Local Operations: The Embodiment of Geometry

Jan-Johan Koenderink, Astrid Kappers and Andrea van Doorn

Buys Ballot Laboratorium, Universiteit te Utrecht

We consider the default structure of the visual front end based on very general symmetry considerations and invariance principles. It turns out that such very general principles constrain the possible sampling structures greatly, and that the resulting default structure is not unlike the actual structure of the primate front end visual system and also not unlike the structures as they have "evolved" in computer vision. The general formalism suggests various useful relations that are an immediate consequence of the front end structure but are generally being rediscovered in ad hoc ways in image processing. The formalism allows a novel interpretation of the concept of a "feature", an otherwise elusive concept in organic and computer vision alike. This interpretation may well turn out to be useful in image understanding.

1 The "Front End Visual System"

Vision is best defined as "optically guided behavior". It is sustained by a dense, hierarchically nested and heterarchically juxtaposed tangle of cyclical processes. On the coarsest scale of description these processes comprise both causal connections in the environment of the agent ("ecological optics"), as well as somatic processes. In this chapter we focus upon the *interface* between the light field and those parts of the brain nearest to the transduction stage. We call this the "visual front end". This interface is an obvious bottleneck of optically guided behavior since all optically specified information has to be passed to higher centers by way of the front end. Thus a thorough analysis of this stage is a prerequisite for the description of higher order processes. Of course, the exact limits of the interface are essentially *arbitrary*, but nevertheless the notion of such an interface is valuable.

We suggest that the demarcation be drawn according to the following considerations (which may be applied with greater or lesser severity according to whim or necessity):

- the front end is a "machine" in the sense of a *syntactical transformer* (or "signal processor");
- there is no semantics (reference to the environment of the agent). The front end merely processes *structure*;

- the front end is *precategorical*, thus – in a way – the front end does not compute anything;
- the front end operates in a *bottom up* fashion. Top down commands based upon semantical interpretations are not considered to be part of the front end proper;
- the front end is a deterministic machine, *i.e.*, it doesn't "hallucinate" or "dream"; but all output depends causally on the (total) input from the immediate past (the "specious moment" as defined in the psychology of time perception).

This roughly outlines the meaning of the term "visual front end". The notion is similar in spirit (though more technical) to Orban's distinction between literal and interpretative representations. (This book.)

The *task* of the front end is to transform, encode and distribute the transduced spatiotemporal irradiance distribution in such a way as to enable efficacious visual processing. That means: processing subserving efficacious optically guided behavior of the organism as a whole. What is not explicitly encoded by the front end is irretrievably lost. Thus the front end should be universal (undedicated) and yet should provide explicit data structures (in order to sustain fast processing past the front end) without sacrificing completeness (everything of potential importance to the survival of the agent has to be represented somehow). Clearly there are some incompatible objectives here: for instance, how can one encode explicit data structures without commitment?

The visual front end is being studied in physiology and psychophysics. We aim at a *general formal framework*, as the only way to proceed from mere factcollecting to natural philosophy. Mathematical structures are used to reorder existing facts. Certain mathematical structures can be used to *represent* given physical structures, for instance geometrical objects represent the local behavior of extended structures. The emphasis will be on these structures themselves, rather than on formal manipulation or representation, *i.e.*, on *geometrical* aspects.

Typically "geometrical objects" are defined as equivalence classes of other objects. For instance, a "tangent vector" is an equivalence class of curves. All tangent vectors at a point span "tangent space", which is a local picture of the space itself. This is the preferred level of description from a physicist's point of view: one concentrates on the essential structure without bothering too much about accidental representation. (*E.g.*, technicalities such as coordinate representation.) The same position is a natural one for the description of the visual system: for instance, every point of the visual field carries a copy of tangent space. Indeed the cortical hypercolumns may be interpreted as their embodiments. The language of "fiber bundles" thus very aptly describes the structure of the visual front end.

The geometrical language provides a universal language and a uniform format in which to describe front end structures. The language is abstract and basically devoid of any *meaning* (*i.e.*, it is pure syntax). Indeed, meaning is due to interpretation, a top down action of the organism. A top down query to the front end entails a "logical format" that bestows meaning on the front end structure. (Thus the same structure may acquire many different meanings.) It's like the format of a "read" command in many computer languages. The same *datum* may be treated as an ASCII-character, a memory address, or an integer number, depending on the format statement associated with the read command. The meaning is not (only) in the structure of the datum, but in the read action performed upon the datum. Of course it doesn't necessarily make sense to interpret a given datum in any old way: it is the responsibility of the process that issues the command to make it a

sensible one, otherwise gibberish results. (An unfortunate effect that is not unfamiliar to computer users!) In the neurosciences this problem is known historically as the problem of "local sign", or of the "homunculus". We will not address such important problems in this chapter.

The front end cannot represent everything. There has to occur some process of *selection.* This is a very serious matter, because what's not represented doesn't so much as *exist* for the agent, in the sense that it cannot codetermine efficacious action of the agent.

By "representing" we mean something like "segregation of quality", putting things in distinct, addressable pigeonholes. This "segregation" implies a parallelism, that is a minimization of lateral connections. The primitive notions of "continuity" and "coherence" make that one particularly useful "segregation of quality" is a division in terms of *locality.* In a local representation one can do without extensive (that is spatial, or geometrical) properties and represent everything in terms of intensive properties. This obviates the need for explicit geometrical expertise. The local representation of geometry is the typical tool of differential geometry. For instance, a *vector* summarizes a *bilocal* property (vector as "arrow with tip at $\mathcal{B}$, tail at $\mathcal{A}$") in a purely *local* manner. The columnar organization of representation in primate visual cortex suggests exactly such a structure. Local sign has to be attached to the (hyper-)columns, whereas the activity within such a (hyper-)column can be *local* in the sense that only intensive, rather than extensive, operations need to be performed.

1.1 Consequences of Non-Commitment

Non-commitment means not making choices: no place is *a priori* different from any other, no orientation special, no level of resolution more important than any other. "Not seeing the wood for the trees" is a serious blindness, but one doesn't want to loose sight of the trees either. Any firm commitment of the front end may allow especially powerful processing of some of the structure, but also necessarily entails a limit on the agent's repertoire of efficacious behavior. In practice extreme dedication of the front end is only found in species for which ecological niches exist that allow them to get away with diminished abilities in other areas. *Homo sapiens* may be the most universal animal around; if so, then its front end must be the least dedicated.

In this paper we explore "ideal", that is completely undedicated front ends. Every real species falls short of this, sometimes for good reasons. Exploration of the ideal limit is a help in understanding real systems nevertheless: it is the only prototype available against which to judge the merit of real systems.

The consequences of non-commitment are that the front end must implement a structure that is invariant under certain basic symmetry groups. The collective symmetries characterize the front end. They are like axioms of a formal system. You can't *deduce* the symmetries from first principles, you *postulate* them. They sum up the common sense notion of non-commitment: the following are a set of symmetries that appear very basic indeed:

homogeneity no location or moment is special. This entails translational symmetry in space and time;

scale invariance no spatial or temporal scale is special (for the eternal eye the eagle's perspective is no more important than the mole's). This entails scale invariance (or self-similarity) in space and time;

isotropy no spatial orientation is singled out. (The vertical is special to us, but how about astronauts?) This entails rotational invariance;

separability various dimensions are independent. This entails separability, *e.g.*, what happens in time doesn't depend on what happens in space, *e.g.*, on whether I happen to look through a telescope or a microscope;

linearity if inputs are superimposed I expect the responses to be likewise superimposed. This symmetry is called "linearity". Nonlinearities always imply distinguished parameter ranges, *i.e.*, dedication to specific phenomena;

semigroup property of scaling scale transformations should combine gracefully: whether I change scale in one go or via a number of stages shouldn't matter at all. However, one can't require that scale change can be undone, thus we require only the "semigroup" property;

contrast invariance TV movies look essentially the same on different TV sets, irrespective the fact that no two sets have exactly the same (nonlinear) transfer, or "gamma" and "brightness" settings. Hence it appears prudent to require that the front end be invariant with respect to contrast transformations, where luminance is scaled by an arbitrary power law.

Clearly one could either add to this list or curtail it. One then obtains front ends of various degrees of generality. The list proposed here is extensive enough to constrain the front end structure considerably, yet it yields a structure that is still more universal than any known biological system. Any *real* system will of course be dedicated to the generic environment and lifestyle in which the species evolved: thus there can be *no general theory* of vision in the strict sense, only theories of specific instances.

The present approach attempts to build a starting platform for such more specific theories. Moreover, we don't even attempt a "theory of everything", but build in essential limitations right from the start: For instance we don't address spectral discrimination or binocular information in this chapter. Our approach is very much akin to that of the physicist: for example, the theory of the "ideal gas" is extremely important in physics, despite the fact that no such a thing exists in nature at all and every real gas is an exception! Moreover, the ideal gas does not even properly "condense" to the fluid state, thus it can't be part of any "complete" theory of material constitution. The scientific advance made possible by such fortunate abstractions is evident enough.

The items discussed above are far more intricate than might appear at first blush. Most of them cannot simply be tested empirically, but their value can only be assessed in a much later stage of development of the theory. We lack the space to develop such important considerations here. Just a simple example: one might strike out the linearity assumption by pointing at the neurophysiological literature which indicates that all neural processes are of a highly nonlinear nature. However, it would be fairly easy to implement our constructs in such a way that it would look likewise highly nonlinear to the superficial eye, yet in no way jeopardize the value of the analysis: Such a simple "test" is not decisive at all. (In order to see at least the possibility of such a state of affairs you may think of a linear problem programmed on a digital computer. To the user the system is linear, though all the digital gates and processes that "implement" the system are of an essential nonlinear nature.) Similar considerations apply to the other items.

1.2 Remapping of Dimensions

The representation of any dimension typically includes taking ratios with some fiducial object, singling out an "origin", *etc.* By assumption all origins are equivalent.

For the spatial domain we typically pick (any) fiducial location and call it "the origin". Distances between locations are given as the ratios to the length of some (arbitrary) fiducial "yardstick". Often the natural thing to do is to pick the resolution as a yardstick. Then distances become pure numbers, whereas only the integer parts of these numbers are relevant. (This is loosely speaking of course: we don't imply that you should truncate coordinates. The problem of discretization is an important one that will not be taken into account here.)

Resolution (or "inner scale") is a dimension on its own right. The resolution is the minimum length over which significant changes may be expected. Thus it is a positive number, expressed in terms of the fiducial yardstick. It is convenient to use the highest available resolution as the fiducial value and treat this scale as uniform. There still is a problem: we don't have self-similarity. Self-similarity implies that you take the logarithm of the ratio of the actual resolution to the fiducial one. We call this the "natural resolution parameter". This parameter ranges from minus to plus infinity. The origin of the scale depends on the fiducial yardstick. Since no yardstick is singled out we regard all origins as equivalent. If you express the spatial distances and the resolution in the indicated manner you can no longer find out whether you look through a microscope or a telescope. Thus scale invariance has been arrived at. This trick should be familiar to the neuroscientist as akin to the "Weber-Fechner law" in various sensory domains. It is also familiar in statistics: the only way to express total ignorance for a parameter that may assume all positive real values is to assume a logarithmically uniform prior probability distribution.

Time is a more complicated dimension than space is. We assume that different observers may agree on simultaneity of events, but we don't assume everyone has the same clock. (With "observers" we indicate parts of the front end here. Different parts may use different temporal scales.) We do assume that all clocks are regular though, they only differ in rate ("unit of time") and epoch. (Days since the birth of Christ or lunar cycles since the battle of Hastings will do equally well.) "Regularity" is taken to mean that if two observers record any three events A, B, C (say), then the ratio $(t_B - t_A)/(t_C - t_A)$ will agree for both observers. Let N ("now") denote the present moment. Suppose we point out two fiducial events P, Q (Q later than P) to all observers. We ask all observers to report the time of occurrence of some event E as the number $\tau = (t_E - t_N)/(t_P - t_Q)$. Now all observers agree (you easily check that the numbers τ reported are equal). These numbers are always positive since all events (including the fiducial one) exist in the past. Self-similarity in time is obtained if we take the logarithm and treat this scale as uniform. Then the origin is delayed $t_P - t_Q$, whereas the present ("now") maps to minus infinity, the infinite past to plus infinity. All choices of origin are to be considered equivalent.

Finally we regard the irradiance domain. Irradiance is always positive. Complications are due to the fact that different observers may use different units (such as lux, or Watt per meter squared) and that we require invariance with respect to contrast transformations. A simple way to handle the problem is the following: we designate two points in the input that are at different irradiances I_A, I_B with $I_B > I_A$ (say). We report the numbers $\log(I_C/I_A)/\log(I_B/I_A)$ for the irradiance at any point C (say). These numbers agree even for different photometers (lux or Watt per meter squared) and even under arbitrary contrast transformations. (That is for transformations of the form $I^* = \alpha I^\beta$, with $\alpha, \beta \in (0, \infty)$.)

The invariance is obtained only if the contrast transformation also affects the fiducial irradiances at A, B. An apt choice for the fiducial anchor points I_A, I_B are the 25% and 75% quartiles of the pixel intensities. An automatic gain control following a logarithmic transducer function (like we find in the visual system) will perform essentially the same task.

In most cases the scale transformations are mere formal devices that are most convenient for the description, and there is no obvious need to implement them, often it is not even clear that such has any meaning. (*E.g.*, in the case of a shift of the origin.) In some cases the transformation is easily incorporated in the structure of the receptive fields. (*E.g.*, in the case of the temporal transfer.) However, in the case of the irradiance we have to insist on a hardware implementation at a very early stage because the non-linear transformation doesn't commute with the linear transformations implemented by the receptive fields. Then we may as well forget about this transform in describing the receptive field structure. The only datum needed at the interpretation stage is the fact that irradiance *ratios* map to internal *differences*.

1.3 Local versus Multilocal Operations

Suppose you want to find the value of some scalar field at some location. All you can ever come up with is the average over an area s that is forced upon you by the measuring apparatus. Different pieces of apparatus may allow you to vary s, thus you obtain a one-parameter family of values. Think of the apparatus as an operator that can be characterized by its location and its inner scale (s). The operator takes a bite of a field and spews out a number, the average value over the area s at the location of the operator. Such an operator might well be called a "point operator". Although the point has a size, it "has no parts" as Euclid would have it. For the special linear fields $X(x, y) = x$, $Y(x, y) = y$, the operator yields the numbers x and y say, which are its "Cartesian coordinates".

A "point" is obviously a "local" entity. However, notice that the inner scale is essentially arbitrary, so the average may well extend over the whole image! Points lack internal structure, but they do have sizes.

A "vector operator" also spews out a number when you feed it a field. The number is the average slope or the slope of the smoothed field (the order makes no difference because of linearity) over an area s in the direction of the vector times the modulus of the vector. Again this is a purely local operation. Contrast this with the following method of finding the result of a vector operation: take two points $\mathcal{A}$, $\mathcal{B}$ ($\mathcal{A}$ denotes the tail, $\mathcal{B}$ the tip of the vector) and let the points operate on the field. Subtract the results and divide by the distance of the points. This is a *multilocal* method, and it is fraught with problems. Not only do you need to be able to point out a location ("find back a point", *i.e.*, Lotze's "local sign"(Koenderink 1990; Lotze 1884)), but you need topological expertise ($\mathcal{A}$ and $\mathcal{B}$ have to be close but not coincident) and even a metric (the distance $\| \vec{\mathcal{AB}} \|$).

We do not assume that the front end is *that* sophisticated. Hence we consider only local operations. Multilocal operations have to be implemented at further stages of processing. We don't loose much, because *e.g.*, all the operations from differential calculus and geometry are of a purely local nature. It is perhaps not superfluous to stress the fact that being *local* in the technical sense has nothing to do with "small size" or "simple structure": thus "local processing" in no way rules out the existence of very large receptive fields with complicated internal structures.

We assume all operators to be centered on the same location. Without loss of generality we may take this location to be the origin.

1.4 Scale Transformations

The symmetries of the front end put a very strong constraint upon the possible implementations of point operators. In fact, it is well known that the only admissible structure is a linear operator with Gaussian profile. Thus the action of a point operator $\mathcal{P}$ on a field $f(\mathbf{r})$ defined on the plane $\mathcal{R}^2$ is explained as follows:

$$\mathcal{P}\langle f\rangle = f(\mathbf{r}_\mathcal{P}, s_\mathcal{P})$$

with

$$f(\mathbf{r}_\mathcal{P}, s_\mathcal{P}) = \int_{\mathcal{R}^2} G(\mathbf{r}, s_\mathcal{P})\, f(\mathbf{r})\, d\mathbf{r},$$

where the kernel G describes the "size" and position of the point $\mathcal{P}$:

$$G(\mathbf{r}, s_\mathcal{P}) = \frac{e^{-\frac{(\mathbf{r}-\mathbf{r}_\mathcal{P})\cdot(\mathbf{r}-\mathbf{r}_\mathcal{P})}{4s_\mathcal{P}}}}{4\pi s_\mathcal{P}}.$$

For points in *any* position we note that the convolution $G \otimes f$ predicts the output in any case.

The operation of a *vector* can be explained in various ways, all basically equivalent. If you conceive of a vector as of a "bilocal object" (arrow with tip and tail), then you may define its action on a field as the difference of the field's values at tip and tail of the arrow. If you conceive of a vector as of a *rate*, then you may think of its action in terms of the rate of change of the field as a moving point experiences it when it moves at a velocity given by the vector. If you are more formally minded, then you may conceive of a vector as of a *directional derivative*. Thus the unit vector in the x-direction operates on the field $F(x,y)$ to yield $F(x+1,y) - F(x,y)$ (vector as bilocal object, tip at $x = 1$, $y = 0$), as $\partial F(x+t,y)/\partial t$ (vector as rate of change for the orbit $(x+t,y)$), or as $\partial F(x,y)/\partial x$ (vector as directional derivative in the x-direction). In the limit all these views merge.

Thus the operator $\mathbf{e}_x$ (the unit vector in the x-direction) is explained through its action:

$$\mathbf{e}_x\langle f\rangle = \frac{\partial f}{\partial x}(\mathbf{r}_{\mathbf{e}_x}, s_{\mathbf{e}_x}),$$

where $\mathbf{r}_{\mathbf{e}_x}$ is the position of the tail of $\mathbf{e}_x$, whereas $s_{\mathbf{e}_x}$ is its "size" (as distinct from its modulus, which is unity). The corresponding weighing function is simply

$$G_{10}(\mathbf{r}, s) = \frac{\partial G(\mathbf{r}, s)}{\partial x},$$

and again the convolution $G_{10} \otimes f$ predicts the output in any case. Because of the linearity of the convolution operation you have formally $\partial(G \otimes f) = \partial G \otimes f = G \otimes \partial f$, *i.e.*, convolution and differentiation commute. This simple observation enables us to construct arbitrary differentiation operators. In figure 1 we depict such differential operators of orders zero, one and two for dimension one.

The formalism presented here is based upon the extreme assumptions that the image is the Euclidean plane and that the resolution may range from $s = 0$ (infinite acuity!) to $s = \infty$ (no detail whatsoever resolved!). Of course real life is different. The typical picture has a finite extent, its *scope*, and is given at a finite resolution, its *grain size*. We run

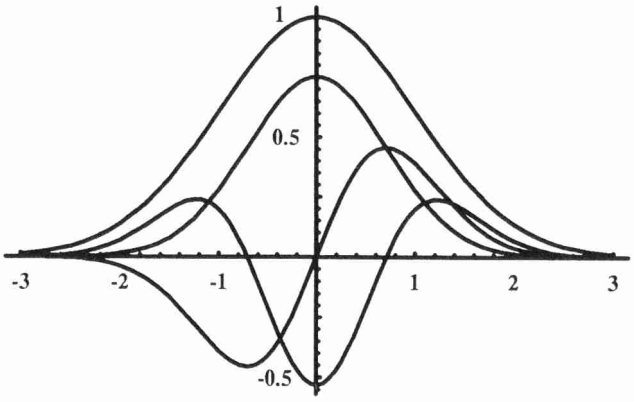

Figure 1: *The differential operators for $s = \frac{1}{2}$ in dimension 1. The operators G_0, G_1 and G_2 are depicted. For very high orders the operators become very wiggly. (The number of zeros equals the order.) Asymptotically you obtain "Gabor functions", i.e., trigonometric functions modulated with a Gaussian envelope. The wide Gaussian in this figure depicts this asymptotic envelope. In a later section we introduce a characteristic function A that exactly equals this asymptotic envelope.*

algorithms that see a limited region of interest (ROI) at a time, the *outer scale*, and use a limited resolution, the *inner scale*. (See figure 2.) If the outer scale approaches the scope you get into the *boundary problem*. If the inner scale approaches the grain size you run into the problem of *insufficient resolution*. Our formalism works fine in cases where resolution and boundary don't pose problems. If one runs into such problems one should strive to increase the scope and/or decrease the grain size, rather than construct fudge "solutions" that will never work well anyway. In the primate visual system algorithms (apparently frozen in pieces of hardware) run in the reasonable regimes (Bijl and Koenderink 1989; Koenderink and van Doorn 1978). When the animal runs against the limits it changes its behavior and strives for more reasonable input (explorative vision). In machine vision many algorithms have been designed to run into the resolution limit. A consequence is the myth that higher order operators are virtually impossible ("not robust"(Horn 1986)). In the primate visual system operators of order four and more are not at all rare and apparently do well (Young 1985; Koenderink and van Doorn 1987; Koenderink 1988).

2 The Blob Hierarchy

Intuitively, the irradiance distribution is a deeply nested set of light and dark regions. It may be compared with a landscape in which hills mimic the light regions, dales the dark regions. (This metaphor is useful if you try to forget the polarization of the typical landscape: Since water runs downwards and eventually has to go somewhere, there are hardly any pot-valleys, whereas isolated hills are numerous. Here we consider "symmetrical" landscapes.) One possible formalization of the hills and dales structure was pioneered by Cayley and later Maxwell. It divides the landscape into "natural districts" by way of the watersheds or the water courses. Thus you obtain a parcellation in either hills, *or* dales. These parcellations interlock and are in many respects dual.

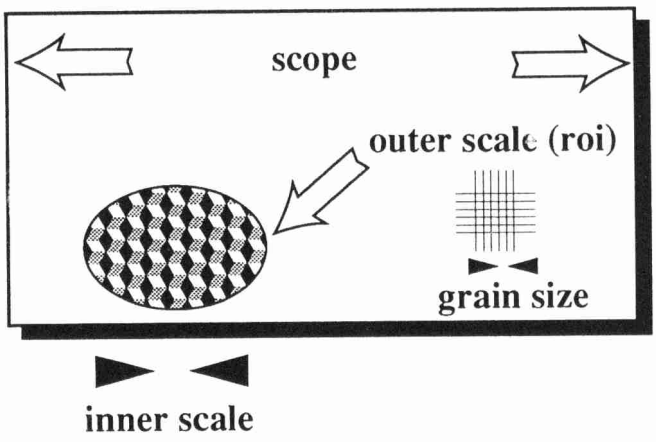

Figure 2: *Definition of cardinal extents. The extent of the whole picture is its scope, the sampling density (e.g., pixel separation) defines the picture's "grain size". Algorithms need input from a region of interest (r.o.i.) that defines the "outer scale", with a resolution that defines the "inner scale". Notice that scope and grain size pertain to the picture, but inner and outer scale to algorithms you may wish to run on the picture. The reasonable regime is that where the grain size is much finer than the inner scale, whereas the scope far exceeds the outer scale.*

For our purposes we need a parcellation into hills *and* dales. The natural way to do this is via the "isophotes", *i.e.*, the loci of equal irradiance. If the landscape is smooth (as we will assume), the isophotes are smooth curves that are either closed, or otherwise end on the boundary. We don't treat the boundary problem here, it is essentially trivial. For a finite number of very special heights, some isophotes may have a double point: this happens exactly if the height is that of a pass. At a pass two hills or two valleys meet. It is also possible that a hill meets a valley, as on the top of a volcano. If the height is taken between two successive pass heights, then the patterns of nearby isophotes are qualitatively identical. (The naïve reader will find these observations substantiated on perusal of a topographic map of some mountainous region containing the lines of constant height above sea level.)

When we say "hill" (or "light blob") we indicate one connected component of a level set, *i.e.*, that part of the landscape that is above a certain (arbitrary) fiducial height. (Similar for the dales or dark blobs.) A hill might well have internal structure (hills and dales), but these are all above the fiducial height. This is the only definition of a "light blob" that makes operational sense. Attempts to define "blobs" in terms of extrema are bound to fail, because of the fact that they are highly sensitive to noise. Even an infinitesimal amount of noise on a hilltop is bound to create any number of spurious hilltops! Thus the only reasonable way to define blobs is via the finite height ranges for which the fiducial isoheight curve landscape contains no critical points (extrema or saddles), but only finite slopes. (Eklundh's chapter in this book develops the notion of "blobs" in more detail.)

For every fiducial height you obtain one possible set of hills and dales. These are juxtaposed and nested to arbitrary depth. The structure is that of a tree (in the sense of computer data structures). For a different fiducial height you obtain a different structure. When two fiducial heights are divided by just one pass height, then the patterns differ

only by a single "blob merge" or "blob split". If they are further apart the structures can be very different indeed. You may characterize the complete landscape in a qualitative fashion if you pick a series of fiducial heights such that each pair of successive heights is divided by a single pass height. Then you may set up a geneological tree for each blob, specifying its parents and offspring as the fiducial level is changed. When you raise the fiducial level hills divide into two or vanish and dales suddenly appear or merge (Possibly with themselves, like a snake biting itself in the tail. Think of a lake that grows from a round basin into a circular one with an island in the middle.) When you lower the level the opposite happens.

Note that you never create a hill out of the blue (that is: creation occurs only via splitting) on raising the level, nor a dale on lowering it.

When you change the *resolution* the landscape itself changes. Intuitively, the landscape should *simplify* when you turn on the blurring. This is very useful because you obtain a "generalized", "abstracted" view of the landscape that will allow you to pick out major features without being bothered with detail. You'd be unpleasantly surprised if blurring would actually introduce detail. Yet this is what happens when you use just any old blurring method (Koenderink 1984). It has been shown that the only way to blur correctly is to let the landscape "diffuse" in the technical sense. (You apply the diffusion equation.) This guarantees that summits decrease and immits (a term introduced by Maxwell in the 19th century as the opposite of a summit) increase on blurring, thus the landscape truly "erodes". The diffusion equation is

$$\triangle I(x, y, s) = \frac{\partial I}{\partial s},$$

where $\triangle I(x, y, s)$ denotes the Laplacean of the illuminance.

This very nice property of scale space is so important that it has given rise to a considerable literature, including reports that "causality" might be violated, even for the diffusion process. Such reports arise from careless interpretations of causality. First of all the very structure of the diffusion equation expresses the fact that summits decrease and immits increase under blurring. This means that for any fiducial level you only loose hills or dales upon blurring, they never arise out of the blue. This condition can never be violated. Please note that this does in no way imply that the number of hills can't increase on blurring. Because splitting processes are not at all rare, it often happens that a hill will give rise to two hills. Such splittings are intuitively reasonable if you think of dumpbell shaped hills (or dales). A saddle-extremum point may also be generated "out of the blue", but this process will *never* introduce new blobs. Progressive blurring will eventually reduce *any* image to a featureless one, thus extrema generated through blurring must vanish again on even further blurring.

The intuitive notion of causality in scale space is that of the natural "erosion" of a landscape: hills wear away and become progressively lower, whereas (pot-)valleys fill up and their bottoms become progressively higher. Examples of "noncausal" behavior would be catastrophic events such as the genesis of a mountain (volcanic eruption), or of a hole (as when a subterranean cavity collapses). The mathematical scale space structure essentially vetoes such events. For our present purposes "causility of scale space" is perhaps best defined as "evolution according to the diffusion equation". However, it would be foolish to forget the intuitive notion altogether.

3 Differential Image Structure

The standard way to study any entity in the neighborhood of a point is to differentiate it. The paradigmatic example is the "first derivative", which is the best linear approximation for arbitrarily small neighborhoods. This immediately generalizes into Taylor's expansion, which expresses the local structure in polynomial form. Can one do anything like that at a finite level of resolution? This is necessary because one can't make operational sense out of derivatives given real (observed) signals.

The gamut of "neighborhood operators" from image processing comes near to the answer, but the standard approach fails miserably on many counts. The reason is that the apparatus used tends to be arbitrary, constrained only by various types of *ad hoc* conditions. If you care to proceed in a principled manner, then there is very little leeway in picking the operators.

One cue on how to progress is to observe that blurring and differentiation commute. As observed above you have formally

$$\partial(G \otimes I) = \partial G \otimes I = G \otimes \partial I.$$

Thus the differential operators are just the derivatives of the blurring kernel. Because the Gaussian is the unique blurring kernel that complies with the basic front end symmetries, it follows that only the derivatives of Gaussian operators respect these basic symmetries. If you use difference operators (difference of Gaussians, Sobel operators, Canny edge finders,...*ad infinitum*), you tie yourself to a specific position, orientation, or scale, and you have lost the universal viewpoint. Thus the common view that the type of edge finder is essentially arbitrary and you may apply further constraints in order to obtain desirable properties (good localization, optimum noise rejection, *etc.*) is just nonsense. There is one exception: linear combinations of derivatives of Gaussians are also appropriate, *e.g.*, the Laplacean operator $\triangle G = G_{xx} + G_{yy}$. Such linear combinations may assume many unexpected guises, and it makes sense to try to find a general characterization of their structure.

What is needed is a way to formalize the structure of allowable operators in a general manner. Then we may use the formalism to derive general rules. It is not obvious how to proceed. We have found that a very simple, but illuminating way to handle this problem is to start by defining the local region through an "aperture". This aperture may be interpreted as a weighing function, or "window function", with weights that are everywhere very small, except in the fiducial region. A judicious choice of the aperture leads to a simple formalism. An obvious choice is to pick an aperture that transforms in a simple manner under diffusion.

Somehow the operators should have a spatially limited support, so let us define this support by way of the aperture $A(\mathbf{r}, s)$. If you have any image $I(\mathbf{r})$, then $A(\mathbf{r} - \mathbf{r}_0, s) \cdot I(\mathbf{r})$ picks out a small *subimage* (of diameter proportional to $\sqrt{s}$) of the image, centered at $\mathbf{r}_0$. One way to pick A is to watch the behavior of derivatives of Gaussians: for a high order of differentiation they are like Gabor functions with envelope $\exp(-\mathbf{r} \cdot \mathbf{r}/8s)/8\pi s$. So let us fix A to this envelope. (See figure 1 and its legend.) The choice is the more a particularly nice one because blurring the window function A clearly leaves its shape invariant. The support of A is conveniently measured by its effective radius, which equals $\sqrt{8s}$.

The next step is to observe that the operators should be scale invariant and that the blurring should be a "causal" one in the above defined sense. Hence the operators should satisfy the diffusion equation. Moreover, they should respect the front end symmetries.

At this point it is convenient to introduce a notion of "natural coordinates". Notice that the factor $\sqrt{4s}$ is a length and can be interpreted as the *equivalent radius*, *i.e.*, the radius of a "pillbox" operator with the same integrated weight and same maximum amplitude as the Gaussian corresponding to the point operator $\mathcal{P}$. We denote this "equivalent radius" with $r_E = \sqrt{4s}$. The natural coordinates are defined as $\xi = \mathbf{r}/r_E$. Please notice that the *point operator* is modulated with the exponential factor $\exp -\xi \cdot \xi$, whereas the *aperture* is modulated with the exponential factor $\exp -\xi \cdot \xi/2$. This probably appears confusing at first blush, however, it is simply a reflection of the fact that the asymptotic envelope of a high order derivative of a Gaussian is *broader* than the width of that Gaussian itself. (The reader who doubts this is in for a somewhat extensive calculation that would be rather out of place here.)

That the operators should respect the front end symmetries means among more that the space dependence must be some function of the natural coordinates, say $\Phi(\xi)\, A(\xi)$. The operator in terms of the normal coordinates and the inner scale will be designated $\Psi(\mathbf{r}, s)$.

Finally, we note that the operators may have various orders. For instance, an "edge detector" is like a first derivative and must have order one. An operator of order n will have the dimension of length to the power minus n. This is illustrated by the derivatives of Gaussians, whose amplitudes vary with scale as r_E^{-n}.

Thus we are let to the *Ansatz*

$$\Psi_n(\mathrm{r}, s) = (\frac{1}{\sqrt{4s}})^n \Phi_n(\frac{\mathbf{r}}{\sqrt{4s}}) A(\mathbf{r}, s) = r_E^{-n-2} \Phi_n(\xi) \frac{\exp \frac{-\xi \cdot \xi}{2}}{2\pi},$$

with the understanding that Ψ_n satisfies the diffusion equation. Since the derivatives of Gaussians can be written in exactly this manner, we may rest assured that solutions exist. The mathematically inclined reader may want to find a few examples of the Φ_n at this point, starting from derivatives of Gaussians as the Ψ_n. A few of the Φ_n are illustrated in figure 4.

Substitution of the expression Ψ_n in the diffusion equation yields a very simple PDE (partial differential equation) for the form factor Φ in terms of the natural coordinates, namely

$$\triangle \Phi_n + ((2n+1) - \xi \cdot \xi) \Phi_n = 0.$$

The exact form of the equation is of little importance (except for the fact that it happens to be a famous equation for mathematical physics (Powell 1961), so the solutions are well known), the important fact is that we have arrived at a PDE in natural coordinates at all. There is a rich theory of PDE's, thus we are in a position to formulate general statements concerning the set of allowable operators at large. This is remarkable and important in view of the fact that no such a general theory of neighborhood operators exists today, and in fact the set of operators in practical use is an odd lot with hardly any internal consistency. We now see that all linear neighborhood operators that respect the front end symmetries are local solutions (*i.e.* vanishing at infinity) of one and the same PDE. Some very important facts we get "for free" from the theory of PDE's are due to the fact that the PDE has a set of solutions that is complete and allows an orthonormal base. This has some important consequences:

- any local image structure ($A \cdot I(\mathbf{r})$) can be represented completely via linear combinations of the Φ_n.
- the coefficients are the result of applying the Ψ_n to the image $I(\mathbf{r})$.

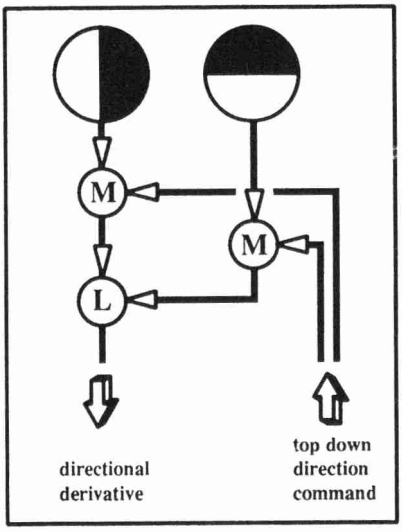

Figure 3: *A simple network that can simulate an edge finder of arbitrary orientation. The network is based upon two orthogonal edge finders, these have to be conceived of as coincident. (They have not been superimposed in the figure for the sake of clarity.) A command input sets the desired orientation. The elements designated "M" are multiplicative nodes, the element marked "L" is a linear (additive) node.*

- when one truncates the representation to some finite order, one obtains the best approximation to the local image structure in the least squares sense. Note that the Φ_n are orthogonal, whereas the Ψ_n are not. Usually one will normalize the Φ_n and thus obtain a convenient orthonormal basis.

- solution of the PDE in various coordinate systems in which the PDE is separable yields families of operators with specific symmetry properties. For instance, we may construct solutions with radial or azimuthal symmetry.

- rotations of the coordinate systems can easily be accommodated via orthogonal transformations of the observations per order. This means, for instance, that you need only two edge finders: any other can be obtained by rotation, that is linear combination. (See figure 3.) Likewise you need only three line finders, *etc.*

- transformations between representations (*e.g.*, polar to Cartesian or vice versa) can be done through linear orthogonal transformations per order.

- The PDE allows us to construct a complete taxonomy of receptive fields.

This immediately yields a tremendous increase in power of the neighborhood operator approach.

When the PDE is solved in Cartesian coordinates we obtain simply the mixed partial derivatives of the Gaussian as solutions. (See figure 4.) Many of such solutions are reminiscent of receptive field profiles as reported in the primate visual system (Jones and Palmer 1987a, 1987b; Young 1985), though by no means all of the possibilities appear to have been reported. When you solve the PDE in other coordinate systems you obtain families of operators with specific types of symmetry properties. For instance, the simplest

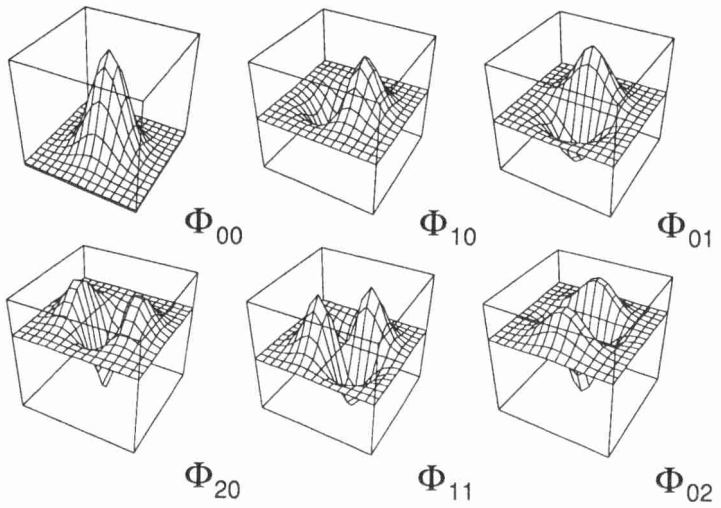

Figure 4: *The solutions of the partial differential equation that governs the allowable operators for orders up to the second in dimension two. The PDE has been separated and solved in Cartesian coordinates. In this simple case the operators are just the mixed spatial derivatives of the zeroth order Gaussian kernel.*

types of center-surround organized receptive fields appear as lowest order radial derivatives of the polar family. The structures of simple symmetry appear to have obvious application in image processing and are perhaps worth a search in neurophysiology: notice how the theory permits a "prediction" of local structure reminiscent of the prediction of elementary particles in high energy physics.

Application of a Cartesian operator, *e.g.*, of order (n, m) yields the exact (!) mixed partial derivative of order (n, m) of the image at the resolution level of the operator. This puts us in a position to implement any formula for differential geometry in terms of these operators: every derivative is the output of an operator, the structure of the formula then specifies how these outputs should be combined. Since the operators are pieces of hardware, we obtain a wiring scheme for a (nonlinear) network that implements the differential geometrical formula.

A simple example is "boundary curvature". This may be taken to mean the curvature of the local isophote. The formula from differential geometry is

$$\kappa(x,y) = \frac{-I_y^2 I_{xx} + 2I_x I_y I_{xy} - I_x^2 I_{yy}}{(I_x^2 + I_y^2)^{\frac{3}{2}}}.$$

(Consider an equation like this to be taken "straight from the book". The reader may want to check, but the derivation is essentially irrelevant to the present discussion.) It is straightforward to "compile" the formula into a little nonlinear network (Koenderink and Richards 1988) (see figure 5), by implementing the I_{xx}, *etc.*, with the Cartesian operators, and addition, multiplication and exponentiation via hardware adders, multipliers and nonlinear transfer elements. However, in practice you would compute the pair of numbers $(-I_y^2 I_{xx} + 2I_x I_y I_{xy} - I_x^2 I_{yy})$ and $(I_x^2 + I_y^2)^{\frac{3}{2}}$. At some higher stage we have to make sure that there *exists* an edge in the first place (*e.g.* that $\| \nabla I \| = (I_x^2 + I_y^2)^{1/2}$ is locally very high) and then the curvature follows immediately. The front end might just compute the

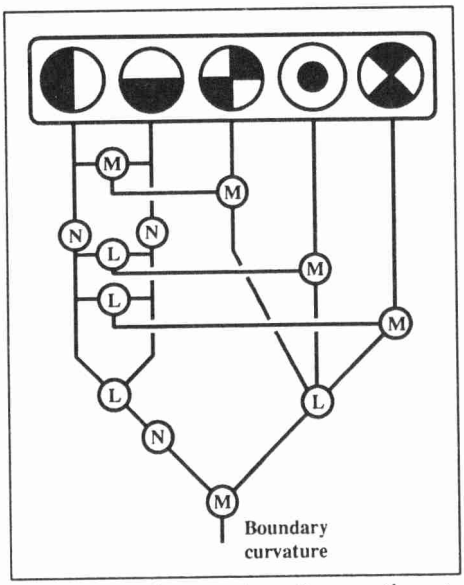

Figure 5: *A network that computes edge curvature. Because the output is the curvature of the local isophote the network is insensitive to arbitrary (non-degenerated) intensity transformations. The input elements are of orders one and two. They must be conceived of as coincident. (They have not been superimposed in the figure for the sake of clarity.) Notice that we picked input operators of polar form. This makes no difference because the polar operators are just linear combinations of the Cartesian ones. The elements marked "M" and "L" have the same meaning as in figure 3, the elements marked "N" are nonlinear transducers.*

two numbers at each point of the image and pass them on as a local description of the illuminance distribution.

The example is of course a trivial one. However, it is easy enough to construct more complicated geometrical invariants with various desirable properties. Such invariants may lead to very complicated constructions indeed. In order to illustrate this we give an almost arbitrary example that is far more complicated than anything being attempted in image processing today: in a (suitably rotated) coordinate system such that $I_x = 0$ and $I_y > 0$, the expression

$$\mathrm{sgn}(I_{xx})\,(-9\,I_{xy}^2\,I_{xx}^2 + 9\,I_{yy}\,I_{xx}^3 - 18\,I_y\,I_{xx}^2\,I_{xxy} +$$

$$18\,I_y\,I_{xy}\,I_{xx}\,I_{xxx} - 5\,I_y^2\,I_{xxx}^2 + 3\,I_y^2\,I_{xx}\,I_{xxxx})/(9\,I_y^{4/3}\,I_{xx}^{8/3})$$

is a curvature measure that is invariant against arbitrary area preserving affinities. (The so called "affine curvature", look for it in any book on advanced differential geometry of the plane. This too is just an example: no need to check if one is not so disposed.) Again, it is straightforward to compile the formula into a network. The resulting network has a very intricate nonlinear structure, and it is *a priori* highly unlikely that standard electrophysiological methods would suffice to unravel its machinery. (This remark may perhaps help to cure the common but mistaken notion that linear operators must lead to trivial results.)

4 What the Front End Should Encode

The front end should encode local image structure allowing the relevant local properties to be computed from intensive, rather than extensive data, thus obviating the need for geometrical expertise (such as local sign or the ability to judge parallellity of orientations at different locations). This means that the front end should encode up to an order that will allow such computations. More complicated properties will have to be computed multilocally, which is bound to require far greater effort and may well turn out to be impossible.

The difference between purely local and multilocal methods perhaps needs further elaboration: we provide an example. Consider boundary curvature again. The method explained above uses second order structure to compute the boundary curvature from purely *local* data. Another method would employ only first order structure. This suffices to find the orientation of the boundary (first order) along the boundary (zeroth order). The rate of change of orientation along the boundary is again the boundary curvature. Notice that this method allows us to compute boundary curvature if we: 1. Can compare orientation at different places, and 2. Are able to measure finite distances between distinct locations. Thus this *multilocal method* asks for rather sophisticated geometrical expertise. (For instance, one needs methods to calibrate distances, transport orientation along curves, *etc.*). In general purely local methods are conceptually much simpler and much more plausible from the viewpoint of physiological implementation. If the brain indeed prefers the simple, local algorithms, then this has important consequences for the repertoire of the front end.

One requirement that has not been introduced so far is that the front end should encode in a *coordinate free* manner. This is again a consequence of the general requirement of non-commitment. For instance, although two edge finders (in x and y-direction say) suffice for the first order, one should not commit oneself to a specific choice of x-y-axis.

In some cases the requirement can be met fairly easily. For instance, in the even orders there exist linear combinations of operators that are isotropic and thus require no special choice of coordinate system. An example is the Laplacean operator. A general way to deal with the problem is not to use the x-y-axis, but an infinite set of axes, caring equally for all directions. Let the azimuth be denoted α (angle with the positive x-axis in direction of positive y-axis say). Then instead of having edge finders in the direction $\alpha = 0, \pi/2$, you have a continuous uniform distribution of them over the range $\alpha \in (0, 2\pi)$. Instead of two numbers you now have an activity function $f(\alpha)$ (say) as the result of running your edge finders locally. (Since the activity is actually present, we may forget about the specific definition of α!) It is easily shown that all the information is contained in the first order Fourier terms of $f(\alpha)$. Thus the activity severely over-represents the first order. However, since the Fourier coefficients are weighted averages of the activity you have obtained a very robust, stable and coordinate independent representation. If hardware is cheap, then the over-representation may well turn out to be an advantage.

The same trick works for any order. Use n^{th}-order directional derivatives in all directions. The information is contained only in the $(n, n-1, n-2, ...)^{th}$-order Fourier coefficients of the resulting activity. (Of course these coefficients depend on the coordinate system.) For the 2^{nd}-order one has a set of "line finders" which closely resembles (part of the structure of) a hypercolumn of simple cells of the primate visual cortex.

The activities of oriented directional derivatives are more complicated than just numbers: there are $(n+1)$ Fourier coefficients of the n^{th}-order that transform as a single entity and depend on the choice of the coordinate system. For many purposes one would

like a set of scalars that have a meaning independent of the coordinate systems. Such scalars exist, they are known as "differential invariants". We have already met a few examples, *e.g.* , $\| \nabla I \|$ that is the magnitude of the gradient, $\triangle I$ that is the Laplacean, and $(-I_y^2 I_{xx} + 2I_x I_y I_{xy} - I_x^2 I_{yy})(I_x^2 + I_y^2)^{-3/2}$ which has to do with isophote curvature. Although the coordinate axes (x, y) formally appear in these expressions, these numbers are actually independent of the specific choice of axes. It can be shown that complete sets of invariants of order n can be constructed, *i.e.*, such that any invariant of order n or less can be expressed in terms of members of the set. Several representations are possible. These invariants are nonlinear combinations of partial derivatives. It is possible to construct polynomial representations, although these need not be the most desirable in any case. Thus they cannot be implemented as simple weighted sums, but only as highly nonlinear combinations of the results of applying variously modulated weights.

5 Spatiotemporal Operators

The basic spacetime operators can easily be constructed via multiplication of spatial and temporal operators. Any operator is characterized by its spatiotemporal location (point and moment), spatial resolution, temporal resolution and delay, spatial and temporal orders and types (*e.g.* polar or Cartesian). Thus the space and time parts are always separable in these linear machines. Separability may be lost in the invariant combinations though. We have not come up with a complete set of spatiotemporal invariants so far. We have merely implemented local operators for motion, and the affine structure of image flow (divergence, vorticity and shear). In the latter case the operators work directly on local spatiotemporal derivatives of the image irradiance, *e.g.*, divergence is not computed via velocities (although this would also be a route worth exploring).

6 The Significance of an Observation

Until now we have mainly discussed the nature of the sampling and representation, *i.e.*, more or less the transformations to be found in the bottom up stage. Now we change gears and perspective and try to specify what can be asserted concerning the input, given a certain activity in the front end. This is more like a top down inference. This is important, because the *meaning* of the front end activity is ultimately contained in the constraints this activity puts on the assertions that can be made concerning what is really "out there". Since we *reverse* the direction of reasoning we have to switch to novel methods of description. In this section we provide a possible way of approaching these important issues. In the literature such problems are typically skipped altogether.

Suppose one has a set of operators up to (and including) order n. To be specific we take $n = 2$, but the reasoning works for any order. There are $(n + 1)(n + 2)/2 = 6$ independent operators of order $n = 2$. Running these operators locally on the image yields an ordered set of 6 numbers, $\{a_{00}, a_{10}, a_{01}, a_{20}, a_{11}, a_{02}\} = \mathcal{O}^2$ say. We refer to the set $\mathcal{O}^n$ as an "observation of order n". In the literature one would probably refer to such a set as a "feature vector', a term that had better be avoided in view of the fact that $\mathcal{O}^n$ doesn't transform as a vector. Notice that $\mathcal{O}^n$ is just a description of the activity in the front end hardware for a given input image.

The numbers depend upon the particular set of operators. If you had happened to pick a rotated set, a polar instead of a Cartesian representation, *etc.*, then you would have obtained different numbers, although the meaning of the observation would have

been the same: the meaning is a function of the input image only and can't depend on the arbitrary choice of representation. The formal way to avoid this flaw is to use the mathematical notion of a *jet*. The "n-jet" of images at a location is the equivalence class of all images that would yield the same response for some given set of n–order operators. Such images necessarily agree among one another in all spatial derivations up to – and including – the order n. The particular set of operators is irrelevant. The images have the same initial terms in their local spatial Taylor series development. The "n-jet-space" at the location is the space of all n-jets, and the "n-jet-bundle" of the visual field is just the visual field with a n-jet space attached to each spatial location. This formal mathematical construct is very similar to the neurophysiological concept of the columnar structure of V-1: the hypercolumns are the hardware implementations of the local (4?-)jets.

Consider the following problem:

Given an observation $\mathcal{O}^n$, what can one assert about the image?

This question is rarely raised, which is remarkable in view of the fact that one most likely wants to come up with some description of the actual input. The reason is probably that very little indeed can be asserted about the image: typically *infinitely* many images can be constructed that would have yielded the same observation.

Some terminology: two images that yield the same observation are called "metameric". The equivalence class of all images that yield some given observation is referred to as the metamere specified by that observation. A metamere can be specified by giving any member, by specification of a canonical member, or by some algorithm that allows one to decide whether any given image belongs to the class. We use a canonical representative.

The question on the significance of an observation then is to solve the following problem (see figure 6):

- find all metameres and how they partition the space of all images into cells,
- find a canonical image (*e.g.*, the simplest one in some sense) for every metamere.

A solution of this problem completely characterizes the "observational power" of the n-jet space.

These problems are related to the question of whether the local operators can be said to be "feature detectors". *E.g.*, does an edge finder find edges? Yes, *if an edge is defined as that what an edge finder finds.* But what if an edge finder responds equally well to bars (as all edge finders do when the bar is broad enough)? Although the questions as posed here are silly, the problem matter is really serious. The present section bears on this.

One way to obtain an insight in the nature of metamery is to consider the construction of "invisible images". An invisible image yields an observation composed of all zeros, just like an image that were constantly zero (flat black) would yield. It is easy enough to construct examples. For example, the image $I(x, y) = x$ is invisible for the point operator at the origin. (The simplest case of the zeroth order jet.) Notice that this "image" has *negative* irradiances at $x < 0$ and thus is a physical impossibility. This is true for every invisible image. Thus the invisible images don't occur in isolation. However, you can *add them* to any given picture, as long as the total intensity remains positive. For instance, the image $I(x, y) = I_0 + \mu x + \nu y, I_0 > 0$ is a possible image near the origin. The point operator can't distinguish such images from the flat image $I(x, y) = I_0$, whatever the values of μ and ν are. Thus we have constructed an *infinity* of metameres for the zeroth order jet.

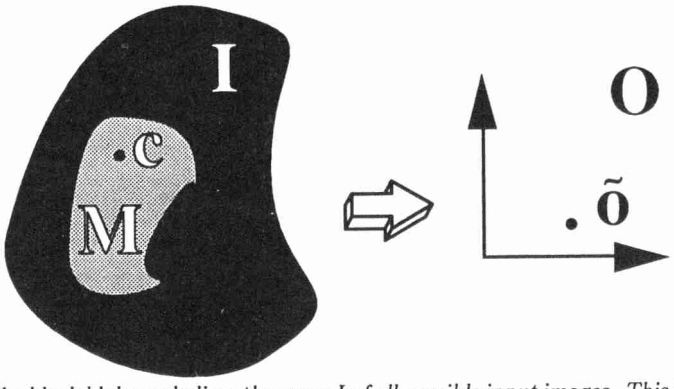

Figure 6: *The black blob symbolizes the space I of all possible input images. This is a very large (infinitely dimensional Hilbert) space. The space O is the space of all possible observations of a given order. These are finitely dimensional Cartesian spaces.* (E.g., *for order two in dimension two the space has dimension six.) An observation* $\tilde{o}$ *(say) could have been caused by many different input patterns. This collection of metameres is a (possibly very large) subset M of the space I. In this subspace you may pick any member c (say) as a canonical representative of the input for this observation. In practice you will characterize the canonical representative through very simple properties.*

The construction of "invisible images" may seem like a mere mathematical oddity. However, because of linearity, any two indistinguishable images must differ by exactly such an invisible image. Thus the images corresponding to given observations are determined only *modulo* the set of all invisible images. Consequently the set of invisible images neatly sums up the degeneracy and it makes a lot of sense to study this set in detail.

Consider the second order jet. There are six degrees of freedom. (*E.g.*, the Taylor expansion of the image, truncated after the second order, has six terms.) Suppose you are handed an *arbitrary* sextuple of numbers. Do they always comprise a possible observation? In order to answer this question we need a few preliminary insights.

First notice that the observations are obtained from the images via purely linear operations. Thus the value of an observation is simply multiplied by a factor when I multiply the image by that factor. If I add two different images, then I can find the observation by addition of the observations of the two original images. Not just any linear combination of images is a possible image though: *E.g.*, minus an image would be an image with negative irradiances, which is physically impossible. If A and B are possible images then the linear interpolate $\lambda A+(1-\lambda)B$ ($\lambda \in [0,1]$) is also a possible image though. But that means that if two observations belong to possible images (have good interpretations), then all observations on the line segment connecting these observations in observation space must also correspond to possible images. Thus the set of observations that corresponds to possible images is geometrically *convex*. (If two points belong to it, then the average of the two also belongs to it.) Thus the boundary of the set of possible observations will be the boundary of a convex body in observation space.

Observations that lie *outside* this convex volume in observation space do not correspond to possible images, whereas interior points admit (in general) of an *infinity* of possible interpretations. The boundary points are special: it can be shown that these are the *only* observations that admit of *unique interpretations*, *i.e.*, there exists only a single possible image that leads to the boundary observation. This makes the boundary obser-

vations important in practice: if an observation is "near" a boundary observation, then it must admit of an interpretation that is "almost unique", *i.e.*, the degree of metamery must be small. That a boundary observation admits of only a single interpretation is shown by explicit construction of the singular image (see below).

If the images are subject to further constraints it is possible to be more specific. Consider the case that is most frequent in practice: you know a priori that the irradiance is strictly limited to the range $(0, I_{\max})$. This further constraint shrinks the set of possible images in image space as well as the set of observations that admit of a possible interpretation to much smaller volumes.

For this constraint, and order two, it is possible to calculate exactly the nature of the boundary observations. It is especially interesting to consider the unique interpretations of the boundary observations. We find that these special images are blobs with a quadric outline, with vanishing irradiance on the exterior, $I_{\max}$ on the interior. (Thus they are *binary* images with elliptic, circular, parabolic or hyperbolic edges.)

The reasoning that leads to such conclusions is related to Schrödinger's and Ostwald's methods in the theory of object colors (Schrödinger 1920a, 1920b; Ostwald 1916, 1917). (Object colors involve the case of order two in dimension one.) Although the proof is not very difficult we do not give it *in extenso*, only the flavor of it.

The proof depends on the fact that the observation space for the second order is six-dimensional. This means – among other things – the following: given any sextuple of lightpoints on a black fond (in general position) you can achieve *any* observation by judicious adjustment of the intensities, although it will often be the case that one or more of the "intensities" will turn out to be negative. It is easy enough to find the required intensities: it boils down to the solution of six linear equations in six unknowns. This entails that if you can find six points in an image that you can perturb as you please, then you can move the corresponding observation into *all* possible directions in observation space. The only obstructions to this freedom occur if you cannot freely pick six points at which to perturb, or if the set of linear equations turn out to be dependent.

Here is a simple proof, based upon these preliminary observations, that the images belonging to boundary observations are *binary images*:

> Assume (for the sake of argument) that there is a region in which the illuminance *does* take on intermediate values. Then we can pick six points in general position in this region. As an immediate consequence we can perturb the observation in *any* direction in observation space by fiddling with the intensities. But then we can't be on the boundary. *Ergo*, there can't be a region of intermediate intensities; the image must be a binary one. **QED.**

The following reasoning leads to the shape of the circumference:

> Assume that the image is a binary one (either 0 or $I_{\max}$, see above) and let the circumference of the light blob be Γ. We can pick 6 points on Γ as we please and slightly perturb the circumference at these points, moving in or out at pleasure. Again, we can move in *any* direction in observation space, thus we can't be on the boundary in observation space. An exception occurs when the six linear equations involved in this problem are dependent. This case defines the boundary points in observation space. Algebraic examination shows that it occurs for binary images for which Γ is a general quadric. **QED.**

As argued above, it is intuitively reasonable that if you meet with an observation that is close to a limiting one, you may make assertions concerning local image structure with

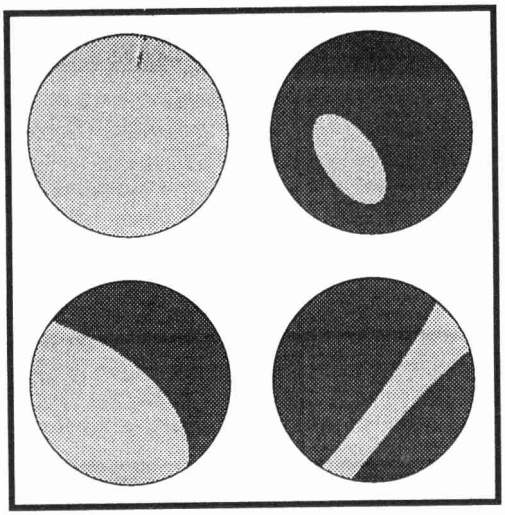

Figure 7: *The essentially different images that can be distinguished by an ensemble of order two in dimension two. These are: the uniform image, the blob, the edge and the bar. Notice that these images may be further characterized through various parameters,* e.g., *illuminance levels, position, orientation, and curvature or shape. Thus the value of an observation may be "a convex light edge of such and so a curvature passing through this or that point at such and so an orientation, separating the levels this and that",* etc.

some degree of certainty (you are clearly certain when the observation is a limiting one). If the observation is at all close to the "medium gray" one (that is $\{I_{\max}/2, 0, 0..., 0)\}$, and is as far from the boundary as you can get), then the degree of uncertainty will be exceedingly high. In that case you commit yourself least if you pick the medium gray image (irradiance $I_{\max}/2$ all over) as the canonical representative. (In general a useful canonical representative is the mixture of an image that belongs to an extreme observation and a gray image. For instance, you may consider mixtures with the image $I = 0$.) It is always possible to find such an image and it is unique.

We now have obtained a handle on the "feature detector" problem: since the collection of all operations up to order 2 can only distinguish between medium gray (by default, rather than evidence!) and quadric blobs (with various degrees of certainty), no single operator (*e.g.*, an "edge detector") can be expected to do better. Since quadric blobs are either ellipses or hyperbolas (Apollonius in Heath 1921; Loria 1910) (with various degenerated singular cases like circles, straight lines, etc.), it is the case that through the 2^{nd}-order the front end encodes only curved edges and bars, or elongated blobs, of various sizes, orientations, and positions. (See figure 7.) Notice that although you may detect an edge, this cannot be said to be the result of running an edge detector on the image.

Because the local jet can only represent image structure *locally*, we only consider the image structure within a disc of total area $8\pi s$.

The term "feature" can be applied to the metameres if you want, or one may further coarsegrain and speak of an "edge", or an "edge at such and so an orientation and position", or even of an "edge at such and so an orientation and position, and such and so a curvature", *etc.* The important thing is that such "features" are canonical names of possibly very large classes of equivalent images. By just declaring something an "edge" one

doesn't change the image, of course. The basic ambiguity in the observation remains pertinent.

If all one knows is that the illuminance is non-negative, then the situation is even more ambiguous. It is still true that possible observations are contained in a certain convex volume, but this volume is now infinitely large. A closer analysis reveals that much of the above analysis remains pertinent though.

It turns out that the general case of order n (say) is not essentially more complicated than the case of the second order ($n = 2$) considered above. The canonical representations are again binary blobs, whereas the boundaries turn out to be general n^{th}-order curves (quadrics for order two, cubics for order three, quartics for order four (Loria 1910; Newton 1706)). Because the general n^{th}-order curves include the m^{th}-order ($m < n$) ones as a proper subset, we obtain a natural hierarchy of features. In the zeroth order you have only the uniformly gray pictures, in the first order you gain straight edges, in the second order bars and convex blobs, *etc.* Once again, it is *not* the case that features are being detected by "feature detectors". For example, suppose we detect an edge in the first order and a bar in the second order for the same input image. Such a case is perfectly possible of course. The first order operators ("edge detectors") will give exactly the same response, irrespective of whether you happen to regard them as members of a first or a second order assembly. Thus the feature "edge" can be "overruled" by the second order operators ("bar detectors"). The response of an edge detector does not by itself define the detected feature.

If we assume the primate visual system to be of at least the 4^{th}-order (Young 1985), then the gamut of possible "features" already runs into the hundreds.

7 Conclusion

We have shown a principled method to obtain a complete characterization of neighborhood operators, their interrelations and transformation properties, as well as their role in inferences concerning the structure of the optical input on the basis of local measurements.

The method is founded upon a limited number of well understood symmetries of the visual front end, such as homogeneity, isotropy, and self similarity. Such symmetries express the preference for minimum commitment, such as preferred positions, directions, or sizes. Together with rather fundamental physical constraints (*e.g.*, the fact that illuminance is necessarily positive and in any given case strictly limited from above) these conditions constrain the possible front end structures to a great extent.

It is remarkable that very general symmetry considerations *constrain* the allowable operators to a great extent. This is essentially due to the "principle of non-commitment": you may well try to design operators with certain desirable qualities that will make their use preferable over the default operators derived from general symmetry principles. Indeed, there has been and still is quite an activity in machine vision in the design of optimum edge and corner detectors, higher order differential operators, and the like. Such operators may well have better localization properties (in the case of edge detectors) or better noise immunity. (Although in the cases studied by us the differences turn out to be hardly worth the effort.) They acquire such new properties by violating the principle of non-commitment though. Indeed, the operators are already constrained by the symmetries and you may only impose other constraints by lifting some of these natural ones. Thus one may gain noise immunity by focusing on a limited scale range, for instance. Then it will turn out that the design of the operators depends on the scale parameter

and that the blurring implicitly involved in any operator will generate spurious detail. To summarize: it may indeed be useful to violate the front end symmetries, especially when you know a lot about the application (*e.g.*, its scale range, *etc.*). You do so at the cost of universality (*e.g.*, true size invariance, *etc.*).

The utility of these notions for image processing and machine vision should be obvious. The applicability as part of a theory of biological front ends is less well established. Clearly the present theory has to be regarded as an "ideal" limiting case. In any real system the symmetries cannot be expected to hold universally, moreover it may be an evolutionary *advantage* to violate these symmetries if the environment is biased towards specific features, sizes, and orientations. The present theory thus appears as a convenient reference from which to depart in all special cases.

References

Apollonius of Pergae, Conics. See: Th. Heath (1921) A history of Greek mathematics, Vol. 2. From Aristarchus to Diophantus. Oxford at the Clarendon Press

Bijl, P., Koenderink, J. J. (1989) Visibility of blobs with a gaussian luminance profile. Vision Res. 29, pp. 447-456

Horn, B. K. P. (1986) Robot Vision. The MIT Press, Cambridge Mass.

Jones, J. P., Palmer, L. A. (1987a) The two-dimensional spatial structure of simple receptive fields in cat striate cortex. J. Neurophysiol. 58, pp. 1187-1211

Jones, J. P., Palmer, L. A. (1987b) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J. Neurophysiol. 58, pp. 1233-1258

Loria, G. (1910) Spezielle algebraische und transzendente ebene Kurven, Theorie und Geschichte. 2nd ed., transl. F. Schütte, Vol. 1: Die algebraischen Kurven. Teubner, Leipzig and Berlin

Lotze, H. (1884) Mikrokosmos. Hirzel, Leipzig

Koenderink, J.J. (1984) The structure of images. Biol. Cybern. 50, pp. 363-370

Koenderink, J. J. (1990) The brain a geometry engine, Psychological Research 52, pp. 122-127

Koenderink, J. J., van Doorn, A. J. (1978) Visual detection of spatial contrast; Influence of location in the visual field, target extent and illuminance level. Biol. Cybern. 30, pp. 157-167

Koenderink, J. J., van Doorn, A. J. (1987) Representation of local geometry in the visual system. Biol. Cybern. 55, pp. 367-375

Koenderink, J. J., Richards, W. (1988) Two-dimensional curvature operators. J. Opt. Soc. Am. A5, pp. 1136-1141

Newton, I. (1706) Enumeratio linearum tertii ordinis, Londini

Ostwald, W. (1916, 1917) Das absolute System der Farben. Z. physik. Chemie 91, p. 132, 1916, and 92, p. 222, 1917

Powell, J. L., Chasemann, B. (1961) Quantum Mechanics. Addison-Wesley, Reading Mass.

Schrödinger, E. (1920a) Theorie der Pigmente von grösster Leuchtkraft. Ann. Physik 62, p. 603

Schrödinger, E. (1920b) Grundlinien einer Theorie der Farbenmetrik im Tagessehen. I, II. Ann. Physik 63, p. 397 and p. 427

Young, R.A. (1985) The gaussian derivative theory of spatial vision: Analysis of cortical cell receptive field line-weighting profiles. General Motors Res. Tech. Rep. GMR-4920, May 1985