Chapter 1
Is Human Vision Any Good?

Jan J. Koenderink

Abstract. Human vision is often referred to as an "existence proof" for challenging targets of machine vision. But in some areas machine vision evidently "beats" human vision, so the questions arise: Is human vision any good, will it be supplanted by machine vision for most tasks soon? I analyze human vision with the aim to provide an answer to such questions. Does machine vision still have anything to learn from human vision? I identify a number of basic principles of biological vision that are likely to be of interest to the machine vision community.

1.1 Introduction

On May 11th, 1997 at 3:00PM EDT, game #6 of the match of Garry Kasparov, the greatest (human!) player in the history of chess, against IBM's Deep Blue Supercomputer put the final score at `[kasparov 2.5: deep blue 3.5]`. So much for chess [1]. Are human chess players any good? Is vision like chess in this respect?

In order to answer the question one needs to make more precise what is exactly being meant by "human vision" and identify potentially interesting aspects. In an attempt to learn from human vision one needs an overview of the various hardware and algorithmic structures that implement human vision. One might (naïvely) believe that it would suffice to ask physiologists and psychologists for the required information. Such is not the case. The best that can be done is to offer creative guesses on the basis of a panoramic knowledge of these fields, framed in the language understood by engineers. This is my aim in this chapter.

1.1.1 What Is "Human Vision"?

"Human vision" (or, more general, "biological vision") stands for an extensive bundle of human or animal capabilities of mutually very different types. The simplest of

Jan J. Koenderink
EEMCS, Delft University of Technology, The Netherlands
e-mail: `jan.koenderink@telfort.nl`

R. Cipolla, S. Battiato, & G.M. Farinella (Eds.): Computer Vision, SCI 285, pp. 1–25.
springerlink.com

these are hardly worth the name "vision", these include many optoreflexes found in the simplest animals, but also present in man. Though vital for survival, they hardly involve anything beyond wiring up a photocell to some effector. More complicated cases of "*optically guided behavior*" include orientation/navigation on the basis of optic flow and so forth. Such mechanisms make up a major part of the "visual system". This part can be understood in the engineering sense and are already emulated in robotic systems. Here machines are going to beat man, the waiting is for the first soccer match of a human team to a robot team that will be lost by the men. In this chapter I am not so much interested in this aspect of vision.

Most of "optically guided behavior" goes on below the level of awareness. This makes sense because of the real–time aspect and the pre–cognitive nature of overt bodily actions. Just try to walk through consciously figuring out the right signals to be sent to your muscular–skeletal system on the basis of the pattern of photons impinging on your retinas and you will understand why. Here I will narrow down the definition of "vision" to the level of awareness, that is *optically induced experience*. Such experiences do not primarily influence your actual (real–time) behavior as in optically guided behavior, but they are relevant for your *potential* behavior. Another way to put this is to say that visual experiences lead to knowledge. For instance, on leaving the door you may grab an umbrella or a raincoat because an earlier (maybe much earlier) look out of the window taught you that it "looked like rain". Or you might do neither and remark "I don't mind to get wet" to a companion. Thus there is *no direct link with action* as there is in optically guided behavior.

As you open your eyes the world appears to you, you cannot voluntarily choose not to see. Visual experiences are *presentations* that *happen to you*, much like sneezing. Through your presentations (including those of your other modalities) you are aware of your world. You are visually aware of the world itself, rather than your thoughts of the world. That is why you are responsible for your thoughts but not for your presentations. Presentations are pre–cognitive. They are "world–directed" though, the technical term is "intentional".

1.1.2 How Is Human Vision Implemented?

I will mainly address the level of the brain (including the retinas) here, although the eye with its optics (cornea, lens, waveguides with photopigments), the eyeball with its muscles, as well as the musculature of head and body play major roles in everyday vision. One learns about the visual system from observations of generic humans (experimental psychology, psychophysics), patients (neuropsychiatry), and dead bodies (neuroanatomy). Nowadays one also records crude brain activity in healthy volunteers and (sporadic) epilectic patients during operations. Most of the more spatiotemporally precise observations are from electrophysiological experiments on animals, sometimes (technically) asleep, sometimes in a waking or (very rarely) even behaving state. Although huge amounts of observations are available, the brain is such a complicated structure that it is fair to say that a synthesis of all this knowledge is sadly lacking. One has to make do with lacunary data, usually

available in potentially misleading formats. What I attempt to do here is to identify some basic principles.

The global neural structure. The global structure of the visual system is made up of a large number of mutually highly (two way) interconnected areas, roughly branching out from the input stage as a hierarchical tree. As one moves into the hierarchy the resolution decreases, in the areas with considerable resolution on finds a retinotopic ordening, probably reflecting a tendency towards wiring economy. The interconnection of the areas is so high that one can do little more than identify a (very) rough division of labor. The two way wiring suggests that the processing is certainly not limited to a "bottom up" stage.

The principle of locality. Most of the wiring inside an area is short range. Most of the processing appears to be local, though with important "top down" (thus probably somewhat global) modulation.

The principle of local selectivity. Most of the short range wiring inside an area is highly selective. A common structure is "center–surround", essentially an isotropic Laplacean operator. Directional wirings are also common, essentially directional derivative operators in space or space–time. Directional units are typically connected to units of similar specificity in their environment.

The principle of global selectivity. Long range (area–to–area) wiring is (at least approximately) somatotopic, that is selective in the sense that topological structure is conserved.

Speculative interpretation. It is hard not to regard the various anatomically and electrophysiologically distinct areas as functional subunits, dedicated to one or more subtasks. The areas early in the stream (closest to the input) are best known and are likely to be dedicated to various "image processing" tasks. The units are likely to function largely independent of each other, possibly at quite different parameter settings (due to local "adaptation"). The "glue" may be provided by the top down queries and (implicitly) through the somatotopic connection to other areas.

A likely interpretation is to regard an area as a "geometry engine" [2], implementing a basis for differential geometric queries. The units world compute a "jet space" composed of directional derivatives up to a certain order (about 4 seems right), possibly at various scales. Then local algorithms might compute various differential invariants (algebraic combinations of derivatives, e.g., curvature) and local statistical measures (e.g., structure tensors) that would be meaningful entities for top down queries. The specific local wiring would be needed to implement geometrical entities like covariant derivatives, curvature tensors, and so forth. This means that a "signal" (e.g., a curvature tensor) cannot be carried by a single nerve fiber, it has to be carried by a (small) ensemble of these. This again entails that present electrophysiological methods all fail to address the meaningful structure of the areas.

The local jets (essentially truncated Taylor series describing local structure) are simply data structures "sitting there", waiting to be queried, they are not necessarily send upstream. Think of this structure as "stuck" in the area as a sample of the world

"in brain readable form". It is available, like a footprint in the sand of a beach. I see no essential difference here even though the latter is "outside" and the former "inside the head". The queries can be of various nature and may expect the jets to be available as local parameters to be used in computing an answer. What is sent upstream by the area is a mere summary account of its overall state. Thus the "bottom up" stream narrows down to what a top level might use to check the present "gist". Details are guessed, perhaps to be checked through a (focussed) stream of "top down" queries.

The global functional structure. The functional structure of the visual system (here "visual system" is meant in a functional sense distinct from the level of the neural substrate) can only be addressed via observations of the interaction of the agent with its world. Thus one needs the observations of experimental psychology or psychophysics. A few observations stand out as especially basic.

The intentionality of visual experiences. Presentations are world–directed, with the immediate implication that they cannot be computed in a purely bottom up fashion. They must be imposed by the agent, hence derive from a centrofugal ("top down" suggests a brain implementation) or "probing" process. Perception is *active*, it is an outward directed probing rather than the passive reception of "data". Without probing there would be no "data", but only structure (see below).

The rock bottom of Gestalts. The smallest parts of presentations (though not necessarily in a spatial sense) are the Gestalts. There is no way to look "into" a Gestalt. The Gestalts are to the human observer what "releasers" are to animals. They are where the buck stops in "explaining" the available structure, avoiding an infinite regress. They both limit and enable presentations.

Speculative interpretation. Much of presentation is essentially "controlled hallucination". Although the word "hallucination" has a bad ring to it, "hallucination" has the advantage to resolve the problem of world–directedness in one go. Intentionality is there from the start, no need to compute it—which is an impossibility anyway. Hallucinations can be "controlled" through comparison with the *facts*, which are *records of optical structure*. Thus the system needs top down queries in order to keep its hallucinations in tune with its world. This *modus operandus* has several advantages. The "binding problem" vanishes since nothing like "binding" is needed; lacking data are only a lack of constraint on the current hallucination; likewise inherent ambiguity (typical for "Shape From X" algorithms) simply means a partial constraint. The presentations are always complete and unique by construction (for hallucinations are constructions). The only problem is how to "hallucinate" productively. Usually this is no problem since most moments are much like the previous ones. Here the generic knowledge of the agent world ("frames", "scripts", "background", "Bayesian priors", etc.) comes into play. It enables the "gist" that seeps in from the bottom up to be used effectively. A wrong hallucination is not a disaster, the agent is almost certain to win any "twenty questions game" with its world. Think of wrong hallucinations as vehicles for learning, much like the hypotheses of scientific research (see below). In that sense the controlled hallucination implements a

selective *probing* process. When probing meets resistance due to unexpected mismatches the agent gains information from direct contact with the world in the context of the (intentional) probing. *No gain without pain:* the mismatches are required in order to be able to gain information at all.

The "Sherlock Model". An apt model for the process of "controlled hallucination" is that so expertly wielded by the famous Sherlock Holmes [3]. The "solution" to a case cannot be "computed" for even the input is undefined. Anything (e.g., a discarded cigarette butt, a hair in the soup, an odd noise at noon) can be either a valuable clue or totally irrelevant. There is no end to this as the world is infinitely structured. It is only a plot that enables Holmes to interpret random structures as clues, or even to actually look for them (is there a bloodstain on the curtain, did someone cut a rose, etc.). Moreover, the plot enables such clues to be interpreted, i.e., to become data (meaningful) instead of structure (meaningless). Questions are like computer formats in that they define the meaning of possible answers. The plot is not "computed from data" either. It is freely invented by Holmes, on the basis of his prior experience and his assessment of the "gist" of the scene. If it doesn't work Holmes discards it for another. How many possibilities are there anyway (the butler did it, or maybe the countess, etc.)? Holmes' method is not different from the way the sciences operate either. No theory was ever "computed from data"! They are freely invented by people like Einstein and checked against the observations. They lead to novel observations to be made ("probing nature") as further checks on the theory. Theories that don't check out are discarded for other (equally freely invented) ones. Thus theories, plots and presentations are subject to a thorougly Darwinnian selection process that soon weeds out ones that fail to explain the world. The ones that remain are current best bets of what the world is like, or—perhaps more apt, the currently most effective *user interfaces* of the agent (scientist, criminal investigator, visual observer, ...) to their world. To the user, the interface is what the world is like, what they *understand* (having constructed it themselves). For the user there is nothing understandable beyond the interface. In that sense perceptions are "mental paint".

The very idea of "bottom up processing" makes no sense in criminal investigation, nor in science. It makes no sense in visual perception either, despite the fact that so many professionals (from philosophy, physiology, psychology and machine vision alike) remain true believers. This (abortive) model assumes that the world is a well defined place even in the absence of any observer and that it is the task for visual systems to compute a representation of it. Thus one can conveniently assess visual systems, their representations should replicate the world in as detailed a manner as possible. Such a weird notion applies (at best) to zombies (no intentionality) in limited settings.

1.2 Frameworks

In discussing vision one has to break down the problems in more or less coherent chunks. In this chapter I discuss two generic "frameworks", but I am by no means

complete, nor did I make a serious attempt at a principled breakdown of the whole of vision. I don't doubt that such an endeavor is both possible and profitable though. I pick these two examples because they lead to a similar formal structure ("geometry"), which enables me to keep the formalism at bay.

1.2.1 The Spatial Framework

"Space", in the senses of "configuration", "shape" or "possibility of movement", is one of the very backbones of presentations. There are many directions from which one might approach the topic of "space", various of them of a very fundamental nature. Here I simply consider one possibility (arguably the simplest case), the structure of visual space as related to the physical space surrounding a single vantage point ("cyclopean, stationary observer"). I will consider the topology of the visual field to be "given" (so called problem of "local sign" [4]). These are strong assumptions, so we're dealing with a toy system. (However, I know of no machine vision work that seriously deals with the local sign problem.)

If you turn the whole world about the vantage point, the observer can undo this change through a voluntary eye movement, a rotation of the eye about its center. In the absence of additional data (e.g., a gravity sensor) such changes generate no information. The global optical structure shifts but the (all important) local spatial configurations are not affected.

If you scale the whole world about the vantage point the optical structure remains invariant. In the absence of additional data (e.g., the presence of Gulliver—who doesn't scale—in Lilliput and Brobdignac) such changes cannot be recorded at all.

We conclude that the optical structure available to a stationary, cyclopean observer is invariant against the group of rotation–dilations about the vantage point. It is easy enough to construct geometries that implement such invariance. Here is an example, consider the Riemann space [5] with line element (metric)

$$\mathrm{d}s^2 = \frac{\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2}{x^2 + y^2 + z^2}. \tag{1.1}$$

This metric is evidently invariant against arbitrary rotation–dilations about the origin $\{x,y,z\} = \{0,0,0\}$. Transforming to polar coordinates

$$\rho = \sqrt{x^2 + y^2 + z^2}, \tag{1.2}$$
$$\vartheta = \arccos(z), \tag{1.3}$$
$$\varphi = \arctan(x,y), \tag{1.4}$$

and setting

$$\zeta = \log \frac{\rho}{\rho_0}, \tag{1.5}$$

where ρ_0 is an arbitrary unit of length ("yardstick") you obtain

$$\mathrm{d}s^2 = \mathrm{d}\zeta^2 + \mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2 = \mathrm{d}\zeta^2 + \mathrm{d}\Omega^2, \tag{1.6}$$

where $\mathrm{d}\Omega^2$ is the line element (metric) of the unit sphere $\mathbb{S}^2$. Since the yardstick is arbitrary, the depth coordinate ζ indicates a point on the affine line $\mathbb{A}$ and we conclude that this space is a vector bundle $\mathbb{S}^2 \times \mathbb{A}$ with base space $\mathbb{S}^2$ and fibers $\mathbb{A}$.

Notice that points $\{\zeta_1, \vartheta, \varphi\}$ and $\{\zeta_2, \vartheta, \varphi\}$ are on the same "visual ray" and thus are imaged on the same "pixel". We conclude that such points are *coincident* in the (physical) *visual field*, though they may be distinct in the (mental) *visual space* of the observer. An example is a glyph like "×" where I may "see" the upslope "／" as being in front of the downslope "＼", thus the point "·" where they intersect as *two* points, one in front of the other, perhaps symbolized as "⊙". Of course I might as well "see" the upslope "／" as being *behind* the downslope "＼", the depth order being fully idiosyncratic. "Visual space" is a mental entity where the mind may shift the depths $\zeta_{1,2,\ldots}$ on the visual rays (directions, points of $\mathbb{S}^2$) as if they were beads on strings.

In order to deal with this in a formal manner you may redefine the metric in such a way that points like $\{\zeta_1, \vartheta, \varphi\}$ and $\{\zeta_2, \vartheta, \varphi\}$ that are on the same "visual ray" are assigned *zero distance* whereas still considered *different*. In the tradition of geometry such points would have to be designated "parallel". This is fully analog to the usage in the case of planes in space. Generically two planes subtend a finite angle (their distance in the angle metric), but it may happen that this angle vanishes without the planes being coincident. In that case one designates the planes to be "parallel".

The way to bring this about is to make the depth dimension *isotropic* [6]. On the "isotropic line" any two points subtend mutual distance zero. This is often useful in science, perhaps the best known example being the special theory of relativity where the light cones have isotropic generators. In a convenient formalism I introduce the "dual imaginary unit ε", where ε is defined as the nontrivial solution of the quadratic equation $x^2 = 0$. That is to say, you have $\varepsilon^2 = 0, \varepsilon \neq 0$. The numbers $u + \varepsilon v$ where $u, v \in \mathbb{R}$ are known as the "dual (imaginary) numbers", an extension of the real number line, much like the conventional imaginary numbers with imaginary unit i, where $\mathrm{i}^2 = -1$. One easily proves that neither $\varepsilon > 0$ nor $\varepsilon < 0$ whereas $\varepsilon \neq 0$. Thus the "Law of the Excluded Third" does not work for the dual number system and one has to adopt intuitionistic logic, which is probably as well in an engineering context. Engineering "proofs" are by nature *constructive*, "proofs by contradiction" play no role.

Writing the metric as

$$\mathrm{d}s^2 = \mathrm{d}\Omega^2 + \varepsilon^2 \, \mathrm{d}\zeta^2, \tag{1.7}$$

solves our problem. Points on different visual rays have finite distances (simply their angular separation), whereas points on a single ray have mutual distance zero, even if they are distinct. However, in the latter case one might define them to subtend the "special distance" $\zeta_2 - \zeta_1$. This is indeed a useful distance measure because it is invariant against arbitrary rotation–dilations. Notice that the special distance applies only to points with zero regular distance. Then one may define "the" distance as either the regular distance or (in case the regular distance vanishes) the special distance.

1.2.1.1 The Case of Narrow Visual Fields

In the case of narrow visual fields the formalism can be simplified. Let the main line of sight be in the Z–direction ($\vartheta \ll 1$). Then the angular coordinates $\{\vartheta, \varphi\}$ may be replaced with the "Riemann normal coordinates" $\{\xi, \eta\}$ [5]:

$$\xi = \vartheta \cos\varphi, \tag{1.8}$$
$$\eta = \vartheta \sin\varphi. \tag{1.9}$$

We obtain a space $\mathbb{E}^2 \times \mathbb{J}$ (where $\mathbb{E}^2$ denotes the Euclidian plane and $\mathbb{J}$ the isotropic affine line). The 8–parameter group

$$\xi' = a(+\xi \cos\beta + \eta \sin\beta) + c_\xi, \tag{1.10}$$
$$\eta' = a(-\xi \sin\beta + \eta \cos\beta) + c_\eta, \tag{1.11}$$
$$\zeta' = f_\xi \xi + f_\eta \eta + g\zeta + h, \tag{1.12}$$

is not unlike that of the similarities (thus including congruencies and movements) of Euclidian space $\mathbb{E}^3$, except for the fact that the latter group is only a 7–parameter group. The group scales distances by the amount a, which may be regarded as due to dilations, because it implements pseudo–perspective scaling; the parameter β results from rotations about the viewing direction, the parameters $\{c_\xi, c_\eta\}$ from rotations about axes orthogonal to the viewing direction; the parameter h from a depth shift. The parameters $\{f_\xi, f_\eta\}$ and g have been added because they conserve distance and special distance, for if $(\xi_2 - \xi_1)^2 + (\eta_2 - \eta_1)^2 > 0$ you have (due to $\varepsilon^2 = 0$)

$$((\xi_2' - \xi_1')^2 + (\eta_2' - \eta_1')^2) = a^2\,((\xi_2 - \xi_1)^2 + (\eta_2 - \eta_1)^2), \tag{1.13}$$

whereas for $(\xi_2 - \xi_1)^2 + (\eta_2 - \eta_1)^2 = 0$ you have

$$(\zeta_2' - \zeta_1') = g(\zeta_2 - \zeta_1). \tag{1.14}$$

This group of similarities defines (in the sense of Felix Klein) the Cayley–Klein space (one of 27) with a single isotropic dimension. The space has a parabolic distance measure (like Euclidian space), but unlike Euclidian space also a parabolic angle measure. This accounts for the additional group parameter: one may scale either distances or angles (or both), whereas Euclidian angles cannot be scaled because periodic (elliptic angle measure).

Since the similarities in the "image plane" (the $\{\xi, \eta\}$–plane) are trivial, as are depth shifts, the subgroup

$$\xi' = \xi, \tag{1.15}$$
$$\eta' = \eta, \tag{1.16}$$
$$\zeta' = f_\xi \xi + f_\eta \eta + g\zeta, \tag{1.17}$$

is perhaps of most immediate interest. It is the group of "bas–relief ambiguities" identified for the "Shape From Shading" problem of machine vision. Apparently the shading setting is irrelevant here, this transformation follows from a very general analysis of stationary, cyclopean vision. The parameter g immediately scales the depth of relief, whereas the parameters $\{f_\xi, f_\eta\}$ describe "additive planes", formally they describe isotropic rotations.

Consider a rotation

$$\xi' = \xi, \tag{1.18}$$
$$\zeta' = f\xi + \zeta, \tag{1.19}$$

in a constant η plane. The "frontoparallel" line $\zeta = 0$ is transformed into the slanted line $\zeta = f\xi$, which has the isotropic slope angle f. (This is a good angle measure because the transformation changes the slope of arbitrarily slanted lines by the same amount.) Apparently the slope angle varies between $\pm\infty$ and is not periodic. Thus you can't rotate the line "upside down". There is no "turning around" in visual space! Notice that this is exactly what is required because these rotations are in mental space: if you see the front of an object you can't see its back, no matter how you shift the depth "beads"along their visual ray "strings". The formalism perfectly captures the condition of a stationary, cyclopean observer.

The geometry and differential geometry of this "visual space" has been developed in detail during the first half of the 20$^{\text{th}}$c. (Not in the context of vision, but as a purely formal endeavor [7].) The resulting geometry is as rich as that of the familiar Euclidean space $\mathbb{E}^3$, though with some surprises. Calculations are typically much simpler than they are for analogous problems in Euclidean space, the main reason being algebraic. The full Taylor expansion of a function $F(u+\varepsilon v)$ being

$$F(u+\varepsilon v) = F(x) + \varepsilon F'(x)\, v, \tag{1.20}$$

(no higher order terms!) really makes life easy. Especially, the trigonometric functions become

$$\sin \varepsilon x = \varepsilon x, \tag{1.21}$$
$$\cos \varepsilon x = 1, \tag{1.22}$$

which enormously simplifies numerous calculations in geometry.

1.2.2 The Photometric Framework

In the "photometric framework" one deals with an "image plane", which is a Euclidian plane $\mathbb{E}^2$, and a scalar "intensity" field $I(x,y)$ (say). The intensity could be the irradiance of the image plane due to some optical system for instance.

The intensity domain. I will only consider two generic properties of the "intensity":

- the intensity is positive $I \in \mathbb{R}^+$ (I consider zero values singular);
- the intensity is a photometric quantity, i.e., its physical dimension is not length but involves radiant power in some way.

In the context of machine vision one typically doesn't care much about physical dimensions, thus it is convenient to decide on some standard intensity I_0 say, and redefine intensity as the dimensionless quantity $\overline{I} = I/I_0$.

In the absence of any prior knowledge the Bayesian prior for the intensity is hyperbolic:

$$\mathrm{P}(\overline{I})\,\mathrm{d}\overline{I} = \frac{\mathrm{d}\overline{I}}{\overline{I}}, \tag{1.23}$$

thus the *natural* representation of intensity is logarithmic, for then the prior probability density is a uniform density. Consequently I redefine intensity yet another time:

$$J = \log \overline{I} = \log \frac{I}{I_0}. \tag{1.24}$$

Since the fiducial intensity I_0 is arbitrary, the J–domain fails a natural origin. One concludes that the natural representation of the intensity domain is the affine line $\mathbb{A}$.

Apparently the objects that one deals with in the photometric framework are cross sections of the fiberbundle $\mathbb{E}^2 \times \mathbb{A}$ with the image plane as base space and the intensity domain as fibers. Such cross sections are conveniently referred to as "images".

The topological structure. In most cases the image will not be defined over the whole of the image plane, though the actual size of the available image is often irrelevant. For most purposes one may define a "region of interest", such that anything outside the region of interest does not affect the calculation. I will refer to the size of the region of interest as the "outer scale" of an image.

In work of a theoretical nature one often thinks of the intensity as defined on any point and one writes $J(\mathbf{r})$, with $\mathbf{r} \in \mathbb{E}^2$, whereas in work of a practical nature one considers intensities "pixel values" and writes J_{ij}, with $\mathrm{ij} \in \mathbb{Z}^2$. The former is nonsense, the latter inconvenient. The former is nonsense because the intensity is a flux density and only defined for finite collecting areas. Thus one needs to settle on some value of the *resolution*. The latter is inconvenient because the pixels ideally are (much) smaller than the size of one's operators (e.g., an "edge detector"). I will assume that a resolution has been decided upon and that it is much larger than the pixel size. Then any pixelation is irrelevant, a mere matter of implementation (e.g., of one's printer: you never hope to see the pixelation).

The formally correct way to deal with resolution is to consider the image a member of a linear scale space. This allows changes of resolution—which are often necessary or desirable—to be defined in a principled way. In this setting the "points" of the image plane are operators that when queried yield the intensity "at that point". These operators are "points" in the sense of "Euclid's Elements": "A point is that which has no parts". You cannot look *into* a point. I will refer to the size of the points as the "inner scale" of the image.

An advantage of the linear scale space setting is that it allows one to introduce partial spatial derivatives in a principled manner. One may actually take the derivative of a point and use the result as an operator that when applied to the image yields the value of the derivative at that point. This avoids the problem that images are not differentiable functions (in fact, not even functions to begin with) and that the approximate numerical differentiation of actual signals is a tricky business at best. (As people say "numerical differentiation is ill posed".) In fact, one may find derivatives of any order of the image *at any given scale*. Whether such (perfectly good!) derivatives are actually *relevant* or *useful* depends upon the current context.

1.2.2.1 The Structure of Image Space

The arena of images is a fiberbundle $\mathbb{E}^2 \times \mathbb{A}$. Can one identify additional structure? This is of appreciable potential interest as uses of popular applications like Adobe's Photoshop© indicate. People are ready to do all kinds of things to images, in many cases claiming to merely "improve" their image, not essentially changing it. One speaks of "straight photography". The transformation admitted in the practice of straight photography apparently play a role not unlike "congruences" or "similarities" and one would like to relate them to the structure of image space.

At first blush one identifies similarities of the image plane and translations along the intensity axis as obvious candidates. Another, perhaps less immediately obvious group of transformations are the similarities of the intensity domain. They correspond to the well known "gamma transformations"

$$I' = I_{\max} \left(\frac{I}{I_{\max}} \right)^{\gamma}, \tag{1.25}$$

of intensities in the range $(0, I_{\max})$.

Next consider transformations that leave the image plane invariant but depend on both space and intensity. Here one meets with an obvious constraint. For instance, I consider the "transformation"

$$x' = J, \tag{1.26}$$
$$y' = y, \tag{1.27}$$
$$J' = x, \tag{1.28}$$

as definitely not allowable. Why? Because the image plane dimensions and the intensity dimension are mutually incommensurable. This transformation violates the condition that images are cross–sections of the fiberbundle $\mathbb{E}^2 \times \mathbb{A}$. On the other hand the transformation

$$x' = x, \tag{1.29}$$
$$y' = y, \tag{1.30}$$
$$J' = ax + J, \tag{1.31}$$

does not represent any problem. The factor a is apparently a *gradient*, so much intensity per unit distance. The fiberbundle structure is not violated.

The situation should look familiar to the reader of this chapter. Images are manipulated by "moving intensities" over the copies of $\mathbb{A}$ at any point of the image plane, like beads on a string. Congruences should look like Euclidean motion in the image plane and leave distances between "beads on a single string" invariant, similarities should scale them by the same factor. This is exactly what is achieved by the group of similarities of a Cayley–Klein space with single isotropic dimension. In this case the isotropic dimension is the intensity domain. Thus I simply set:

$$x' = a(+x\cos\beta + y\sin\beta) + c_x, \tag{1.32}$$
$$y' = a(-x\sin\beta + y\cos\beta) + c_y, \tag{1.33}$$
$$J' = f_x x + f_y y + gJ + h, \tag{1.34}$$

as it does precisely the right things. The subgroup that leaves the image plane invariant is evidently the most interesting. It is

$$x' = x, \tag{1.35}$$
$$y' = y, \tag{1.36}$$
$$J' = f_x x + f_y y + gJ + h. \tag{1.37}$$

The parameter h controls overall brightness, whereas parameter g implements the gamma–transformations (usually denoted "contrast control"). The parameters $\{f_x, f_y\}$ are often applied by landscape photographers as "grad filters".

These transformations have many applications in vision. For instance, consider the local image structure

$$J(x,y) = a_{00} + (a_{10}x + a_{01}y) + \frac{1}{2!}(a_{20}x^2 + 2a_{11}xy + a_{02}y^2) + \dots \tag{1.38}$$

Using a congruency of image space it can be transformed into canonical form

$$J'(u,v) = \frac{1}{2!}(\kappa_1 u^2 + \kappa_2 v^2) + \dots \tag{1.39}$$

With an additional similarity one may even achieve

$$\sqrt{\frac{\kappa_1^2 + \kappa_2^2}{2}} = 1, \tag{1.40}$$

that is unit "curvedness". The ratio κ_1/κ_2 is a pure (second order) shape measure. The four coefficients of the cubic term are also interesting *cubic shape measures* because obviously differential invariants. It seems likely that such local measures are taken in the human visual system.

1.3 A Case Study: "Shape from Shading"

"Shape From Shading" is not exactly the biggest success of machine vision. It is not so clear that human vision is doing much better though. The issue remains undecided because the very aims of machine vision and human vision appear to be widely different. This makes shape from shading of some interest as a case study.

1.3.1 The So Called "Shape from Shading Problem"

The so called "Shape From Shading Problem" as conventionally construed is rather artificial and relies on numerous shaky, even *a priori* unlikely, occasionally even plainly wrong prior assumptions. Here is a rough outline:

An observer views a smooth, generically curved surface that is being illuminated such as to produce a pattern of light and shade. The task is to report the shape (that is the curvature landscape) of the surface. Sometimes the observer may also be asked for the nature of the illuminating beam (e.g., the spatial configuration and photometric properties of the "primary sources").

As stated the problem is probably an impossible one to tackle, thus one lists any number of simplifying prior assumptions. Among these may be:

- the surface is smooth, no edges, no contours, no 3D texture (roughness);
- the surface is uniform, i.e., the same at all places;
- the surface is characterized by a single bidirectional reflectance distribution function (BRDF). Thus effects of translucency do not play a role;
- the BRDF is constant, in other words, the surface is Lambertian;
- the illuminating beam has a uniform cross section, it will illuminate a set of concentric spherical surfaces uniformly. ("A homocentric, collimated beam");
- the illuminating beam will illuminate planes uniformly. ("A parallel beam");
- the illumination is by primary sources only. I.e., there are no mutual interreflections;
- each point of the surface is illuminated by the same primary sources. I.e., there is no "vignetting" or "cast shadow".

Some of these assumptions are mutually exclusive, others imply each other. Accepting some may have strong consequences, e.g., the absence of interreflections implies that the surface is either black or non–concave; the Lambertian assumption implies that the viewing geometry is irrelevant.

It is possible to frame certain limiting cases in which some of the assumptions are automatically (though approximately) satisfied. One interesting example is to assume "low relief". This automatically takes care of the vignetting and interreflection issues. Whether such limiting cases are of any interest depend on one's goals. If the goal is applications the assumption of low relief is likely to be frequently violated. If the goal is theoretical understanding such a limiting case may be of considerable use.

Fig. 1.1 The Asam house at Munich. Illumination by the overcast sky from above. The material is whitewashed stucco, roughly Lambertian. There are various regions of low relief where our simplifying assumptions hold reasonably well (the clock face, the sitting putto) though there are also parts that are modeled "in the round" and where effects of vignetting and interreflection are evident. In cases like this frontal viewing is a natural condition (I made the photograph from he opposite side of he street).

A special case that appears rather limiting, yet is often applicable is that of frontal viewing (see Figure 1.1). Expecially when combined with the low relief assumption this often applied to cases of real life importance, just think of viewing bas relief murals.

Another special case is that of frontal illumination. This case is completely different from frontal viewing. Moreover, it has few other applications than photographs taken with flash on the camera. Although this situation is avoided like the plague by professional photographers (it "flattens" the scene, thus works against regular shape from shading for the human observer), it is (perhaps perversely) a case for which dedicated computer algorithms have been designed.

Some of the most useful assumptions are almost certainly wrong. A key example is the assumption of Lambertian surfaces. It is a highly desirable assumption because the influence of viewing geometry vanishes, thus greatly simplifying the problem. But from physics one knows that Lambertian surfaces don't exist. It is not that basic physics forbids them, but they can be only approximately produced even

under laboratory conditions. The non–Lambertian character of surface scattering becomes especially obvious for very oblique viewing and/or illumination directions.

Here I will assume Lambertian surfaces, low relief, frontal viewing and parallel, collimated illumination. In this case most of the usual assumptions apply (no vignetting, no interreflections, ...), even the Lambertian assumption is not problematic. This is easily the simplest setting imaginable that still holds some interest.

1.3.2 Setting up the Problem

Consider a relief

$$z(x,y) = z_0 + \mu w(x,y), \tag{1.41}$$

where μ keeps track of the depth of relief. It is a convenient parameter because we simply carry calculation to 1st–order in μ. Here is an example, the surface normals are

$$\mathbf{n}(x,y) = -\mu\left(\frac{\partial w(x,y)}{\partial x}\mathbf{e_x} + \frac{\partial w(x,y)}{\partial y}\mathbf{e_y}\right) + \mathbf{e_z} + \mathbf{O}[\mu]^2. \tag{1.42}$$

Here we obtained a significant gain in simplicity because the usual normalization factor affects only 2nd and higher orders in μ and thus can be ignored.

I will set

$$\frac{\partial z(x,y)}{\partial x}\|_{x=y=0} = \frac{\partial z(x,y)}{\partial y}\|_{x=y=0} = 0, \tag{1.43}$$

throughout the computation because of the assumption of frontal viewing.

Assume the direction of the illuminating beam is $\mathbf{i}$ and that it causes a normal illimination E_0. Then Lambert's Cosine Law yields the illumination pattern:

$$E(x,y) = E_0\,\mathbf{i}\cdot\mathbf{n}(x,y) = E_0\left(-\mu\left(i_x\frac{\partial w(x,y)}{\partial x} + i_y\frac{\partial w(x,y)}{\partial y}\right) + i_z\right) + \mathbf{O}[\mu]^2. \tag{1.44}$$

Notice that the absolute value of the illuminance is irrelevant. The visual system will merely record the *spatial contrast* $C(x,y)$, which is

$$C(x,y) = \frac{E(x,y) - E(0,0)}{E(0,0)} = -\frac{\mu\left(i_x\frac{\partial w(x,y)}{\partial x} + i_y\frac{\partial w(x,y)}{\partial y}\right)}{i_z} + \mathbf{O}[\mu]^2. \tag{1.45}$$

Writing

$$\mathbf{i} = -\left(\cos\vartheta\left(\cos\varphi\,\mathbf{e_x} + \sin\varphi\,\mathbf{e}_y\right) + \sin\vartheta\,\mathbf{e}_z\right), \tag{1.46}$$

where ϑ denotes the elevation and φ the direction of the illumination, we finally obtain

$$C(x,y) = \mu\cot\vartheta\left(\cos\varphi\frac{\partial w(x,y)}{\partial x} + \sin\varphi\frac{\partial w(x,y)}{\partial y}\right) + \mathbf{O}[\mu]^2. \tag{1.47}$$

What can be observed locally is the contrast gradient $\nabla C = C_x \mathbf{e}_x + C_y \mathbf{e}_y$, it can be found by straight differentiation. The differentiation will generate second order derivatives of the height $z(x,y)$. Dropping higher order terms in μ you obtain

$$C_x = \mu \cot\vartheta (\cos\varphi \frac{\partial^2 w(x,y)}{\partial x^2} + \sin\varphi \frac{\partial^2 w(x,y)}{\partial x \partial y}) \tag{1.48}$$

$$C_y = \mu \cot\vartheta (\cos\varphi \frac{\partial^2 w(x,y)}{\partial x \partial y} + \sin\varphi \frac{\partial^2 w(x,y)}{\partial y^2}). \tag{1.49}$$

In the "Shape From Shading Problem" the "unknowns" are ϑ, φ, and the three 2^{nd}–order partial derivatives of the height of relief $z(x,y)$. For this we have two equations, the observables C_x and C_y. The problem is evidently underdetermined, even in this simplest setting.

One ambiguity that is clearly unavoidable is the mix up between the height of contrast and the elevation of the source as expressed through the factor $\mu \cot\vartheta$. A scaling $\mu w(x,y)$ can be undone by adjusting ϑ. We may as well notice this relation and proceed to eliminate ϑ, obtaining a homogeneous equation for the three partial derivatives.

Writing $\nabla C = G(\cos\gamma\, \mathbf{e_x} + \sin\gamma\, \mathbf{e_y})$ we obtain

$$\sin\gamma\cos\varphi\, z_{xx} - \cos(\gamma+\varphi)\, z_{xy} - \cos\gamma\sin\varphi\, z_{yy} = 0, \tag{1.50}$$

where I have introduced a more concise notation for the partial derivatives. Apart from this we have that the height of relief is undefined. This may be expressed through the equation

$$\frac{1}{2}(z_{xx}^2 + 2z_{xy}^2 + z_{yy}^2) = \text{constant}, \tag{1.51}$$

the height being absorbed in the elevation of the source. (The expression is the "curvedness", see below.)

Thus we end up with one parameter (the elevation of the source) remaining fully unspecified and two equations for four unknowns (the illumination direction and three partial derivatives of the height). It would help to know the direction of illumination, but even then we still have only two equations for the three partial derivatives. The problem is evidently very underspecified.

Ways to proceed. There are various ways to proceed from here. Well known methods from machine vision recognize the fact that the local conditions are insufficient to find the local shape (curvature) and reformulate the Shape From Shading Problem into a global problem, using either partial differential equations or a variational method. Thus one introduces surface integrity constraints to force a solution. These methods are well known and I will not pursue them here because they are quite unlike anything that might be attempted by the human visual system. I will stubbornly pursue the purely local problem in an attempt to guess what the visual system might be doing.

Posing the local problem in a more symmetrical way: describing 2^{nd}–order shape. The description of 2^{nd}–order surface shape in terms of partial derivatives in a Cartesian frame in the tangent plane is often convenient, but masks the symmetries of the 2^{nd}–order structure. Here is a better adapted description:

Notice that a term like x^2+y^2 is rotationally symmetric whereas terms like xy and x^2-y^2 have two lines of bilateral symmetry. The latter two terms are very similar and can be transformed into each other through a rotation of the coordinate system over $\pi/4$. Hence the transformation

$$z_{xx} = r+t, \tag{1.52}$$
$$z_{xy} = s, \tag{1.53}$$
$$z_{yy} = t-r, \tag{1.54}$$

thus we obtain

$$z(x,y) = \frac{1}{2}(z_{xx}x^2 + 2z_{xy}xy + z_{yy}y^2) = r\frac{x^2-y^2}{2} + sxy + t\frac{x^2+y^2}{2}. \tag{1.55}$$

I treat the surface as in visual space, that is to say, differential invariants like the mean and Gaussian curvature are calculated as in singly isotropic (the z–direction) space. This meshes perfectly with the assumption of "low relief". (This even allows one to introduce an overall surface slant without any complication.)

The principal curvatures are

$$\kappa_{1,2} = t \pm \sqrt{r^2+s^2}, \tag{1.56}$$

and the principal directions are

$$\{r \pm \sqrt{r^2+s^2}, s\}. \tag{1.57}$$

Thus the mean curvature H and Gaussian curvature K are

$$H = \frac{1}{2}(\kappa_1+\kappa_2) = t, \tag{1.58}$$
$$K = \kappa_1\kappa_2 = -r^2-s^2+t^2. \tag{1.59}$$

The expression $\frac{1}{2}(\kappa_1-\kappa_2)$ may be called the "non–sphericity" as it measures the deviation from rotational symmetry. It equals $\sqrt{r^2+s^2}$, thus sphericity implies $r=s=0$.

The "curvedness" $\chi = \sqrt{\frac{1}{2}(\kappa_1^2+\kappa_2^2)}$ measures the deviation from planarity and turns out to be $\sqrt{r^2+s^2+t^2}$. The "shape index" specifies the pure shape and is defined as

$$\sigma = \arctan\frac{\kappa_1+\kappa_2}{\kappa_1-\kappa_2} = \arctan\frac{t}{\sqrt{r^2+s^2}}. \tag{1.60}$$

The shape index takes values on $[-\frac{\pi}{2}, +\frac{\pi}{2}]$.

All this can be summarized in an intuitively very attractive manner. The space of all 2nd–order surface shapes is best represented by a Cartesian $\{r,s,t\}$–space. The origin represents planarity, i.e., shapelessness, whereas distance from the origin, the "curvedness" implies deviation from the tangent plane. On the surface of the unit sphere the latitude is the shape index, whereas the longitude indicates twice of the direction of principal curvature. The natural representation is in polar coordinates

$$r = \chi \cos\sigma \cos\psi, \tag{1.61}$$
$$s = \chi \cos\sigma \sin\psi, \tag{1.62}$$
$$t = \chi \sin\sigma. \tag{1.63}$$

This direction of principal curvature (that is $\frac{\psi}{2}$) is undefined at the poles because the t–axis represents the spherical shapes. Notice that antipodes are mutually related as a cast and its mold.

1.3.3 The Local Shape from Shading Problem

The local Shape From Shading problem is best recast in terms of the symmetrical parameters $\{r,s,t\}$ introduced above.

The ambiguity due to the elevation of the illumination means that shape inferences have to be done *modulo* the curvedness, which again means that the space of possible inferences is reduced to the lines through the origin of shape space, a projective plane. We may represent it by the unit sphere in shape space with pairs of antipodal points identified.

The observation of the contrast gradient yields the constraint

$$r\sin(\gamma+\varphi) - s\cos(\gamma+\varphi) + t\sin(\gamma-\varphi) = 0, \tag{1.64}$$

which is a homogeneous, linear equation thus a plane through the origin of shape space, which meets the unit sphere in a great circle. At this point we may simplify the expression by specializing the coordinate system, letting the first frame vector coincide with illuminance surface flow direction. Thus, setting $\gamma \to 0$ the constraint is

$$r\sin\varphi - s\cos\varphi - t\sin\varphi = 0. \tag{1.65}$$

The pole of this great circle is

$$\mathbf{p}(\varphi) = \frac{\sin\varphi\,\mathbf{e_1} - \cos\varphi\,\mathbf{e_2} - \sin\varphi\,\mathbf{e_3}}{\sqrt{1+\sin^2\varphi}}. \tag{1.66}$$

Of major interest is the colatitude of the pole, because it specifies the extreme values of the possible shape index inference. It is

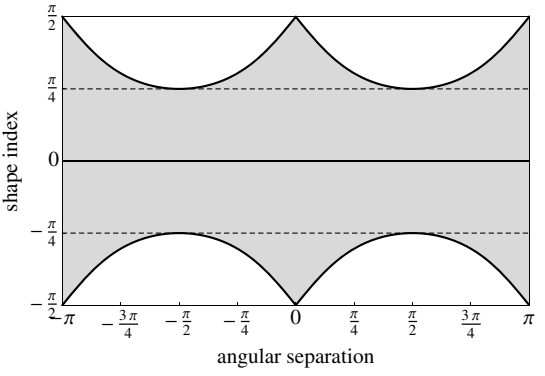

Fig. 1.2 The range of feasible shape indices (dark) as a function of the angular separation between the direction of illumination and the direction of the contrast gradient.

$$\sigma_{\max} = \frac{\pi}{2} - \left| \arcsin \frac{\sin \varphi}{\sqrt{1 + \sin^2 \varphi}} \right| . \tag{1.67}$$

Although it is quite possible to infer any hyperbolic shape, no matter what the value of φ might be, there is a maximum to the shape index of elliptic inferences (see Fig. 1.2). For instance, a spherical inference implies $\gamma = \varphi$ (of course modulo π).

1.3.3.1 The "Observables" for the Shape from Shading Problem

In the literature on the Shape From Shading Problem the generic assumption is that the relevant observable is the spatial contrast. (For the local problem this reduces to the contrast gradient, though this plays no role in the machine vision literature.) However, this assumption may well be questioned.

Various attempts to find the illuminance direction from the image are found in the literature, most of them ad hoc, some of of them mere shots in the dark.

A principled manner to find the illuminance direction from the image is available if the surface is corrugated such as to yield a visible illuminance induced texture. Such a method works if the statistical structure of the corrugations is isotropic. The basic idea is simple enough. An isotropic protrusion will yield a dipole pattern, light on the side facing the source, dark on the side facing the other direction. An isotropic indentation will also yield a dipole pattern, it will have the same axis as that of the protrusion, but the opposite polarity. Thus the gradients have the same *orientation*, but opposite *directions*. The average gradient of an isotropic texture will indeed tend to zero, but the average *squared* gradient will have the correct orientation. This is the crux of the structure tensor method. One computes the eigensystem of the "structure tensor"

$$S = \langle \nabla C^{\dagger} . \nabla C \rangle = \begin{pmatrix} \langle C_x C_x \rangle & \langle C_x C_y \rangle \\ \langle C_y C_x \rangle & \langle C_y C_y \rangle \end{pmatrix} . \tag{1.68}$$

The direction of the largest eigenvector is the orientation of the illumination flow [8].

This method has been shown to work very well with a large variety of 3D textures. Isotropy is essentially the only requirement. It has also been shown that the human observer uses this method and typically finds the orientation with an accuracy of about 5°.

1.3.3.2 Human Vision and Shape from Shading

We have solved the Local Shape From Shading Problem above and we have introduced the possibility of additional observational evidence. This should be sufficient to investigate the angle human vision takes on the problem.

We identify three cases:

- the observer lacks any prior information, except for the general setting: the observer is looking at a frontoparallel, Lambertian (e.g., plaster or marble) plane with low relief modulation, illuminated by a uniform, parallel, collimated beam (e.g., the sun);
- the observer additionally has prior knowledge concerning the illumination direction (e.g., through observation of 3D texture induced contrast);
- the observer has prior information concerning the shape (e.g., knows it to be spherical).

In the first case the observer is supposed to estimate both the shape and the illumination direction, in the second case only the shape and in the third case only the direction of illumination. The first case is the most interesting, though especially the second case may be expected to have frequent application. An overview of the various relations is graphically illustrated in figure 1.3.

Consider the first case. The observer observes the direction of the contrast gradient, we specialize the coordinate system such that $\mathbf{e_x}$ is in the contrast gradient direction. Of course there remains a $\pm\pi$ ambiguity here. Then we know that the shape index is limited as

$$\sigma \leq \frac{\pi}{2} - \left| \arcsin \frac{\sin\varphi}{\sqrt{1+\sin^2\varphi}} \right|, \tag{1.69}$$

where the direction of illumination φ is supposed to be fully unknown.

It is *always* possible to infer a symmetrical saddle ($s = 0$), for *any* direction of illumination. It is also possible to infer a spherical shape ($s = \pm\frac{\pi}{2}$), though this implies that the illumination direction coincides with the contrast gradient direction ($\varphi = 0$), a very specific condition. Of course the saddle would have to be in a specific orientation, whereas the sphere looks the same in *all* orientations. All considering, it is hard to make a principled choice. Of course *any* shape is possible, with more complicated constraints on the directions.

Consider the second case. If the direction of illumination is known, there is a constricted range of possible values of the shape index. Granted a preference for elliptical shapes (see below) one expects the system to select the largest possible value, then the best guess is

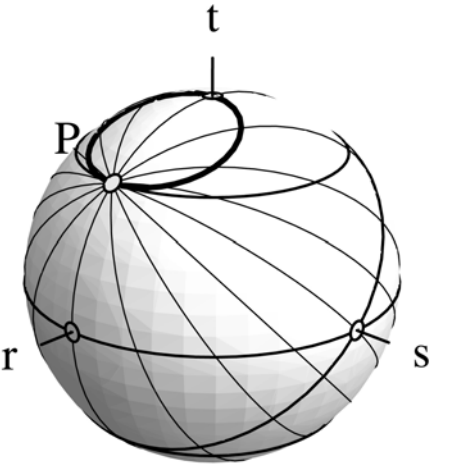

Fig. 1.3 The unit sphere in rst–space (equation 1.51 that is $\chi = 1$), where the r–dimension is the direction of the relative contrast gradient. The sheave of great circles through P (midpoint of the arc rt) are the constraint planes (equation 1.50) for the various illumination directions (the poles lie on the great circle passing through s.) The fat small circle is the locus of "most spherical inferences". At P the inference is cylindrical and all illumination directions go, at t the inference is spherical and the illumination direction coincides with the gradient direction.

$$\sigma = \frac{\pi}{2} - \left| \arcsin\left(\frac{\sin\varphi}{\sqrt{1+\sin^2\varphi}} \right) \right|. \tag{1.70}$$

Thus knowledge of the illumination direction fails to nail the shape, but does constrain the possibilities. (See Figure 1.4 .)

Finally consider the third case. Knowing the shape means that only points on the latitude circle of that shape are feasible. Thus the solutions lie on the intersection of the great circle defined by the constraint and this latitude (small) circle. It is possible that no solution exists, otherwise there are two distinct ones. If so, the solution is

$$\varphi = \arccos \frac{|\sin\sigma - \cos\psi\cos\sigma|}{\sqrt{1-\cos 2\sigma\cos\psi}}. \tag{1.71}$$

In this case one might run into a *contradiction*, which is the strongest constraint possible. Otherwise such a prior knowledge pretty much nails the direction of illumination (see Figure 1.5). A key example is the spherical shape which may act like a "wind sack" for the flow of light.

How well is the human visual system doing? In the absence of prior knowledge it appears to be the case that the human observer invariably reports a spherical shape. Apparently the visual system considers the spherical inference the best bet. This may be due to the fact that (overall) smooth objects are likely to be predominantly convex. The remaining convex/concave ambiguity is usually resolved by the prior assumptions that illumination tends to be from above. If the illumination is actually

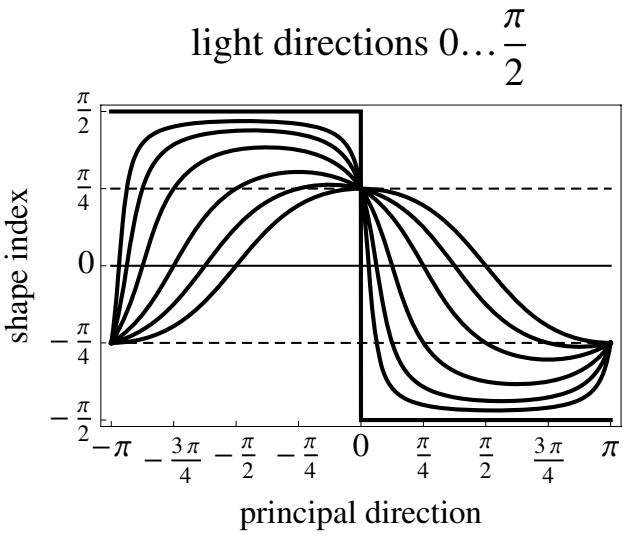

Fig. 1.4 Assume the direction of illumination is known (each curve is for a specific direction of illumination), then the shape (as specified by the shape index) still depends upon the (unknown) direction of principal curvature. Notice that an elliptical inference (convex or concave) is always possible, though generically not spherical.

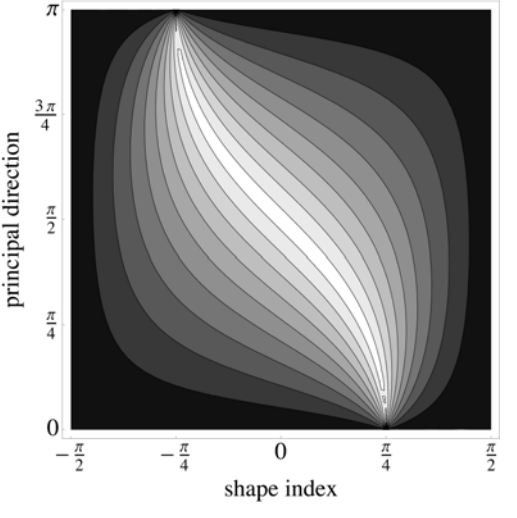

Fig. 1.5 Suppose the shape (specified by the shape index) is known. If the direction of principal curvature is also known the light direction is fully determined, otherwise there exists a one–parameter ambiguity. The shades run from black $= 0$ to white $= \frac{\pi}{2}$.

from the side humans typically report convexities, thus an additional bias seems to be on convexity as opposed to concavity.

Notice that these "inferences" are in fact "hallucinations". In reality any shape goes! Yet human observers rarely feel that they are *guessing*, the experience is that of *seeing* a specific shape.

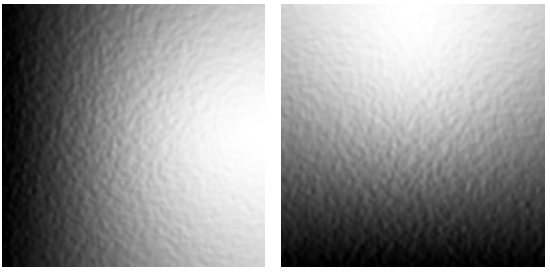

Fig. 1.6 Two renderings of a rough surface illuminated from the right. Both surfaces are quadrics viewed frontally at the center of the figure. The surface at right is spherically convex, that at left is a symmetric saddle. Notice that the 3D–texture *visually* reveals the direction of illumination.

If the direction of illumination is clearly visible one expects the human observer to use this. (It has been established that human observers indeed see the illumination direction based on 3D texture.) Perhaps surprisingly, *they don't*. People tend to report sphericity, even when there is no elliptic solution at all, that is when the contrast gradient is orthogonal to the illumination direction! (see Figure 1.6) Thus there can be no doubt that a machine algorithm would do better. What frequently happens in such cases is that the presentation "splits" into two layers (like it does often as in cases of apparent transparency). One layer has the 3D texture, illuminated veridically whereas the other layer gets the (spherical) shape, illuminated from a direction perpendicular to the (clearly visible!) veridical direction.

We conclude that the visual system attempts to solve the local problem (where machine vision gives up), but doesn't do too well on it. One could easily beat it with a simple machine algorithm.

Remaining questions. As always in the study of human visual perception, many questions remain. One is whether the visual system does anything global in addition to purely local inferences. Whereas it is very unlikely that the system does anything remotely like the current machine algorithms, there remain a number of intriguing possibilities.

It is likely that the visual system pursues "local" inferences on various levels of resolution. Whereas this would hardly be of much interest if there were no "added value" to it, this would be of interest because the assumption that the illumination directions are the same irrespective of the level of resolution is a very reasonable one. For articulated surfaces (a globally quadric surface would not profit at all) this is a very promising proposition. Such methods would be in between local and global, though quite different from the global algorithms in use by the machine vision community today.

It is also likely that the system would not stop at the 2^{nd}–order surface structure. There is much reason to believe that the 3^{rd}–order surface structure has to be very relevant for shading based inferences. For instance, the singular points of the illuminance pattern (extrema and saddle points) occur at the parabolic points of the surface [9]. At such points the surface is locally degenerate (cylindrical) and for a generic description one has to take the cubic terms into account. The cubic terms will also affect the Hessian (in addition to the gradient) of the contrast and it is very

likely that the visual system is sensitive to those, indeed, perhaps more so than to the gradient. As we have argued above, at any given point the gradient can be transformed away through a congruence in image space. Little is known about the way cubic structure appears in shading, although the potential importance cannot be in doubt.

1.4 Final Remarks

"Final remarks" is more apt than "conclusions" at this point, because there isn't any real "conclusion". What I have tried to convey in this chapter is the enormous gap between the attempts to understand the functioning of the human visual system in a formal way and the attempts to implement machine systems that "see". One might expect these endeavors to overlap appreciably because both the initial "data" and the final "goals" are very similar or even identical. Moreover, the generic human observer and the generic machine vision system are supposed to function in very similar (often identical) worlds. However, the differences are very significant. They are mainly due to hardware constraints. The major bottlenecks of biological as opposed to artificial systems are:

- whereas absolute calibrations or at least fixed operation points are usually no problem in artificial systems, such luxuries are not available in biological systems. In biological systems any level has to dynamically shift its operation level in order to keep signals within the (very limited) dynamical range and these levels are not known to other parts of the system (or even the individual subsystem itself);
- in biological systems local processing is the rule, global processing only works in very coarse grained (sub–)systems (that is to say, low resolution is substituted as a cheap replacement of true globality). This rules out most algorithms that have made machine vision into a viable technology.

As opposed to these limitations biological systems also have major strengths, the main one being the full integration of the "background". Biological systems are part of their biotopes and background knowledge of the structure of the biotope is evident at all levels of implementation, from the optics of the eye to the nature of the "hallucinations". This is an aspect that machine vision has hardly touched upon.

The examples I gave in this chapter I believe to be typical in showing up such differences.

The limitations of biological systems may be a burden to an engineer designer, but evolution has done remarkably well given these constraints. Thus I believe that machine vision has something to learn from biological implementations even though I also believe that biological systems are bound to be beaten in many subdomains by well designed artificial systems, certainly on the long run. The lessons will be (of course) very general design principles and in this chapter I have tried to outline a few.

Areas where the human visual system is unlikely to give way to machine implementations are those of the visual arts. However, such achievements are very

difficult to measure up. There are essentially no yardsticks for the products of the creative arts. The only way to assess this is to try to find whether there is a market (in the art galleries circuit) for machine generated products. If human artists are eventually muscled out of these circuits, machines will finally be in power. That'll be the day.

References

1. http://www.research.ibm.com/deepblue/
2. Koenderink, J.J.: The brain a geometry engine. Psychological Res. 52, 22–127 (1990)
3. http://nl.wikipedia.org/wiki/Sherlock_Holmes
4. Lotze, R.H.: Medicinische Psycholohie oder Physiologie der Seele. Weidman'sche Buchhandlung, Leipzig (1852)
5. Riemann, B.: Über die Hypothesen, welche der Geometrie zu Grunde liegen. Habilitationsschrift, Abhandlungen der Kniglichen Gesellschaft der Wissenschaften zu Gttingen 13 (1854)
6. Sachs, H.: Ebene isotrope Geometrie. Vieweg-Verlag, Wiesbaden (1987)
7. Yaglom, I.M.: A simple non–Euclidean geometry and its physical basis: an elementary account of Galilean geometry and the Galilean principle of relativity (Translated from the Russian by Abe Shenitzer). Springer, New York (1997)
8. Koenderink, J.J., Pont., S.C.: Irradiation direction from texture. J. Opt. Soc. Am. A 20, 1875–1882 (2003)
9. Koenderink, J.J.: Photometric invariants related to solid shape. Optica. Acta. 27, 981–996 (1980)