Biol. Cybern. 50, 363–370 (1984)

Biological Cybernetics

The Structure of Images

Jan J. Koenderink

Department of Medical and Physiological Physics, Physics Laboratory, State University Utrecht, The Netherlands

Abstract. In practice the relevant details of images exist only over a restricted range of scale. Hence it is important to study the dependence of image structure on the level of resolution. It seems clear enough that visual perception treats images on several levels of resolution simultaneously and that this fact must be important for the study of perception. However, no applicable mathematically formulated theory to deal with such problems appears to exist. In this paper it is shown that any image can be embedded in a one-parameter family of derived images (with resolution as the parameter) in essentially only one unique way if the constraint that no spurious detail should be generated when the resolution is diminished, is applied. The structure of this family is governed by the well known diffusion equation (a parabolic, linear, partial differential equation of the second order). As such the structure fits into existing theories that treat the front end of the visual system as a continuous stack of homogeneous layers, characterized by iterated local processing schemes. When resolution is decreased the images becomes less articulated because the extrem ("light and dark blobs") disappear one after the other. This erosion of structure is a simple process that is similar in every case. As a result any image can be described as a juxtaposed and nested set of light and dark blobs, wherein each blob has a limited range of resolution in which it manifests itself. The structure of the family of derived images permits a derivation of the sampling density required to sample the image at multiple scales of resolution. The natural scale along the resolution axis (leading to an informationally uniform sampling density) is logarithmic, thus the structure is apt for the description of size invariances.

1 The Problem of Scale and Resolution

In every imaging situation you have to face the problem of *scale*: a given image has a limited extent or window (the "outer scale") as well as a limited resolution (the "inner scale"). These limits are set by the "format" of the image, e.g. by the size of the photographic plate and the graininess of the emulsion, the number and spacing of photosensitive elements of a CCD array, or, in the case of the visual system, the discrete structure of the retinal receptive fields and the extent of the retina. In a number of situations the inner scale is determined by the structure of the radiation itself, e.g. in low-luminance situations (night vision, image intensifiers) or scintigraphy (where the number of gamma – quanta available is limited by dosimetry). In a great many applications the inner and outer scales are set by the subject matter rather than the image format, e.g. a treetop does not exist on the scale of the leaves nor on that of the forest. (You typically define treetops as features in volumes with an outer scale of 10 m and an inner scale of 10 cm say.)

In all of the latter cases the problem of setting outer scale (that is finding the subject matter, "identification") and inner scale (morphometric characterization or "localization") can be acute. This is especially true in automatic image processing, much less so in vision: the human eye seems to possess an uncanny aptitude to "zoom in" on the right range of scale. Thus, for instance, to locate the heart on a cardioscintigram you blur the image, then to study the shape of the left ventricle you increase resolution until the photon noise becomes really objectionable (Hay and Chesters, 1977). Thus you probe what may be called the "deep structure" before dealing with the "superficial" structure (at one level of resolution). Similar problems are well known in other fields, e.g. biology, astronomy.

If you have no a priori reasons to look for certain features, then you cannot decide on the "right scale". (Except in certain trivial cases, e.g. once you resolve individual quantum events it is useless to increase resolution any further – regardless of subject matter.) Thus if you aim to retain *all* available structure, and yet

want to vary the resolution (e.g. in order to be able to identify global objects through blurring), then you must treat the image on all levels of resolution simultaneously. Several attempts to do so have been published (Hay and Chesters, 1977; Burt et al., 1981; Witkin, 1983). The challenge is to understand the image really on all these levels *simultaneously*, and not as an unrelated set of derived images at different levels of blurring: this presupposes the existence of links, or "projections" between the different levels of resolution. The obvious way to proceed appears to be:

1. Embed the original (or "primal") image in an one-parameter family of "derived" images. The parameter measures resolution, or inner scale. The outer scale determines how far to proceed. (For inner scale can never exceed outer scale, the simplest derived image contains just one logon or structural degree of freedom.)
2. Study the family as a family, i.e. define deep structure, the relations between structural features of different derived images.
3. In a latter phase of this program (not covered in the present paper) these mathematical structures may be incorporated in more detailed mechanistic models of the visual system composed of homogeneous processing layers with a specific across-layer structure.

In the sequel I show that under a few rather general constraints there exists really only one reasonable way to generate the one-parameter family and that the induced deep structure can be used to define the "projections" unequivocally.

2 The Unique One-Parameter Family Generated by an Image

For the present discussion an "image" is just a real function of two real variables:

$$L: R^2 \to R$$

$$L(\mathbf{r}) = L(x, y) = \lambda \qquad \mathbf{r} \in R^2,\ \lambda \in R\,.$$

The coordinates (x, y) are understood as the Cartesian coordinates in the image plane, the value λ will be called the "luminance" here – for ease of reference – but may be interpreted in many different ways. I shall not require λ be positive, e.g. a "reference luminance" may be subtracted.

The aim is to define a real function K of three variables

$$K: R^3 \to R$$

$$K(\mathbf{R}) = K(x, y, z) = \Lambda \qquad \mathbf{R} \in R^3,\ \Lambda \in R$$

in such a way that $K(x, y, 0) = L(x, y)$ for all x, y, and such that the parameter z measures inner scale. I require that the family depends "causally" on the primal image, i.e. I require the vertical derivative K_z at any level to be given by a functional that depends solely on the function (or derived image) $K(x, y, z = \text{const})$. The problem then is how to express K_z in terms of the derived image at a given level. It will be shown that this can only be done in essentially a single sensible way.

Most persons experience no difficulties when asked to point out "the same" features in two photographs that differ with respect to the amount of blurring if these features are sufficiently coarse. It seems natural to identify light spots when they occur at similar locations as really the *same spot*, and our confidence is increased when we identify configurations of light and dark spots that show similar spatial relations. Let us then start by identifying a pixel (x', y') at resolution z' with a pixel (x, y)

$$K(x', y', z') = K(x, y, z) \quad \text{(metrical identity)}$$

$$[(x'-x)^2 + (y'-y)^2]$$

is a local minimum (structural proximity).

Note that it is not at all guaranteed that such a mapping always exists, in fact a given luminance at some level of resolution need not at all survive if you blur that image. Here I introduce the first hypothesis, that of *causality*: any feature at a coarse level of resolution is required to possess a (not necessarily unique) "cause" at a finer level of resolution although the reverse need not be true. This asymmetry leads to a rather strong constraint. The hypothesis in effect forbids the generation of "spurious resolution". Let me formalize the constraint first: Consider a surface $K(x, y, z) = \Lambda_0$ (a constant) in (x, y, z)-space [or "scale space" (Witkin, 1983)]. Then you can formulate the constraint for the stationary points of the derived images (the points $K_x = K_y = 0$). Note that you have extrema if the Hessian $K_{xx}K_{yy} - K_{xy}^2$ is positive (a minimum or dark blob if $K_{xx} + K_{yy}$ is positive, a light blob if it is negative) and a saddle if the Hessian is negative. If the primal image is generic (an assumption that is easily eliminated later on), then the Hessian never vanishes at the stationary points for $z = 0$. It may vanish at stationary points for certain finite values of z, however. Now the assumption of causality implies that the surface $K = \Lambda_0$ should point its convex side towards the direction of decreasing resolution at the extrema. For otherwise the more blurred image would possess luminance values that could not be traced to the less blurred images, contrary to the hypothesis.

The curvature of the surface $K(x, y, z) = \Lambda_0$ is easily obtained with standard methods (Spivak, 1975). First note that the unit surface normal $\mathbf{n}$ may be defined as:

$$\mathbf{n} = \mathbf{p}/p \quad \text{with} \quad \mathbf{p} = (K_x, K_y, K_z)\,.$$

The signs of the principal curvatures are defined with respect to this choice of orientation of the surface.

The principal curvatures are

$$k_i = \frac{\lambda_i}{\sqrt{K_x^2+K_y^2+K_z^2}} \qquad i=1,2,$$

where the λ_i are the roots of the (quadratic!) equation

$$\det \begin{vmatrix} K_{xx}-\lambda & K_{xy} & K_{xz} & K_x \\ K_{yx} & K_{yy}-\lambda & K_{yz} & K_y \\ K_{zx} & K_{zy} & K_{zz}-\lambda & K_z \\ K_x & K_y & K_z & 0 \end{vmatrix} = 0$$

or (because we consider the case $K_x=0$, $K_y=0$, $K_z=0$):

$$\lambda^2 - \lambda(K_{xx}+K_{yy}) + (K_{xx}K_{yy}-K_{xy}^2) = 0 .$$

By hypothesis $K_{xx}K_{yy}-K_{xy}^2$ is positive (I consider extrema, not saddle points), thus both roots have equal sign. This sign is given by the sign of $K_{xx}+K_{yy}=\Delta K$, whereas convexity (concavity) is defined relative to the sign of the third component of the surface normal. (That is the sign of K_z.) Thus the constraint can finally be written

$$\Delta K = \alpha^2(x,y,z)K_z ,$$

where α denotes an arbitrary but nowhere vanishing real function. [Note that this equation has really been derived at the location of the extrema solely. But then, for images that are not a priori known, these extrema might be anywhere! Thus the equation must hold at all points of the image, which is why I introduced the function $\alpha(x,y,z)$.] Consequently I have arrived at a partial differential equation that has to be satisfied by the family of derived images.

In order to proceed I introduce a second hypothesis at this point: *homogeneity and isotropy*. The inner scale depends only on the parameter z, and in no way on x or y. Thus I do not permit space variant blurring. Clearly this is not essential to the issue, but it simplifies the analysis greatly. The hypothesis means that $\alpha(x,y,z)$ depends only on z. Then I can introduce a new scale parameter t (say) in such a way that $t=\varphi(z)$ where φ is a monotonically increasing function, and $\Delta K = K_t$.

This is the well known heat conduction or diffusion equation. This equation governs the deep structure of the image.

Consequently, I *define* the family of derived images $K(x,y,t)$ as the solution of the heat conduction equation with the boundary condition $K(x,y,0) = L(x,y)$. This works fine if the image extends over the whole of R^2. If the primal image is only defined over a finite region S, say a square or disc, etc. (the usual case), I proceed a little different. First I define $L^*(x,y)$ as the solution of $\Delta L^* = 0$ with as boundary condition that $L^*(x,y) = L(x,y)$ restricted to ∂S (the boundary of the image). Then I define $K(x,y,t)$ as the solution of the heat conduction equation with as boundary condition $K(x,y,0) = L(x,y) - L^*(x,y)$ and $K(\partial S,t)=0$. [Note that L^* would lead to $K_t(x,y,0)=0$ anyway: it is an invariant component of the primal image.]

In retrospect you can obtain any derived image directly from the primal image through convolution with the gaussian kernel

$$K(\mathbf{r},\mathbf{r}') = \exp(-|\mathbf{r}-\mathbf{r}'|^2/4t)/4\pi t .$$

In fact any derived image at level t can be derived from any other derived image at level $t'<t$ through convolution with a suitable gaussian kernel (or point spread function). Thus if spurious resolution is prohibited (the first hypothesis), then the family of gaussians is unique (Note 1). Gaussian blurring is the only sensible way to embed a primal image into a one-parameter family.

Interestingly enough the structure proposed here has several features that can be traced to well known models of the visual system. For instance, the study of zero crossings for images subjected to different degrees of blurring (Marr et al., 1977) and the studies on processing in layered media (Marko, 1969; Roehler, 1976). The latter study even explicitely incorporates the diffusion equation.

3 Image Structure – The Superficial Structure

In the preceding paragraphs I have glibly spoken of light and dark spots in the image. Obviously such image features are of importance, but how do you delimit a light blob (say) in a blurred image? In one-dimensional images, such as time signals, one defines peaks and troughs either by way of extrema (Ehrich and Foith, 1976; e.g. a "peak" is a region between two successive minima) or through points of inflexion (Witkin, 1983). Both methods are not easily transposed to two dimensions. The two-dimensional equivalent of a point of inflexion would be a parabolic curve ($K_{xx}K_{yy}-K_{xy}^2=0$), but parabolic curves sometimes fail to enclose single extrema. People have attempted to use "zero-crossings ($\Delta L=0$)" (Marr et al., 1977) for the purpose, but these curves suffer from the same drawback. One nice method is the one commonly used in geography, and introduced in mathematics by Cayley (1859) and Maxwell (1870) ("Hills" and "Dales", separated by "watercourses" and "watersheds"). A quite similar method – that seems more natural for the present purpose – is to employ the foliation of the image induced by the family of equiluminance curves (Koenderink and van Doorn, 1979).

From differential topology it is known (Guillemin and Pollack, 1974) that almost all (in a precise sense)

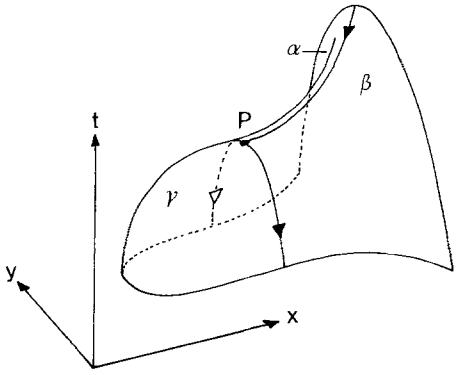

Fig. 1. The surface $K=\Lambda_0$. The point P is (x_0, y_0, t_0). In the regions alfa and beta the lines of steepest descent have the singular paths through P as asymptotes. In the region gamma these lines issue from P. Surfaces $K=\Lambda_1$ with $\Lambda_1>\Lambda_0$ have extrema nor saddle points, whereas surfaces $K=\Lambda_2$ with $\Lambda_2<\Lambda_0$ have one extremum and one saddle point: at point P you have a "collision" of a saddle and an extremum

images are generic, that is:
- stationary points ($K_x=K_y=0$) are isolated,
- $K_{xx}K_{yy}-K_{xy}^2=0$ at stationary points,
- stationary values are distinct.

Then singular equiluminance curves are points (at the extrema) and curves with self-intersections (at the saddle-points, Maxwell's "false extrema"). The extrema and false extrema can be put into a natural partial order (of inclusion) as follows: Each saddle point defines a closed equiluminance curve with a single self-intersection, the two loops define two disjunct families of closed equiluminance curves that contain either extrema or false extrema (containing other – possibly false – extrema, etc.). In this manner you obtain a nested family of (false) extrema and the inclusion defines a partial order. The boundary of the image does not lead to complications if you first subtract the invariant image (as noted earlier): then the boundary itself is a closed equiluminance curve. From the vantage point of visual perception this method of treating an image in terms of a hierarchy of nested and juxtaposed light and dark blotches appears as a very natural one.

4 The Deep Structure

When you blur an image, you loose structure: the total number of extrema cannot increase, and generally decreases if the blurring is sufficiently strong. A single process accounts for this (an immediate consequence of Thom's theorem (Thom, 1972): when t is increased it may happen that an extremum merges with a saddle-point, whereon both are annihilated. An example is (this is at the same time the general affine model of this singularity, see Fig. 1):

$$K(x_0+\delta x, y_0+\delta y, t_0+\delta t) = \Lambda_0 + \frac{\delta x^3}{6} + \frac{\delta y^2}{2} + \delta t(\delta x+1).$$

(Note that K satisfies the diffusion equation.)

For $\delta t<0$ the extremum is at $\delta y=0$, $\delta x=\sqrt{-2\delta t}$, the saddle at $\delta y=0$, $\delta x=-\sqrt{-2\delta t}$. For $t>0$ both have vanished. Note that $K(x, y, t_0)$ is not a generic image. In all practical cases the family of derived images is *versal*; that is all but a finite number of isolated derived images are generic.

The non-generic images occur as images in which an extremum merges with a saddle-point. Thus you can unequivocally assign extrema to saddle-points. The isoluminance curve through the saddle-point must encircle that extremum, and thus serves to define the boundary of the light or dark blob. There exists an even more natural method to do this, however.

The requirement that in two "successive" derived images, say $K(x, y, t)$ and $K(x, y, t+\delta t)$ (with x, y variable), corresponding points have equal luminance and are as close as possible, yields a simple rule of *projection* between images: the orbits of the projection are the integral curves of the vector field

$$\mathbf{s}=(-K_tK_x, -K_tK_y, K_x^2+K_y^2).$$

This is easily proved as follows: The point $\mathbf{r}+d\mathbf{r}$ at the image $t+dt$ that is connected to the point $\mathbf{r}$ at the image t, must satisfy $dL=\nabla K\cdot d\mathbf{r}+K_t dt=0$. Moreover, the steepest descent is in the direction of the gradient (∇K). Thus

$$d\mathbf{r}/dt=-(K_t/\nabla K\cdot\nabla K)\cdot\nabla K.$$

The vector $(\nabla K\cdot\nabla K)\dfrac{d\mathbf{r}}{dt}$ has everywhere the same direction as $d\mathbf{r}/dt$, and its singularities coincide with those of $d\mathbf{r}/dt$: thus the integral curves of these vector fields are the same.

The stationary points of the images are just the singularities of the vector field $\mathbf{s}$ (because $K_x^2+K_y^2=0$). When you project some region of a derivative image towards the primal image plane, it is apparent that not all points in the latter plane can be reached by the integral curves of $\mathbf{s}$: each extremum-saddle-point pair defines a region that remains blank. These regions are described through the integral curves that pass through the extremum and those through the saddle that do reach the plane $t=0$ (Fig. 2). These regions are topologically equivalent to discs, in the primal image plane the saddle-point lies on the border, the extremum inside it.

I propose to call these regions the "ranges" of the extrema, they can be taken to define the light and dark blobs defined by the extremum-saddle-point pairs.

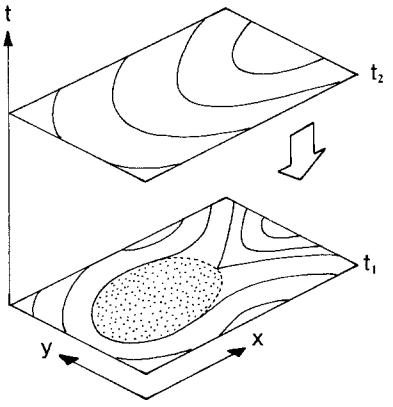

Fig. 2. If the derived image at $t=t_2$ is down-projected to the plane $t=t_1$, the dotted region is left open: it contains detail that is not present in the blurred image

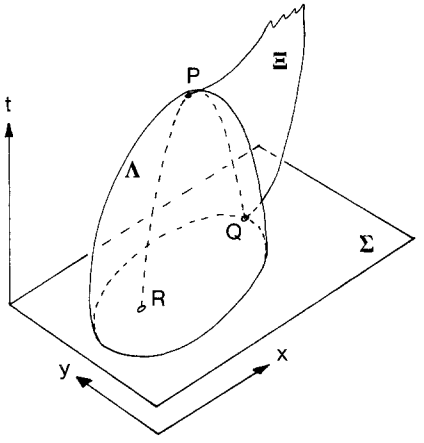

Fig. 3. A tube Λ defined by the saddle point Q, extremum R — pair in the primal image Σ. The top of the tube is the singular point P where saddle point and extremum meet. (It has a horizontal tangent plane.) The surface Ξ contains orbits that end on saddle points on the arc PQ and from there split into two branches on Λ. Any orbit inside the tube ends on an extremum on the arc PR. No orbit from outside Λ can enter its inside. Thus downprojection from a level above P leaves the realm on the primal image Σ uncovered

If you don't project down to the primal image plane, but to some intermediary image plane, you obtain the range at that level of resolution – at least if it exists there. These ranges sweep out tube-like volumes (with t as parameter, Fig. 3) in scale space. The tubes are closed on one side. (This highest point being the merge of the extremum-saddle-point pair.) These tubes define the volumes in scale space at which the blobs manifest themselves, I propose to call them the "realms" of the extrema. In complicated images many different realms coexist, both juxtaposed and nested to arbitrary depth. (Because of the structure of the **s** field the boundaries of different realms can never meet.) Thus you may really speak of light blobs containing other light or dark blobs, containing... etc.

For a certain finite range of resolution the blobs can be *identified* (that is if t is less than the value at which the extremum meets its saddle-point), and in a still more limited range the blob exists in its pure form, unarticulated. For too high a resolution the blob may be difficult to detect because it is articulated with irrelevant smaller detail (e.g. blurring really helps to find objects in scintigrams), whereas for too low a resolution the blobs loose identity (e.g. in a cardioscintigram the left and right ventricles may merge). Details thus have a limited range of resolution in which they can be said to exist. We can define this range from the top of the realm to the next lower top of any included subrealm.

Some details exist over a long range of resolution, others are more ephemeral and at once desintegrate once you identify them. There is some evidence that "stable features" (those that exist over long ranges) are the visually most conspicuous ones (Witkin, 1983).

Note that you cannot "reconstruct" the primal image from a highly blurred image through the device of downprojection: surely this sharpens or "deblurs" the image, but at the cost of the introduction of blank spaces (the ranges of extrema on the primal image). Thus you have to bring in extra information at the levels of resolution where – by downprojection – new realms appear. A complete description of the image on the coarsest possible scale entails:

1) the image at some (coarse) level of resolution,

2) the luminance values on the loci of extrema (a set of curves in (x, y, t) space). Downprojection from these entities completely fills the primal image plane, thus if you add,

3) the geometrical structure of the family of downprojecting paths, you have completely characterized the image.

Concerning the geometrical structure of the downprojecting paths, they alone are sufficient description! For the structure of the **s**-field determines the surfaces $K=\text{const}$: $\mathbf{s}\wedge(\mathbf{s}-(\mathbf{s}\cdot\mathbf{e}_t)\mathbf{e}_t)$ yields the direction of the normal to these surfaces. ($\mathbf{e}_t$ a unit vector in the t-direction.) Consequently, the image is determined except for a transformation of the type $K'(x,y,t)=\Phi(K(x,y,t))$. Obviously this transformation must conserve the property that $\Delta K=K_t$, thus $\Delta K'=K'_t$. This latter equation can be shown to be equivalent to:

$$\frac{\partial^2\Phi}{\partial K^2}|\nabla K|^2+\frac{\partial\Phi}{\partial K}(\Delta K-K_t)=0\,.$$

Thus $\dfrac{\partial^2\Phi}{\partial K^2}=0$, or $K'=\alpha K+\beta$ (α and β const).

But then the image is determined up to a multiplicative and an additive constant through the

projection orbits in scale space alone! In fact you obtain $\nabla \ln|\nabla K|$ through the **s**-field, and thus by integration K except for scale and offset.

One nice feature of this description is that it permits a *logical filtering in the scale domain.* For every range in the primal image plane you may solve $\Delta L = 0$ within the range with the boundary value $L = L$ on the boundary of the range. Then you may "lift off" the detail by defining it as $L - L$ within the range and zero outside. In this way the whole primal image can be written as a superposition of the light and dark blobs. A subfamily may be defined for each subimage, and because the diffusion equation is linear the original family is just the superposition of the subfamilies. Now you may choose, for instance, to use only summands belonging to features existing in a certain range of scales. This is in effect a logical filtering in the scale domain. You may even compose images in which details in different scale ranges have been blurred differentially, etc.

Finally, note that the diffusion equation may also be used backwards to *enhance* the image. This process may end, however. E.g. the primal image $\exp(-(x^2+y^2)/4\mu)/(4\pi\mu)$ can only be sharpened to $t = -\mu$, then it has been shrunk to an impulse.

5 The Sampling of Images in Scale Space

Two basic solutions of the heat equation are

$$\varphi(\mathbf{r}, t) = \exp(-|\mathbf{r}|^2/4t)/4\pi t \qquad (\int \varphi d\mathbf{r} = 1, \varphi(\mathbf{r}, 0) = \delta(\mathbf{r}))$$

$$\psi(\mathbf{r}, t) = \mathrm{Re}\exp(-i\mathbf{k}\cdot\mathbf{r} - k^2 t).$$

Both are convenient when you want to construct solutions of the heat conduction equation through the principle of superposition. I use these simple solutions here to demonstrate some principles that pertain to the *sampling* of the image in scale space. This is of obvious importance to practical (i.e. numerical) applications.

Let the metrical resolution be given, e.g. the luminance (or rather the flux in a resolution cell) is measured with a relative accuracy of $\exp(-R)$ ($R \gg 1$, thus the accuracy is $R/\ln 2$ "bits"). Take $\varphi(\mathbf{r}, t)$ as a basic solution, then if you require that at any level of resolution a cell centered at the origin samples at least $(1-\exp(-R))th$ part of the total flux, such a cell must have a radius of $2\sqrt{tR}$. At a center spacing of $\sqrt{tR}$ such cells sample uncorrelated fluxes if the points in the ground plane were uncorrelated. You may also inquire after the required Nyquist sample frequency. Consider the basic solution $\psi(\mathbf{r}, t)$: a spatial frequency component with wavelength $\lambda = 2\pi/k$ damps exponentially with characteristic decay length

$$\Delta t_{1/e} = 1/k^2 = \lambda^2/4\pi^2.$$

If you start with a Gaussian spectrum $\Psi(\mathbf{k}, t_0) = \Psi_0 \exp(-k^2/2k_0^2)$, you have that

$$\Psi(\mathbf{k}, t) = \Psi_0 \exp[(-k^2/2)\cdot(2(t-t_0) + 1/k_0^2)].$$

Thus the spectrum remains Gaussian but the width decreases as

$$(1/k_0^2 + 2(t-t_0))^{-1/2}.$$

If you start out with a white spectrum ($k_0 \to \infty$), the width just goes as $1/\sqrt{2t}$. (I will set $t_0 = 0$ in the sequel.) The highest significant frequency is obviously $k_{\max}(t)$, for which $\Psi(k_{\max}, t) = \exp(-R)\psi(0, t)$. Thus $k_{\max} = \sqrt{R/t}$, or in other words the Nyquist sample density must use a spacing $d = \pi/k_{\max} = \pi\sqrt{t/R}$.

Another problem concerns the spacing that is required along the t-axis. The characteristic decay length for the highest frequency component is d^2/π^2 with d as defined above. Thus this wave damps with a factor

$$(1 - \pi^2\delta t/d^2) \text{ over a distance } \delta t (\delta t \ll d^2/\pi^2).$$

Now there are two problems to consider: that of the *accuracy* of the representation and that of the *stability* (in the numerical sense) of the representation. Let us consider accuracy first. The approximation $I(\mathbf{r}, \delta t) \approx I(\mathbf{r}, 0) + \delta t \cdot \Delta I(\mathbf{r}, 0)$ can easily be shown to have a relative error bounded by

$$\varepsilon \leqq \frac{d^4 \delta t^2}{2\pi^4}.$$

The requirement that $\varepsilon < \exp(-R)$ then yields the condition

$$\delta t < \frac{\sqrt{2} e^{-R/2}}{\pi^2} d^2.$$

Next consider stability. For a spatial frequency ω the transfer function from layer $t = 0$ to layer $t = \delta t$ is $(1 - \omega^2 \delta t)$. Stability requires that the absolute value of the transfer function remains less than unity: otherwise arbitrarily small errors will soon grow without bounds. This leads to the requirement $\delta t < \frac{2}{\pi^2} d^2$. For any reasonable value of R stability is guaranteed when

$$\delta t = (\sqrt{2}/\pi^2) \exp(-R/2) \cdot d^2.$$

This can also be written (making use of the relation $d = \pi\sqrt{t/R}$) as

$$\delta t/t = \sqrt{2} \exp(-R/2)/R = \text{const}.$$

Thus you need a *logarithmic* spacing of sample planes along the t-axis. This is in accord with the intuitive notion that there can be *no preferred scale*, thus a uniform sampling density on a logarithmically scaled axis is indicated.

From these basic results it appears that the reciprocal of t (say $q=t^{-1}$) is an even more natural measure of resolution from the standpoint of structural information theory: in $(\mathbf{r}, q)$-space the resolution cells have constant volume. This volume is

$$\xi = \Delta q \Delta \mathbf{r} = \Delta t \cdot d^2/t^2 = \sqrt{2}\pi^2 \exp(-R/2)/R^2;$$

it depends on the *metrical* resolution alone. This is a basic "uncertainty relation" for scale space: a "blob" of area ΔA exists only over a resolution interval $\Delta q = \xi/\Delta A$.

The so-called hierarchical "pyramid" structures that are in widespread use today for multiresolution image analysis are all much coarser than this (Burt et al., 1981). Consequently the family of derived images cannot be derived simply from these structures, except by the trivial measure of starting all over from the primal image. Thus quantization effects must be rather severe. Yet algorithms based on these structures are admittedly powerful and these structures behave at least qualitatively very much like the system discussed in this paper. Note that the correctly sampled image is also a "pyramid", but one that tapers much less swiftly. The total number of samples needed to represent the structure can be easily obtained as follows: For a square image with sides L and resolution δ the total resolution space has a volume $L^2 q_{\max} = \pi^2 L^2/R^2\delta^2 = \pi^2 N/R^2$, where N is the total number of independent image elements in the original images. Dividing the volume by the volume of a resolution cell, we find the number of samples (M) needed: $M = N \exp(R/2)/\sqrt{2}$. This number is seen to grow exponentially with the required metrical accuracy (R). For small values of R, however, M is of the order of N, e.g. for an accuracy of 1% you have $M \approx 7N$. Thus the human visual system contains certainly sufficient hardware as far as mere numbers are concerned to accommodate the retinal image in this manner!

6 Conclusions

One main result is that there appears to be essentially a single sensible way to embed an image into a one-parameter family of derived images, with resolution as a parameter: namely by a diffusion process, or convolution with a family of Gaussian point spread functions. This result must have seemed obvious to some previous investigators in this field who started out from the family of Gaussians (or rather "DOG's": "difference of Gaussians") or from iterated blurrings (which asymptotically leads to diffusion) in an apparently ad hoc fashion (Marr and Hildreth, 1980). The relation to the diffusion equation appears to have been overlooked previously, although it is this equation that explicitely defines the deep structure of the image.

Another main result is that if the mutual immutability of details with respect to blurring is taken into consideration, then you are able to define a true (or "linear") order of extrema: the image can be described unambiguously as a set of nested and juxtaposed light and dark blobs that vanish in a well defined sequence on progressive blurring. Note that such a linear order cannot be established at just one single level of resolution: e.g. for a pseudo-maximum consisting of two maxima (that is a light blob containing two smaller light blobs) it cannot be decided which of the sub-maxima is actually subordinate to the other, whereas on blurring this becomes clear: at some degree of blurring one of the two must vanish and yield to the other. Thus the image can be truely segmented into nested and juxtaposed light and dark blobs. Moreover, to each blob can be assigned three characteristic ranges of resolution: in one of them the blob is non-existent (or unresolved), in another it manifests itself purely as a simple blob, and in a final one finer detail intrudes on its territory. Thus the effects of progressive erosion clarify the deep structure of the image. In typical image processing applications this structure can be used for "logical filtering with respect to scale". Such a filtering can for instance be based on the relative stability of the blobs with respect to erosion.

In the family of derived images as described in this paper the structural information is not coded very efficiently: the primal image – thus the values of the luminances in just one plane of scale space – contains already all information! This may be remedied by considering K_t, the derivative with respect to scale, instead of K. This function is equivalent to ΔK, the Laplacian of the derived images, and thus it is just Marr and Hildreth's (1980) scheme (although these authors arrived at their method in a different, rather ad hoc, manner). Obviously, K_t contains the same information as K, except for a possible difference that is invariant against blurring. If you consider a primal image with detail in a very limited range of scale, e.g. the function $\psi(\mathbf{r}, t)$, you find that

$$K_t = -k^2 \exp(-k^2 t) \cos \mathbf{k} \cdot \mathbf{r}.$$

Thus at a given level (fixed t) you find the spatial spectrum of the primal image filtered with a bandpass filter with a relative (half-height) bandwidth of 0.5. The curves $K_t = 0$ are the "zero-crossings" of Marr and Hildreth (1980). In our scheme they are singled out through the property that they are (locally) stable with respect to erosion.

As I have shown before (Koenderink et al., 1978, 1982) the visual system is extensive enough to be able to represent the retinal image *at all levels of resolution simultaneously*. The initialization of this data structure is simple diffusion which can be effected in an extremely simple manner by layered neural structures (Roehler, 1976; Marko, 1969). In this paper I have shown how to use such a structure, that is how to "read it out". This requires projections between different layers of the structure, guided by the activity in the network itself.

Note

The theorem that Gaussian blurring uniquely avoids spurious resolution, but only for the case of one-dimensional images, was brought to my attention by Andrew Witkin at the Marr Conference held at Cold Spring Harbor, 1983. Apparently the proof was complicated and yielded no intuitive insight. I immediately realized that an existing proof by myself – that diffusion only destroys structure but cannot generate it – could easily be adjusted to proof the theorem in a very simple manner in the more general 2-dimensional case. (The one-dimensional case follows directly from the 2-dimensional case.)

References

Burt, P.J., Hong, Tsai-Hong, Rosenfeld, A.: Segmentation and estimation of image region properties through cooperative hierarchical computation. IEEE Trans. SMC-**11**, 802–825 (1981)

Cayley, A.: On contour and slope lines. The London, Edinburgh, and Dublin Philosophical Magazine and J. of Science 18 **(120)**, 264–268 (Oct. 1859)

Maxwell, J.C.: On hills and dales. The London, Edinburgh, and Dublin Philosophical Magazine and J. of Science 4th Series 40 **(269)**, 421–425 (Dec. 1870)

Ehrich, R.W., Foith, J.P.: Representation of random waveforms by relational trees. IEEE Trans. Comput. **25**, 725–736 (1976)

Guillemin, V., Pollack, A.: Differential topology. Englewood Cliffs, NJ: Prentice-Hall 1974

Hay, G.A., Chesters, M.S.: A model of visual threshold detection. J. Theor. Biol. **67**, 221–240 (1977)

Koenderink, J.J., Doorn, A.J. van: The structure of two-dimensional scalar fields with applications to vision. Biol. Cybern. **33**, 151–158 (1979)

Koenderink, J.J., Doorn, A.J. van: Visual detection of spatial contrast: influence of location in the visual field, target extent and illuminance level. Biol. Cybern. **30**, 157–167 (1978)

Koenderink, J.J., Doorn, A.J. van: Invariant features of contrast detection: an explanation in terms of self-similar detector arrays. J. Opt. Soc. Am. **72**, 83–87 (1982)

Marko, H.: Die Systemtheorie homogener Schichten. Kybernetik **5**, 221 (1969)

Marr, D., Poggio, T., Ullman, S.: Bandpass channels, zero-crossings, and early visual information processing. J. Opt. Soc. Am. **69**, 914–916 (1977)

Marr, D., Hildreth, E.: Theory of edge detection. Proc. Royal Soc. Lond. B **207**, 187–217 (1980)

Roehler, R.: Ein Modell zur örtlich-zeitlichen Signalübertragung im visuellen System des Menschen auf der Basis der linearen Systemtheorie kontinuierlichen Medien. Biol. Cybern. **22**, 97–105 (1976)

Spivak, M.: A comprehensive introduction to differential geometry, Vol. III. Berkeley, CA: Publish or Perish Inc. 1975

Thom, R.: Stabilité structurelle et morphogenésè. Reading, MA: Benjamin 1972

Witkin, A.P.: Scale-space filtering. Proc. of IJCAI, 1019–1021, Karlsruhe 1983

Received: April 20, 1984

Prof. Dr. J. J. Koenderink
Rijksuniversiteit Utrecht
Fysisch Laboratorium
Princetonplein 5
Postbus 80.000
3508 TA Utrecht
The Netherlands