photometric non-Lambertian stereo
  - reciprocal viewpoint/illumination: Helmholtz stereopsis

2.2 Lambertian reflection

Lambertian surfaces essentially look the same regardless of the viewpoint: $\beta(v,l) = \beta(w,l)\ \forall w \in \mathbb{H}^2$. This yields to major simplifications of the image formation model. Moreover, in the case of constant illumination, it allows relating different views of the same scene to one another directly, bypassing the image formation model. This is known as the *correspondence problem*, which relies crucially on the Lambertian assumption and the resulting brightness constancy constraint.[8] We address this case first.

2.2.1 Constant illumination

In this case we have $L(t) = L$ and $dE(q,l;t) = dE(q,l)$. We consider two simple light source models first.

Ambient light

Ambient light is due to inter-reflection between different surfaces in the scene. Since modeling such inter-reflections is quite complicated,[9] we will approximate it by assuming that there is a constant amount of energy that "floods" the ambient space. This can be approximated by a sphere radiating constant energy: $L = \mathbb{S}^2$ and $dE = E_0 dL$. In this case, the imaging equation reduces to

$$I(\mathbf{x},t) = \rho_d(p) E_0 \int_{\mathbb{S}^2} \langle \nu_p, l \rangle d\Omega(l) \tag{18}$$

Due to the symmetry of the light source, assuming there are no shadows, we can always change the global reference frame so that $\nu_p = e_3$; therefore, the integral does not depend on p, and is a constant that, together with E_0, can be lumped into ρ_d, yielding the simplest possible model that, when written with respect to a moving camera, gives

$$\boxed{\begin{cases} I(\mathbf{x},t) = \rho(p) \\ \mathbf{x} = \pi(g(t)p); \quad p = S(\mathbf{x}_0). \end{cases}} \tag{19}$$

Note that this model effectively neglects illumination, for one can think of a scene S that is self-luminous, and radiates an equal amount of energy $\rho(p)$ in all directions. Even for such a simple model, however, performing visual inference is non-trivial. It has been done for a number of special cases:

Constant albedo: silhouettes When $\rho(p)$ is constant, the only information in equation (19) is at the discontinuities between $\mathbf{x} = \pi(g(t)p), p \in S$ and $p \notin S$, i.e. at the occluding boundaries. Given suitable conditions, that have been first studied by Aström et al. [1], motion $g(t)$ and shape S can be recovered. The reconstruction of shape S and albedo ρ has been addressed in an infinite-dimensional optimization framework by Yezzi and Soatto [23, 24] in their work on stereoscopic segmentation.

Smooth albedo The stereoscopic segmentation framework has been extended to allow the albedo to be smooth, rather than constant. The algorithm in [12] provides an estimate of the shape of the scene S as well as its albedo $\rho(p)$ given its motion relative to the viewer, $g(t)$.

Piecewise constant/piecewise smooth albedo The same framework has been recently extended to allow the albedo to be piecewise constant in [10]. This amount to performing region-based segmentation a' la Mumford-Shah [15] on the scene surface S. Although it has not been done yet, the same ideas could be extended to piecewise smooth albedo.

[8] Although the constraint is often used *locally* to approximate surfaces that are *not* Lambertian.
[9] There is some admittedly sketchy evidence that inter-reflections are not perceptually salient [4].

Nowhere constant albedo When $\nabla\rho(p) \neq 0$ everywhere in p, the image formation model can be bypassed altogether, leading to the so-called correspondence problem which we will see shortly. This is at the base of most traditional stereo reconstruction algorithms and structure from motion. Since these techniques apply without regard to the illumination, we will address this after having relaxed our assumptions on illumination.

Point light(s)

A countable number of stationary point light sources can be modeled as $L = \{L_1, L_2, \ldots, L_k\},\ L_i \in \mathbb{R}^3$, $dE = \sum_{i=1}^{k} E_i \delta(q - L_i)$. In this case the imaging equation reduces to

$$I(\mathbf{x}, t) = \sum_{i=1}^{k} E_i \rho_d(p) \langle \nu_p, p - L_i / \|p - L_i\| \rangle. \tag{20}$$

Note that, if we neglect occlusions and cast shadows, the sum can be taken inside the inner product and therefore there is no loss of generality in assuming that there is only one light source. If the light sources are at infinity, p can be dropped from the inner product; furthermore, the intensity of the source E multiplies the light direction, so the two can be lumped into the vector L. We can therefore further simplify the above model to yield, taking into account camera motion,

$$\boxed{\begin{cases} I(\mathbf{x}, t) = \rho(p) \langle \nu_p, L \rangle \\ \mathbf{x} = \pi(g(t)p); \quad p = S(\mathbf{x}_0) \end{cases}} \tag{21}$$

Inference from this model has been addressed for the following cases.

Constant albedo Yuille et al. [26] have shown that given enough viewpoints and lighting positions one can reconstruct the shape of the scene. Jin et al. [10] have proposed an algorithm for doing so, which estimates shape, albedo and position of the light source in a variational optimization framework. If the position of the light source is known and there is no camera motion, this problem reduces to classical shape from shading [7].

Smooth/piecewise smooth albedo In this case, one can easily show that albedo and light source cannot be recovered since there are always combinations of the two that generate the same images. However, under suitable conditions shape can still be estimated, as we discuss next.

Nowhere constant radiance If the combination of albedo and the cosine term (the inner product in (21)) result in a radiance function that has non-zero gradient, we can think of the radiance as an albedo under ambient illumination, and therefore this case reduces to multi-view stereo, which we will discuss shortly. Naturally, in this case we cannot disentangle reflectance from illumination, but under suitable conditions we can still reconstruct the shape of the scene, as we discuss shortly in the context of the correspondence problem.

Cast shadows If the visibility terms are included, under suitable conditions about the shape of the object and the number and nature of light sources, one can reconstruct an approximation of the shape of hte scene.

General light distribution

An arbitrary distant light distribution can be modeled as a positive density on the sphere at infinity: $L = \mathbb{S}^2$. Any positive density on the sphere can be approximated arbitrarily well by a sum of Gaussians, a result known to Wiener, slightly modified to take into account the spherical ambient space. However, each Gaussian can be represented as a convolution of a delta measure with a canonical Gaussian (zero-mean). When inserted into the imaging equation, the effect of the Gaussian kernel and the BRDF compound in a way that cannot be discerned from the data alone. Therefore, we can lump the Gaussian kernel into the BRDF and be left with point light sources with no loss of generality. Naturally, in practice each each light may have a different

dispersion matrix, which in general results in an empirical coefficient (α in the Torrance-Sparrow model) that is direction-dependent. If we allow the BRDF to be anisotropic, we can never distinguish reflectance from illumination. Consider for instance a polished sphere illuminated by a Gaussian light sources, compared to a rougher sphere illuminated by a point.

Note that most current work on general representation of illumination uses a series expansion of the distribution dE on $L = \mathbb{S}^2$ into spherical harmonics [18]. This is problematic for two reasons: first, spherical harmonics are *global*, so the introduction of another term in the series affects the entire image. Second, while any function on the sphere can be approximated with spherical harmonics, there is no guarantee that such a function be *positive*. Indeed, the harmonic terms in the series are themselves not positive, and therefore each individual component does not lend itself to be interpreted as a valid illumination, and there is no guarantee except in the limit where the number of terms goes to infinity that the truncated series will be a vaild illumination. The advantage of a sum of Gaussian approximation is that one can approximate any positive function, and given any truncation of the series one is guaranteed to have a positive distribution dE.

Given these considerations, we restrict our attentions to illumination models that consist of the sum of a constant ambient term and a countable number of point light sources. The general case, therefore, reduces to the special cases seen above:

$$L = \mathbb{S}^2; \quad dE(q) = E_0 dL(q) + \sum_{i=1}^{k} E_i \delta(q - L_i). \tag{22}$$

Note that the energy does not depend on the direction, since for distant lights (sphere of infinite radius) all directions pointing towards the scene are normal to L.

Multi-view stereo and the correspondence problem

If the radiance of the scene $R_S(p)$ is not constant, under suitable conditions one can do away with the image formation model altogether. Consider in fact the irradiance equation (16). Under the Lambertian assumption, given (at least) two viewpoints, indexed by t_1 and t_2, we have that

$$I(\mathbf{x}_1, t_1) = R_S(p, \pi(g(t_1)p)) = I(\mathbf{x}_2, t_2) \tag{23}$$

without regards to how the radiance R_S comes to existence. The relationship between $\mathbf{x}_1$ and $\mathbf{x}_2$ depends solely on the shape of the scene S and the relative motion of the camera between the two time instants, $g_{12} \doteq g(t_1)g(t_2)^{-1}$:

$$\mathbf{x}_1 = \pi(g_{12}\pi_S^{-1}(\mathbf{x}_2)) \doteq w(\mathbf{x}_2; S, g_{12}). \tag{24}$$

Therefore, one can forget about how the images are generated, and simply look for the function w that satisfies (substitute the last equation into the previous one)

$$\boxed{I(w(\mathbf{x}_2; S, g_{12}), t_1) = I(\mathbf{x}_2, t_2)}. \tag{25}$$

Finding the function w from the above equation is known as the *correspondence problem*, and the equation above is known as the *brightness constancy constraint*.

It is easy to see that if the gradient of I is zero around a given point $\mathbf{x}$, the equation above is satisfied for any S, g, and therefore it provides no useful constraint on w. Similarly, if I is a periodic function, w is not a one-to-one function. Even if I was different at every pixel, due to noise, deviation from Lambertian reflection and other accidents, the equation above is not per se a useful vehicle to recover S and g_{12}. Nevertheless, the equation is broadly employed in computer vision, mostly due to its simplicity. The most common approach consists of comparing the left-hand side to the right-hand side of the equation via some discrepancy measure K, integrated in a neighborhood of a collection of "feature" points, assuming that w can be approximated within each such neighborhood with a simple parametric function, typically an affine transformation:

$$\phi(w(S, g_{jk})) = \sum_{j,k} \int_{\cup_i W(\mathbf{x}_i; A)} K(I(\mathbf{x}, t_j), I(w(\mathbf{x}; S, g_{jk}), t_k)) d\mathbf{x} \tag{26}$$

where $W(\mathbf{x}; A)$ is a neighborhood of the pixel $\mathbf{x}$, which can be parameterized by an affine transformation A, and K is a discrepancy measure, such as the L^1 or L^2 norm, normalized cross-correlation, or Kullback-Leibler divergence. *Feature points* $\mathbf{x}_i$ can then be defined as those for which minimizing the functional ϕ allows the inference of the transformation parameters in w. Various "robust" versions of this program have been investigated. They consists of comparing *not* the intensity of neighborhood of the images directly, but instead various *statistics* of the images, such as the sum of square differences, various gradient orientation or color histograms, normalized to account for local geometric of photometric variation:

$$\phi(w(S, g_{jk})) = \sum_{j,k} \|F(\{I(\mathbf{x}, t_j) \mid \mathbf{x} \in W(\mathbf{x}_i)\}) - F(\{I(\mathbf{x}, t_k) \mid \mathbf{x} \in W(w(\mathbf{x}_i; S, g_{jk}))\})\| \tag{27}$$

for some choice of image statistic F. This is essentially what is done in the majority of the stereo reconstruction algorithms, as well as in structure from motion (see [5] and references therein). Given enough feature points, the conditions under which one can reconstruct the motion of the camera and the position of feature points are well known [14].

More recently, Faugeras and Keriven have cast the problem of stereo reconstruction in an infinite-dimensional optimization framework, where the equation above is integrated over the entire image, rather than just in a neighborhood of feature points, and the correspondence function w is estimated implicitly by estimating the shape of the scene S, with a given motion g. This works even if ρ is constant, but due to a non-uniform light and the presence of the Lambertian cosine term (the inner product in equation (21)) the radiance of the surface is nowhere constant (shading effect, or attached shadow) and even in the case of cast shadows, if the light does not move. In the presence of regions of constant radiance, the algorithm interpolates in ways that depend upon the regularization term used in the infinite-dimensional optimization (see [6] for more details).

2.2.2 Constant viewpoint: photometric stereo

When the viewpoint is fixed, but the light changes, inverting the model above is known as photometric stereo [8]. If the light configuration is not known and is allowed to change between views, Belhumeur and coworkers have shown that this problem cannot be solved [2]. In particular, given two images one can pick a surface S at will, and construct two light distributions that generate the given images, even if the scene is known to be Lambertian. However, this result relies on the presence of a single point light source. We conjecture that if the illumination is allowed to contain an ambient term, these results do not apply, and therefore reconstruction could be achieved. Note that psychophysical experiments suggest that face recognition is extremely hard for humans under a point light source, whereas a more complex illumination term greatly facilitates the task.

2.3 Non-Lambertian reflection

In this subsection we relax the assumption on reflectance. While, contrary to intuition, a more complex reflectance model can in some cases facilitate recognition, in general it is not possible to disentangle the effects of shape, reflectance and illumination. We start by making assumptions that follow the taxonomy used for the Lambertian case in the previous subsection.

2.3.1 Constant illumination

Ambient light

In the presence of ambient illumination, the specular term of an empirical reflection model, for instance Phong's, takes the form

$$\rho_s(p) \int_{-\pi}^{\pi} \int_{0}^{\pi/2} \frac{\cos^k \delta}{\cos \theta_o} \sin \theta_i d\theta_i d\phi_i \tag{28}$$

If the exponent $c \to \infty$, only one point on the light surface $\mathbb{S}^2$ contributes to the radiance emitted from the point p. Since the distribution dE is uniform on L, we conclude that, if we exclude occlusions and cast

shadows, this term is a constant. This can be considered as a limit argument to conjecture that, in the presence of ambient illumination, the specular term is negligible compared to the diffuse albedo. Naturally, if an object is perfectly specular, it renders the viewer an image of the light source, so in this case inter-reflection is the dominant contribution, and the ambient illumination approximation is no longer justified. See for instance figure 5.

Figure 5: In the presence of strongly specular materials, the image is essentially a distorted version of the light source. In this case, modeling inter-reflections with an ambient illumination term is inadequate.

Point light(s)

In the presence of point light sources, the specular component of the Phong models becomes

$$\sum_i E_i \rho_s(p) \frac{\left\langle g^{-1}(t)\mathbf{x} + L_i/\|L_i\|, \nu_p \right\rangle^c}{\langle g^{-1}(t)\mathbf{x}, \nu_p \rangle} \tag{29}$$

where the arguments of the inner products are normalized. In this case, assuming that a portion of the scene is Lambertian and therefore motion and shape can be recovered, one can invert the equation above to estimate the position and intensity of the light sources. This is called "inverse global illumination" and was addressed by Yu and Malik [25]. If the scene is dominantly specular, so no correspondence can be established from image to image, we are not aware of any general result that describes under what condition shape, motion and illumination can be recovered. Savarese and Perona [19] study the case when assumptions on the position and density of the light, such as the presence of straight edges at known position, can be exploited to recover shape.

General light

In general, one cannot separate reflectance properties of the scene with distribution properties of the light sources. Jin et al. [11] showed that one can recover shape S as well as the radiance of the scene, which mixes the effects of reflectance and illumination.

2.3.2 Constant viewpoint

In the presence of multiple point light sources, Many have studied the conditions under which one can recover the position and intensity of the light sources, see for instance [16] and references therein. Variations

of photometric stereo have also been developed for this case, starting from [9].

2.3.3 Reciprocal viewpoint and light source

Zickler et al. [27] have developed techniques to exploit a very peculiar imaging setup where a point light source and the camera are switched in pairs of images, which allows to eliminate the BRDF from the imaging equation.

3 Conclusions: what does recognition have to do with all this?

In practice, to be able to determine whether an object is present in the scene, or whether two or more images portray the same scene, reconstructing the shape, motion and reflectance of the scene may not be necessary. However, it is important to understand the image formation process because, in general, no function of the image can be computed that is invariant with respect to all nuisance factors, such as viewpoint, illumination, material properties etc. Therefore, any recognition system will have to rely on assumptions on some of the nuisances or prior models derived from data. The imaging equation allows the researcher to elucidate these assumptions and ultimately is aimed at designing better recognition systems.

References

[1] K. Astrom, R. Cipolla, and P. J. Giblin. Motion from the frontier of curved surfaces. pages 269–275, 1995.

[2] P. Belhumeur, D. Kriegman, and A. L. Yuille. The generalized bas relief ambiguity. *Int. J. of Computer Vision*, 35:33–44, 1999.

[3] A. Do Carmo. *Differential Geometry of Curves and Surfaces*. Prentice Hall, 1976.

[4] J. Enns and R. Rensink. Influence of scene-based properties on visual search. *Science*, 247:721–723, 1990.

[5] O. Faugeras. *Three dimensional vision, a geometric viewpoint.* MIT Press, 1993.

[6] O. D. Faugeras and R. Keriven. Variational principles, surface evolution pdes, level set methods and the stereo problem. *INRIA Technical report*, 3021:1–37, 1996.

[7] B. Horn and M. Brooks (eds.). *Shape from Shading*. MIT Press, 1989.

[8] B. K. P. Horn. *Robot vision*. MIT press, 1986.

[9] K. Ikeuchi. Determining surface orientations of specular surfaces by using the photometric stereo method. *IEEE Trans. on Pattern Analysis and Machine Intelligence,*, 3:661–669, 1981.

[10] H. Jin, D. Cremers, A. Yezzi, and S. Soatto. Shedding light in stereoscopic segmentation. In *Proc. of the IEEE Intl. Conf. on Comp. Vis. and Patt. Recog.*, (submitted) 2004.

[11] H. Jin, S. Soatto, and A. J. Yezzi. Multi-view stereo beyond lambert. 2003.

[12] H. Jin, R. Tsai, L. Chen, A. Yezzi, and S. Soatto. Estimation of 3d surface shape and smooth radiance from 2d images: A level set approach. *J. of Sci. Comp.*, 19(1-3):267–292, 2003.

[13] D. G. Kendall. Shape manifolds, procrustean metrics and complex projective spaces. *Bull. London Math. Soc.*, 16, 1984.

[14] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An invitation to 3D vision, from images to models*. Springer Verlag, 2003.

[15] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. on Pure and Applied Mathematics*, 42:577–685, 1989.

[16] S. Nayar, K. Ikeuchi, and T. Kanade. Surface reflection: physical and geometrical perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(7):611–634, 1991.

[17] B. T. Phong. Illumination for computer generated pictures. In *Communications of the ACM*, volume 18(6), pages 311–317, 1975.

[18] R. Ramamoorthi and P. Hanrahan. On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object. *J. Opt. Soc. Am.*, pages 2448–2459, Oct. 2001.

[19] S. Savarese and P. Perona. Local analysis for 3d reconstruction of specular surfaces. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, 2001.

[20] S. Soatto and A. Yezzi. Deformotion: deforming motion, shape average and the joint segmentation and registration of images. In *Proc. of the Eur. Conf. on Computer Vision (ECCV)*, volume 3, pages 32–47, 2002.

[21] K. E. Torrance and E. M. Sparrow. Theory for off-specular reflection from roughed surfaces. *J. of the Opt. Soc. of Am.*, 57(9):1105–1114, 1967.

[22] G. Ward. Measuring and modeling anisotropic reflection. In *SIGGRAPH*, pages 265–272, 1992.

[23] A. Yezzi and S. Soatto. Stereoscopic segmentation. In *Proc. of the Intl. Conf. on Computer Vision*, pages 59–66, 2001.

[24] A. J. Yezzi and S. Soatto. Structure from motion for scenes without features. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, pages I–525–532, June 2003.

[25] Y. Yu, P. Debevec, J. Malik, and T. Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proc. of the AMS SIGGRAPH*, 1999.

[26] A. Yuille, J. M. Coughlan, and S. Konishi. Kgbr viewpoint-lighting ambiguity. *J. Opt. Soc. Am. A*, 20(1):(in press), 2003.

[27] T. Zickler, P. N. Belhumeur, and D. J. Kriegman. Helmholtz stereopsis: exploiting reciprocity for surface reconstruction. In *Proc. of the ECCV*, pages 869–884, 2002.