Scale-Space and Edge Detection Using Anisotropic Diffusion

PIETRO PERONA AND JITENDRA MALIK

Abstract—**The scale-space technique introduced by Witkin involves generating coarser resolution images by convolving the original image with a Gaussian kernel. This approach has a major drawback: it is difficult to obtain accurately the locations of the "semantically meaningful" edges at coarse scales. In this paper we suggest a new definition of scale-space, and introduce a class of algorithms that realize it using a diffusion process. The diffusion coefficient is chosen to vary spatially in such a way as to encourage intraregion smoothing in preference to interregion smoothing. It is shown that the "no new maxima should be generated at coarse scales" property of conventional scale space is preserved. As the region boundaries in our approach remain sharp, we obtain a high quality edge detector which successfully exploits global information. Experimental results are shown on a number of images. The algorithm involves elementary, local operations replicated over the image making parallel hardware implementations feasible.**

Index Terms—**Adaptive filtering, analog VLSI, edge detection, edge enhancement, nonlinear diffusion, nonlinear filtering, parallel algorithm, scale-space.**

I. Introduction

THE importance of multiscale descriptions of images has been recognized from the early days of computer vision, e.g., Rosenfeld and Thurston [20]. A clean formalism for this problem is the idea of scale-space filtering introduced by Witkin [21] and further developed in Koenderink [11], Babaud, Duda, and Witkin [1], Yuille and Poggio [22], and Hummel [7], [8].

The essential idea of this approach is quite simple: embed the original image in a family of derived images $I(x, y, t)$ obtained by convolving the original image $I_0(x, y)$ with a Gaussian kernel $G(x, y; t)$ of variance t:

$$I(x, y, t) = I_0(x, y) * G(x, y; t). \tag{1}$$

Larger values of t, the scale-space parameter, correspond to images at coarser resolutions. See Fig. 1.

As pointed out by Koenderink [11] and Hummel [7], this one parameter family of derived images may equivalently be viewed as the solution of the heat conduction, or diffusion, equation

$$I_t = \Delta I = (I_{xx} + I_{yy}) \tag{2}$$

Manuscript received May 15, 1989; revised February 12, 1990. Recommended for acceptance by R. J. Woodham. This work was supported by the Semiconductor Research Corporation under Grant 82-11-008 to P. Perona, by an IBM faculty development award and a National Science Foundation PYI award to J. Malik, and by the Defense Advanced Research Projects Agency under Contract N00039-88-C-0292.

The authors are with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720.

IEEE Log Number 9036110.

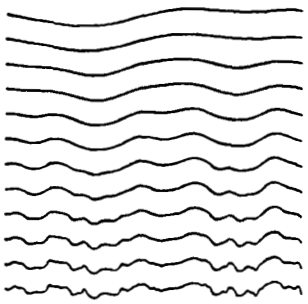


Fig. 1. A family of 1-D signals $I(x, t)$ obtained by convolving the original one (bottom) with Gaussian kernels whose variance increases from bottom to top (adapted from Witkin [21]).

with the initial condition $I(x, y, 0) = I_0(x, y)$, the original image.

Koenderink motivates the diffusion equation formulation by stating two criteria.

1) 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. In other words, no spurious detail should be generated when the resolution is diminished.

2) Homogeneity and Isotropy: The blurring is required to be space invariant.

These criteria lead naturally to the diffusion equation formulation. It may be noted that the second criterion is only stated for the sake of simplicity. We will have more to say on this later. In fact the major theme of this paper is to replace this criterion by something more useful.

It should also be noted that the causality criterion does not force uniquely the choice of a Gaussian to do the blurring, though it is perhaps the simplest. Hummel [7] has made the important observation that a version of the maximum principle from the theory of parabolic differential equations is equivalent to causality. We will discuss this further in Section IV-A.

This paper is organized as follows: Section II critiques the standard scale space paradigm and presents an additional set of criteria for obtaining "semantically meaningful" multiple scale descriptions. In Section III we show that by allowing the diffusion coefficient to vary, one can satisfy these criteria. In Section IV-A the maximum principle is reviewed and used to show how the causality criterion is still satisfied by our scheme. In Section V some

experimental results are presented. In Section VI we compare our scheme with other edge detection schemes. Section VII presents some concluding remarks.

II. Weaknesses of the Standard Scale-Space Paradigm

We now examine the adequacy of the standard scale-space paradigm for vision tasks which need "semantically meaningful" multiple scale descriptions. Surfaces in nature usually have a hierarchical organization composed of a small discrete number of levels [13]. At the finest level, a tree is composed of leaves with an intricate structure of veins. At the next level, each leaf is replaced by a single region, and at the highest level there is a single blob corresponding to the treetop. There is a natural range of resolutions (intervals of the scale-space parameter) corresponding to each of these levels of description. Furthermore at each level of description, the regions (leaves, treetops, or forests) have well-defined boundaries.

In the standard scale-space paradigm the true location of a boundary at a coarse scale is not directly available in the coarse scale image. This can be seen clearly in the 1-D example in Fig. 2. The locations of the edges at the coarse scales are shifted from their true locations. In 2-D images there is the additional problem that edge junctions, which contain much of the spatial information of the edge drawing, are destroyed. The only way to obtain the true location of the edges that have been detected at a coarse scale is by tracking across the scale space to their position in the original image. This technique proves to be complicated and expensive [5].

The reason for this spatial distortion is quite obvious—Gaussian blurring does not "respect" the natural boundaries of objects. Suppose we have the picture of a treetop with the sky as background. The Gaussian blurring process would result in the green of the leaves getting "mixed" with the blue of the sky, long before the treetop emerges as a feature (after the leaves have been blurred together). Fig. 3 shows a sequence of coarser images obtained by Gaussian blurring which illustrates this phenomenon. It may also be noted that the region boundaries are generally quite diffuse instead of being sharp.

With this as motivation, we enunciate [18] the criteria which we believe any candidate paradigm for generating multiscale "semantically meaningful" descriptions of images must satisfy.

1) Causality: As pointed out by Witkin and Koenderink, a scale-space representation should have the property that no spurious detail should be generated passing from finer to coarser scales.

2) Immediate Localization: At each resolution, the region boundaries should be sharp and coincide with the semantically meaningful boundaries at that resolution.

3) Piecewise Smoothing: At all scales, intraregion smoothing should occur preferentially over interregion smoothing. In the tree example mentioned earlier, the leaf regions should be collapsed to a treetop *before* being merged with the sky background.

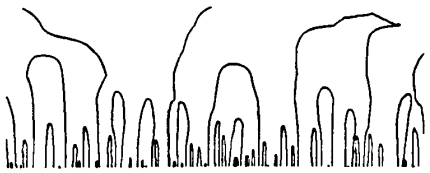


Fig. 2. Position of the edges (zeros of the Laplacian with respect to x) through the linear scale space of Fig. 1 (adapted from Witkin [21]).

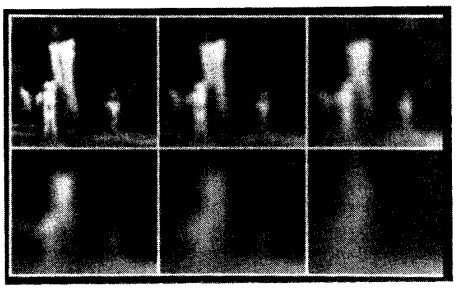


Fig. 3. Scale-space (scale parameter increasing from top to bottom, and from left to right) produced by isotropic linear diffusion (0, 2, 4, 8, 16, 32 iterations of a discrete 8 nearest-neighbor implementation. Compare to Fig. 12.

III. Anisotropic Diffusion

There is a simple way of modifying the linear scale-space paradigm to achieve the objectives that we have put forth in the previous section. In the diffusion equation framework of looking at scale-space, the diffusion coefficient c is assumed to be a constant independent of the space location. There is no fundamental reason why this must be so. To quote Koenderink [11, p. 364], ". . . I do not permit space variant blurring. Clearly this is not essential to the issue, but it simplifies the analysis greatly." We will show how a suitable choice of $c(x, y, t)$ will enable us to satisfy the second and third criteria listed in the previous section. Furthermore this can be done without sacrificing the causality criterion.

Consider the anisotropic diffusion equation

$$I_t = div\left(c(x, y, t)\nabla I\right) = c(x, y, t)\Delta I + \nabla c \cdot \nabla I \quad (3)$$

where we indicate with *div* the divergence operator, and with ∇ and Δ respectively the gradient and Laplacian operators, with respect to the space variables. It reduces to the isotropic heat diffusion equation $I_t = c\Delta I$ if $c(x, y, t)$ is a constant. Suppose that at the time (scale) t, we knew the locations of the region boundaries appropriate for that scale. We would want to encourage smoothing *within* a region in preference to smoothing *across* the boundaries. This could be achieved by setting the conduction coefficient to be 1 in the interior of each region and 0 at the boundaries. The blurring would then take place separately in each region with no interaction between regions. The region boundaries would remain sharp.

Of course, we do *not* know in advance the region boundaries at each scale (if we did the problem would already have been solved!). What can be computed is a

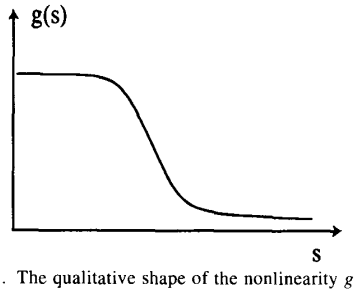


Fig. 4. The qualitative shape of the nonlinearity $g(\cdot)$.

current best estimate of the location of the boundaries (edges) appropriate to that scale.

Let $\boldsymbol{E}(x, y, t)$ be such an estimate: a vector-valued function defined on the image which ideally should have the following properties:

1) $\boldsymbol{E}(x, y, t) = \boldsymbol{0}$ in the interior of each region.

2) $\boldsymbol{E}(x, y, t) = K\boldsymbol{e}(x, y, t)$ at each edge point, where $\boldsymbol{e}$ is a unit vector normal to the edge at the point, and K is the local contrast (difference in the image intensities on the left and right) of the edge.

Note that the word *edge* as used above has not been formally defined—we mean here the perceptual subjective notion of an edge as a region boundary. A completely satisfactory formal definition is likely to be part of the solution, rather than the problem definition!

If an estimate $\boldsymbol{E}(x, y, t)$ is available, the conduction coefficient $c(x, y, t)$ can be chosen to be a function $c = g(\|\boldsymbol{E}\|)$ of the magnitude of $\boldsymbol{E}$. According to the previously stated strategy $g(\cdot)$ has to be a nonnegative monotonically decreasing function with $g(0) = 1$ (see Fig. 4). This way the diffusion process will mainly take place in the interior of regions, and it will not affect the region boundaries where the magnitude of $\boldsymbol{E}$ is large.

It is intuitive that the success of the diffusion process in satisfying the three scale-space goals of Section II will greatly depend on how accurate the estimate $\boldsymbol{E}$ is as a "guess" of the position of the edges. Accuracy though is computationally expensive and requires complicated algorithms. We are able to show that fortunately the simplest estimate of the edge positions, the gradient of the brightness function, i.e., $\boldsymbol{E}(x, y, t) = \nabla I(x, y, t)$, gives excellent results.

There are many possible choices for $g(\cdot)$, the most obvious being a binary valued function. In the next section we show that in case we use the edge estimate $\boldsymbol{E}(x, y, t) = \nabla I(x, y, t)$ the choice of $g(\cdot)$ is restricted to a subclass of the monotonically decreasing functions.

IV. Properties of Anisotropic Diffusion

We first establish that anisotropic diffusion satisfies the causality criterion of Section II by recalling a general result of the partial differential equation theory, the maximum principle. In Section IV-B we show that a diffusion in which the conduction coefficient is chosen locally as a function of the magnitude of the gradient of the brightness function, i.e.,

$$c(x, y, t) = g\left(\left\|\nabla I(x, y, t)\right\|\right) \tag{4}$$

will not only preserve, but also sharpen, the brightness edges if the function $g(\cdot)$ is chosen properly.

A. The Maximum Principle

The causality criterion requires that no new features are introduced in the image in passing from fine to coarse scales in the scale-space. If we identify "features" in the images with "blobs" of the brightness function $I(x, y, t)$ for different values of the scale parameter t, then the birth of a new "blob" would imply the creation of either a maximum or a minimum which would have to belong either to the interior or the top face $I(x, y, t_f)$ of the scale space (t_f is the coarsest scale of the scale-space). Therefore the causality criterion can be established by showing that all maxima and minima in the scale-space belong to the original image.

The diffusion equation (3) is a special case of a more general class of elliptic equations that satisfy a maximum principle. The principle states that all the maxima of the solution of the equation in space and time belong to the initial condition (the original image), and to the boundaries of the domain of interest provided that the conduction coefficient is positive. In our case, since we use adiabatic boundary conditions, the maximum principle is even stronger: the maxima only belong to the original image. A proof of the principle may be found in [17]; to make the paper self-contained we provide an easy proof in the Appendix, where the adiabatic boundary case is also treated, and weaker hypotheses on the conduction coefficient are used. A discrete version of the maximum principle is proposed in Section V.

B. Edge Enhancement

With conventional low-pass filtering and linear diffusion the price paid for eliminating the noise, and for performing scale space, is the blurring of edges. This causes their detection and localization to be difficult. An analysis of this problem is presented in [4].

Edge enhancement and reconstruction of blurry images can be achieved by high-pass filtering or running the diffusion equation backwards in time. This is an ill-posed problem, and gives rise to numerically unstable computational methods, unless the problem is appropriately constrained or reformulated [9].

We will show here that if the conduction coefficient is chosen to be an appropriate function of the image gradient we can make the anisotropic diffusion enhance edges while running *forward* in time, thus enjoying the stability of diffusions which is guaranteed by the maximum principle.

We model an edge as a step function convolved with a Gaussian. Without loss of generality, assume that the edge is aligned with the y axis.

The expression for the divergence operator simplifies to

$$\operatorname{div}\left(c(x, y, t)\nabla I\right) = \frac{\partial}{\partial x}\left(c(x, y, t)I_x\right).$$

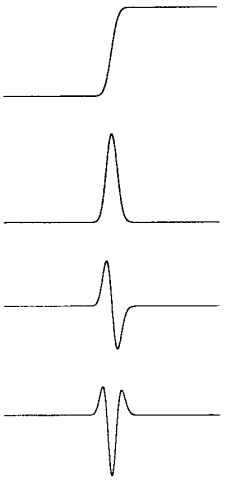


Fig. 5. (Top to bottom) A mollified step edge and its 1st, 2nd, and 3rd derivatives.

We choose c to be a function of the gradient of I: $c(x, y, t) = g(I_x(x, y, t))$ as in (4). Let $\phi(I_x) \doteq g(I_x) \cdot I_x$ denote the flux $c \cdot I_x$.

Then the 1-D version of the diffusion equation (3) becomes

$$I_t = \frac{\partial}{\partial x}\phi(I_x) = \phi'(I_x) \cdot I_{xx}. \tag{5}$$

We are interested in looking at the variation in time of the slope of the edge: $\partial/\partial t(I_x)$. If $c(\cdot) > 0$ the function $I(\cdot)$ is smooth, and the order of differentiation may be inverted:

$$\frac{\partial}{\partial t}(I_x) = \frac{\partial}{\partial x}(I_t) = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial x}\phi(I_x)\right)$$
$$= \phi'' \cdot I_{xx}^2 + \phi' \cdot I_{xxx}. \tag{6}$$

Suppose the edge is oriented in such a way that $I_x > 0$. At the point of inflection $I_{xx} = 0$, and $I_{xxx} << 0$ since the point of inflection corresponds to the point with maximum slope (see Fig. 5). Then in a neighborhood of the point of inflection $\partial/\partial t(I_x)$ has sign opposite to $\phi'(I_x)$. If $\phi'(I_x) > 0$ the slope of the edge will decrease with time; if, on the contrary $\phi'(I_x) < 0$ the slope will increase with time.

Notice that this increase in slope cannot be caused by a scaling of the edge, because this would violate the maximum principle. The edge becomes sharper.

There are several possible choices of $\phi(\cdot)$, for example, $g(I_x) = C/(1 + (I_x/K)^{1+\alpha})$ with $\alpha > 0$ (see Fig. 6). Then there exists a certain threshold value related to K, and α, below which $\phi(\cdot)$ is monotonically increasing, and beyond which $\phi(\cdot)$ is monotonically decreasing, giving the desirable result of blurring small discontinuities *and* sharpening edges. Notice also that in a neighborhood of the steepest region of an edge the diffusion may be thought of as running "backwards" since $\phi'(I_x)$ in (5) is

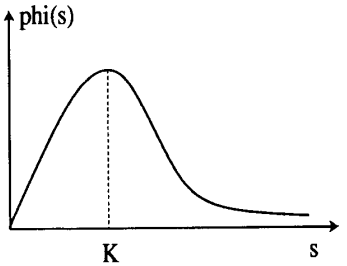


Fig. 6. A choice of the function $\phi(\cdot)$ that leads to edge enhancement.

negative. This may be a source of concern since it is known that constant-coefficient diffusions running backwards are unstable and amplify noise generating ripples. In our case this concern is unwarranted: the maximum principle guarantees that ripples are not produced. Experimentally one observes that the areas where $\phi'(I_x) < 0$ quickly shrink, and the process keeps stable.

V. Experimental Results

Our anisotropic diffusion, scale-space, and edge detection ideas were tested using a simple numerical scheme that is described in this section.

Equation (3) can be discretized on a square lattice, with brightness values associated to the vertices, and conduction coefficients to the arcs (see Fig. 7). A 4-nearest-neighbors discretization of the Laplacian operator can be used:

$$I_{i,j}^{t+1} = I_{i,j}^t + \lambda[c_N \cdot \nabla_N I + c_S \cdot \nabla_S I + c_E \cdot \nabla_E I + c_W \cdot \nabla_W I]_{i,j}^t \tag{7}$$

where $0 \le \lambda \le 1/4$ for the numerical scheme to be stable, N, S, E, W are the mnemonic subscripts for North, South, East, West, the superscript and subscripts on the square bracket are applied to all the terms it encloses, and the symbol ∇ (not to be confused with ∇, which we use for the gradient operator) indicates nearest-neighbor differences:

$$\nabla_N I_{i,j} \equiv I_{i-1,j} - I_{i,j}$$
$$\nabla_S I_{i,j} \equiv I_{i+1,j} - I_{i,j}$$
$$\nabla_E I_{i,j} \equiv I_{i,j+1} - I_{i,j}$$
$$\nabla_W I_{i,j} \equiv I_{i,j-1} - I_{i,j}. \tag{8}$$

The conduction coefficients are updated at every iteration as a function of the brightness gradient (4):

$$c_{N_{i,j}}^t = g\left(\left\|(\nabla I)_{i+(1/2),j}^t\right\|\right)$$
$$c_{S_{i,j}}^t = g\left(\left\|(\nabla I)_{i-(1/2),j}^t\right\|\right)$$
$$c_{E_{i,j}}^t = g\left(\left\|(\nabla I)_{i,j+(1/2)}^t\right\|\right)$$
$$c_{W_{i,j}}^t = g\left(\left\|(\nabla I)_{i,j-(1/2)}^t\right\|\right). \tag{9}$$

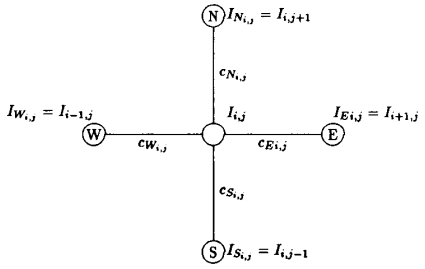


Fig. 7. The structure of the discrete computational scheme for simulating the diffusion equation (see Fig. 8 for a physical implementation). The brightness values $I_{i,j}$ are associated with the nodes of a lattice, the conduction coefficients c to the arcs. One node of the lattice and its four North, East, West, and South neighbors are shown.

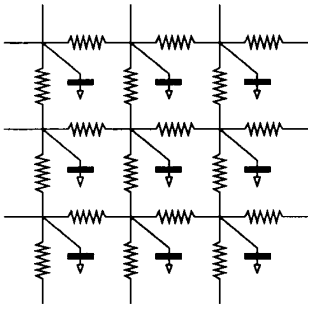


Fig. 8. The structure of a network realizing the implementation of anisotropic diffusion described in Section V, and more in detail in [19]. The charge on the capacitor at each node of the network represents the brightness of the image at a pixel. Linear resistors produce isotropic linear diffusion. Resistors with an I-V characteristic as in Fig. 6 produce anisotropic diffusion.

The value of the gradient can be computed on different neighborhood structures achieving different compromises between accuracy and locality. The simplest choice consists in approximating the norm of the gradient at each arc location with the absolute value of its projection along the direction of the arc:

$$\begin{aligned} c^t_{N_{i,j}} &= g(|\nabla_N I^t_{i,j}|) \\ c^t_{S_{i,j}} &= g(|\nabla_S I^t_{i,j}|) \\ c^t_{E_{i,j}} &= g(|\nabla_E I^t_{i,j}|) \\ c^t_{W_{i,j}} &= g(|\nabla_W I^t_{i,j}|). \end{aligned} \tag{10}$$

This scheme is not the exact discretization of (3), but of similar diffusion equation in which the conduction tensor is diagonal with entries $g(|I_x|)$ and $g(|I_y|)$ instead of $g(\|\nabla I\|)$ and $g(\|\nabla I\|)$. This discretization scheme preserves the property of the continuous equation (3) that the total amount of brightness in the image is preserved. Additionally the "flux" of brightness through each arc of the lattice only depends on the values of the brightness at the two nodes defining it, which makes the scheme a natural choice for analog VLSI implementations [19]. See Fig. 8. Less crude approximations of the gradient yielded perceptually similar results at the price of increased computational complexity.

It is possible to verify that, whatever the choice of the approximation of the gradient, the discretized scheme still satisfies the maximum (and minimum) principle provided that the function g is bounded between 0 and 1.

We can in fact show this directly from (7), using the facts $\lambda \in [0, 1/4]$, and $c \in [0, 1]$, and defining $(I_M)^t_{i,j} \doteq \max\{(I, I_N, I_S, I_E, I_W)^t_{i,j}\}$, and $(I_m)^t_{i,j} \doteq \min\{(I, I_N, I_S, I_E, I_W)^t_{i,j}\}$, the maximum and minimum of the neighbors of $I_{i,j}$ at iteration t. We can prove that

$$(I_m)^t_{i,j} \le I^{t+1}_{i,j} \le (I_M)^t_{i,j} \tag{11}$$

i.e., no (local) maxima and minima are possible in the interior of the discretized scale-space. In fact:

$$\begin{aligned} I^{t+1}_{i,j} &= I^t_{i,j} + \lambda[c_N \cdot \nabla_N I + c_S \cdot \nabla_S I \\ &\quad + c_E \cdot \nabla_E I + c_W \cdot \nabla_W I]^t_{i,j} \\ &= I^t_{i,j}(1 - \lambda(c_N + c_S + c_E + c_W)^t_{i,j}) \\ &\quad + \lambda(c_N \cdot I_N + c_S \cdot I_S + c_E \cdot I_E + c_W \cdot I_W)^t_{i,j} \\ &\le I_{M^t_{i,j}}(1 - \lambda(c_N + c_S + c_E + c_W)^t_{i,j}) \\ &\quad + \lambda I_{M^t_{i,j}}(c_N + c_S + c_E + c_W)^t_{i,j} \\ &= I_{M^t_{i,j}} \end{aligned} \tag{12}$$

and, similarly:

$$I^{t+1}_{i,j} \ge I_{m^t_{i,j}}(1 - \lambda(c_N + c_S + c_E + c_W)^t_{i,j}) + \lambda I_{m^t_{i,j}}(c_N + c_S + c_E + c_W)^t_{i,j} = I_{m^t_{i,j}}. \tag{13}$$

The numerical scheme used to obtain the pictures in this paper is the one given by equations (7), (8), and (10), using the original image as the initial condition, and adiabatic boundary conditions, i.e., setting the conduction coefficient to zero at the boundaries of the image. A constant value for the conduction coefficient c (i.e., $g(\cdot) \equiv 1$) leads to Gaussian blurring (see Fig. 3).

Different functions were used for $g(\cdot)$ giving perceptually similar results. The images in this paper were obtained using

$$g(\nabla I) = e^{(-(\|\nabla I\|/K)^2)}$$

(Fig. 9), and

$$g(\nabla I) = \frac{1}{1 + \left(\frac{\|\nabla I\|}{K}\right)^2}$$

(Figs. 12–14). The scale-spaces generated by these two functions are different: the first privileges high-contrast edges over low-contrast ones, the second privileges wide regions over smaller ones.

The constant K was fixed either by hand at some fixed value (see Figs. 9–14), or using the "noise estimator" described by Canny [4]: a histogram of the absolute values of the gradient throughout the image was computed,

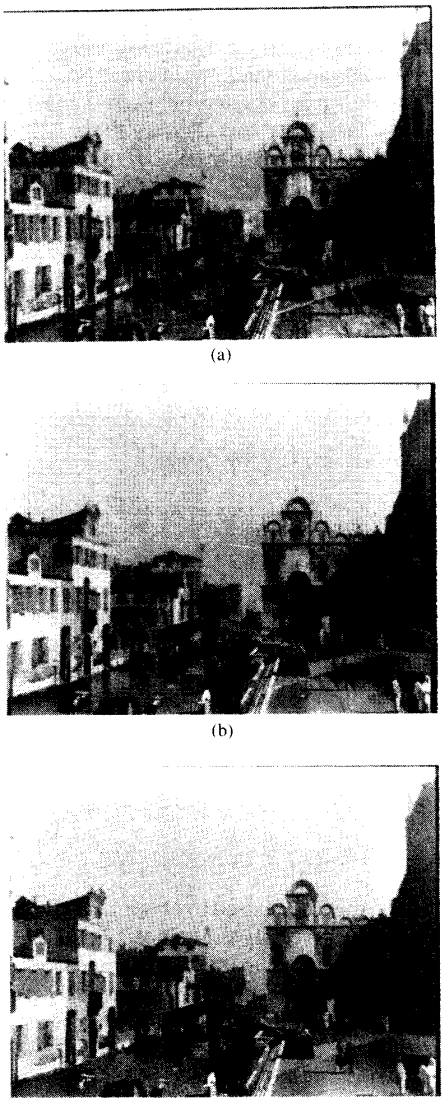


Fig. 9. Effect of anisotropic diffusion (b) on the Canaletto image (a) [3].

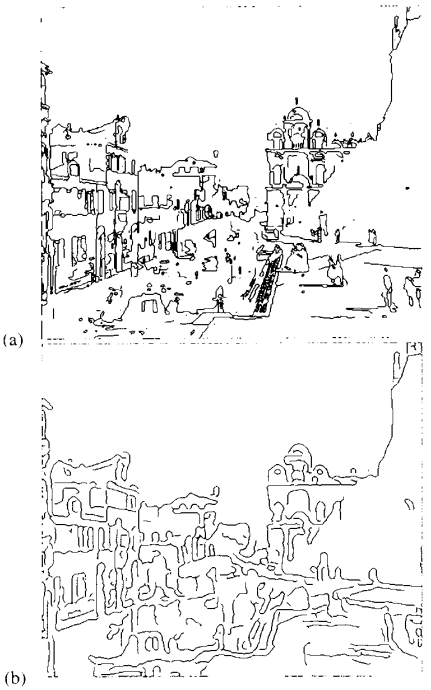


Fig. 10. Edges detected using (a) anisotropic diffusion and (b) Gaussian smoothing (Canny detector).

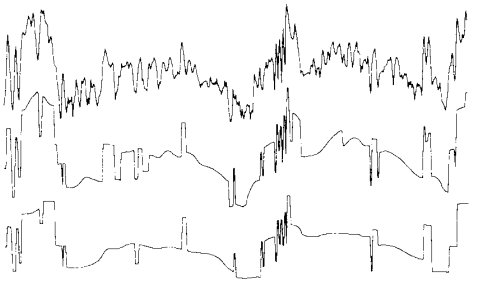


Fig. 11. Scale space obtained with anisotropic diffusion. The diffusion was performed in 2-D on the Canaletto image of which one line (the horizontal line number 400 out of 480—just above the gondola) is shown. Notice that the edges remain sharp until their disappearance.

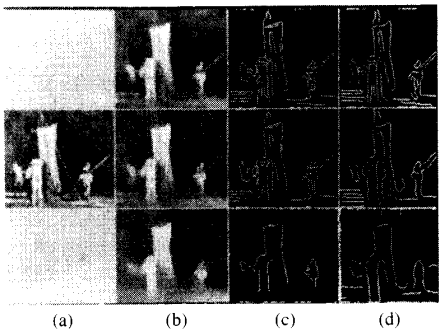


Fig. 12. From left to right (a) original image, (b) scale-space using anisotropic diffusion (10, 20, 80 iterations), (c) edges of the same, (d) edges at comparable scales detected using the Canny detector (convolution kernels of variance 1, 2, 4 pixels).

and K was set equal to the 90% value of its integral at every iteration (see Fig. 12(b)).

The computational scheme described in this section has been chosen for its simplicity. Other numerical solutions of the diffusion equation, and multiscale algorithms may be considered for efficient software implementations.

VI. Comparison to Other Edge Detection Schemes

This section is devoted to comparing the anisotropic diffusion scheme that we present in this paper with previous work on edge detection, image segmentation, and image restoration.

We will divide edge detectors in two classes: fixed-neighborhood edge detectors, and energy/probability "global" schemes.

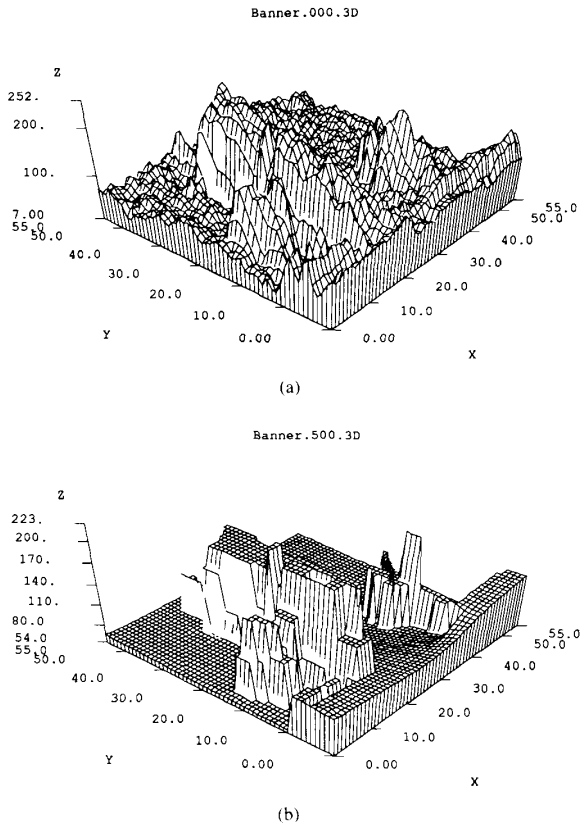


Fig. 13. Scale-space using anisotropic diffusion. Three dimensional plot of the brightness in Fig. 12. (a) Original image, (b) after smoothing with anisotropic diffusion.

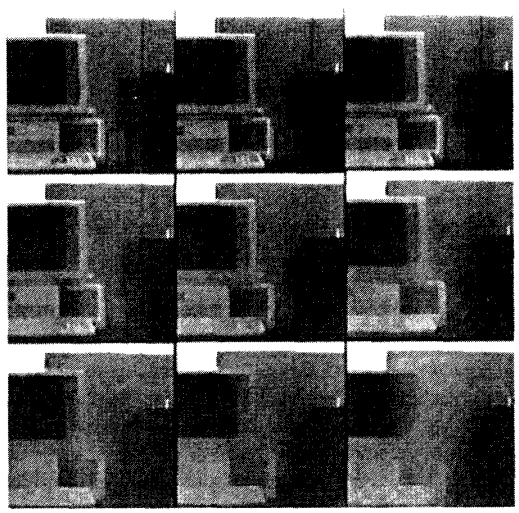


Fig. 14. Scale-space using anisotropic diffusion. Original image (top left) and coarser scale images after (left to right, top to bottom) 20, 60, 120, 160, 220, 280, 320, 400 iterations.

A. Fixed Neighborhood Detectors

This class of detectors makes use of local information only—they typically examine a small window of the image and try to be clever about deciding if and where there is an edge. This decision is ambiguous and difficult.

We pick Canny's scheme [4] as a representative of this class of detectors. The image is convolved with directional derivatives of a Gaussian—the idea is to do smoothing parallel to the edge and thus reduce noise without blurring the edge too much. There are two major difficulties: 1) the inevitable tradeoff between localization accuracy and detectability that comes from using linear filtering 2) the complexity of combining outputs of filters at multiple sales. Anisotropic diffusion is a nonlinear process, hence in principle is not subject to limitation 1). The complication of multiple scale, multiple orientation filters is avoided by locally adaptive smoothing.

We can thus summarize the advantages of the scheme we propose over linear fixed-neighborhood edge detectors.

Locality: The shape and size of the neighborhood where smoothing occurs are determined locally by the brightness pattern of the image, and adapt to the shape and size of the regions within which smoothing is required. In schemes based on linear smoothing or fixed-neighborhood processing the shape and size of the areas where smoothing occurs are constant throughout the image. This causes distortions in the shape of the meaningful regions, and in the loss of structures like edge junctions (see Figs. 10(b), 12(d), 15) which contain much of the information that allows a three-dimensional interpretation of the edge line-drawing [12].

Simplicity: The algorithm consists in identical nearest-neighbor operations (4–8 differences, a function evaluation or a table look-up, and 4–8 sums) iterated over the nodes of a 4 (8) connected square lattice. By comparison the Canny detector requires a number of convolutions (each involving large neighborhoods at a time) as a preprocessing stage, and a stage of cross-scale matching. Moreover with our algorithm the edges are made sharp by the diffusion process discussed in Section IV-B, so that edge thinning and linking are almost unnecessary, especially at coarse scales of resolution (compare Fig. 17 to Fig. 16). For edge detectors based on convolution this is an essential, delicate, and expensive step since linear low-pass filtering has the effect of blurring the edges. The simplicity of the computations involved in anisotropic diffusion makes it a good candidate for digital hardware implementations.

Parallelism: The structure of the algorithm is parallel which makes it cheap to run on arrays of simple parallel processors.

On sequential machines, anisotropic diffusion is computationally more expensive than convolution-based detectors. This is because in the diffusion process a continuum of scales are generated instead of a small fixed number.

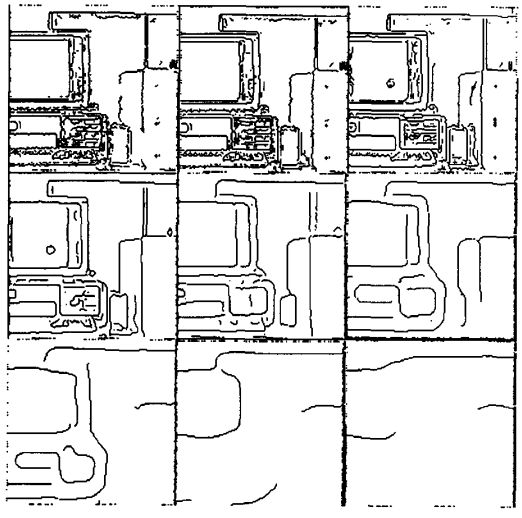


Fig. 15. Scale-space using linear convolution. The edges are distorted and the junctions disappear. Images generated using the Canny detector and smoothing Gaussian kernels of variance (top left to bottom right) 1/2, 1, 2, 4, 8, 16 pixels. Compare to Fig. 17 where anisotropic diffusion preserves edge junctions, shape, and position.

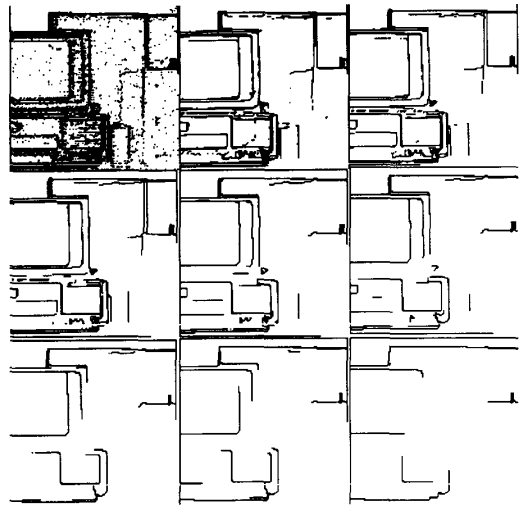


Fig. 16. Edges detected by thresholding the gradient in Fig. 14. Linking is not necessary. Thinning is only for the finer scales. Compare to Fig. 17 where thinning and linking have been used.

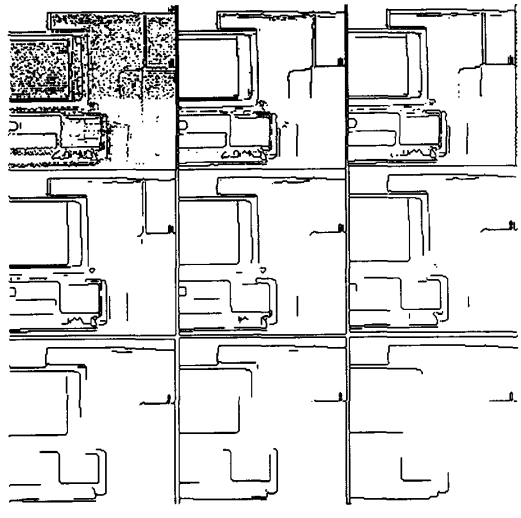


Fig. 17. Edges detected in Fig. 14 using a thinning and linking stage [4].

B. Energy-Based Methods for Image Reconstruction and Segmentation

A number of methods have appeared in the literature where the edge detection/image segmentation process is performed by the minimization of an energy function of type

$$U(\vec{z}) = \sum_{i \in I, j \in N(i)} V(z_i, z_j) + \sum_{i \in I} W_i(z_i) \qquad (14)$$

with I indicating the set of the nodes of a lattice, $N(i) \subset I$ indicating the nodes neighboring node i, and z a function defined on the lattice, typically the brightness function [2]. An equivalent formulation is based on finding maxima of a Markov probability distribution function defined on the space of all images:

$$P_Z(\vec{z}) = \frac{1}{K} e^{-U(\vec{z})} \qquad (15)$$

where the function $U(\cdot)$ has the form of (14) [6], [14]. Because the exponential function is monotonic the maxima of the probability distribution and the minima of the energy function coincide, and we can limit our attention to the schemes based on minimizing the energy.

The energy function (14) is the sum of two terms: the *a priori* term (the sum of the "clique" functions V containing the *a priori* knowledge about the image space—see any one of [6], [16], [2] for a complete discussion), and a term depending on the data available (the sum of the functions W_i). $V(\cdot, \cdot)$ is typically an even function depending only on the value of the difference of its arguments (with abuse of notation $V(z_i, z_j) = V(z_i - z_j)$). It has minimum at zero and it is monotonic on the positive and negative semilines assigning higher energy ($\Leftrightarrow$ lower probability) to the pairs i, j of lattice nodes whose brightness difference $\| z_i - z_j \|$ is bigger. We will show that the approximation of anisotropic diffusion that we suggest in Section V may be seen as a gradient descent of the *a priori* part of the energy function

$$U(\vec{z}) = \sum_{i \in I, j \in N(i)} V(z_i, z_j). \qquad (16)$$

The steepest descent strategy for finding minima of a function consists of starting from some initial state, and then changing iteratively the state following the opposite of the gradient vector. The gradient of the energy function, which may be computed from (16) differentiating with respect to the z_i, is the vector of components

$$\nabla U_i(z) = 2 \sum_{j \in N(i)} \dot{V}(z_i - z_j) \qquad (17)$$

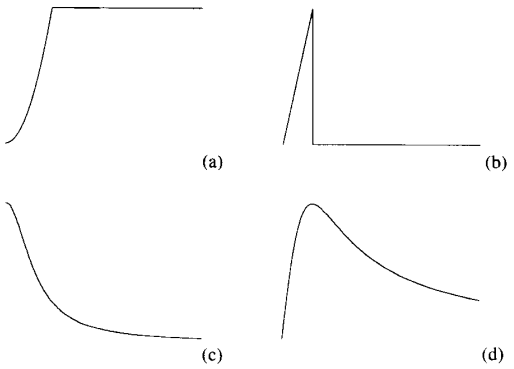


Fig. 18. (a) The local energy function proposed by [6], [2], [14] typically is equal to the square of the nearest-neighbor brightness difference, and saturates at some threshold value. (b) The first derivative of the energy function (a). (c), (d) The anisotropic diffusion conduction coefficient and flux function as a function of the brightness gradient magnitude, proportional to the nearest neighbor brightness difference in the discrete case. (b) and (d) have the same role.

therefore the gradient descent algorithm is

$$\frac{\partial}{\partial t} z_i = -A \cdot \sum_{j \in N(i)} \dot{V}(z_i - z_j) \qquad (18)$$

where A is some "speed" factor.

Suppose that $V(\cdot)$ is differentiable in the origin and define $\phi(\cdot) = -\dot{V}$. Since $V(\cdot)$ is even, $\phi(\cdot)$ is an odd function and $\phi(0) = 0$. Then we may write $\phi(s) = s \cdot c(s)$ for some function $c(\cdot)$ even and positive. Substituting into (18) we obtain

$$\frac{\partial}{\partial t} z_i = A \cdot \sum_{j \in N(i)} c(z_j - z_i) \cdot (z_j - z_i) \qquad (19)$$

which is exactly the anisotropic diffusion algorithm defined by (7), (8), and (10) if the neighborhood structure is given by natural nearest-neighbor cliques of a square lattice. The flux functions obtained by differentiating the local energy functions $V(\cdot)$ of [6], [15], [2] are similar to the shape of flux function that the analysis in Section IV-B suggests. See Fig. 18.

To summarize: anisotropic diffusion may be seen as a gradient descent on some energy function. The data (the original image) are used as the initial condition. In the energy-based methods [6], [16], [2] the closedness of the solution to the data is imposed by a term in the energy function. This makes the energy function nonconvex and more complicated optimization algorithms than gradient descent are necessary. Most of the algorithms that have been proposed (simulated annealing for example) appear too slow for vision applications. Perhaps the only exception is the GNC algorithm proposed by Blake and Zisserman [2] which does not guarantee to find the global optimum for generic images, but appears to be a good compromise between speed and accuracy.

VII. Conclusion

We have introduced a tool, anisotropic diffusion, that we believe will prove useful in many tasks of early vision. Diffusion based algorithms involve simple, local, identical computations over the entire image lattice. Implementations on massively parallel architectures like the connection machine would be almost trivial. Implementations using hybrid analog-digital networks also seem feasible.

We have shown that the simplest version of anisotropic diffusion can be applied with success to multiscale image segmentation. As a preprocessing step it makes thinning and linking of the edges unnecessary, it preserves the edge junctions, and it does not require complicated comparison of images at different scales since shape and position are preserved at every single scale.

In images in which the brightness gradient generated by the noise is greater than that of the edges, and the level of the noise varies significantly across the image the scheme that we have described proves insufficient to obtain a correct multiscale segmentation. In this case a global noise estimate does not provide an accurate local estimate, and the local value of the gradient provides too partial a piece of information for distinguishing noise-related and edge-related gradients. Moreover, the abscissa K of the peak of the flux function $\phi(\cdot)$ has to be set according to the typical contrast value, if this changes considerably through the image the value of K has to be set locally. To tackle these difficulties anisotropic diffusion should be implemented using local contrast and noise estimates.

Appendix
Proof of the Maximum Principle

Call $\boldsymbol{A}$ an open bounded set of $\mathbb{R}^n$ (in our case $\boldsymbol{A}$ is the plane of the image, a rectangle of $\mathbb{R}^2$), and $\boldsymbol{T} = (a, b)$ an interval of $\mathbb{R}$. Let $\boldsymbol{D}$ be the open cylinder of $\mathbb{R}^{n+1}$ formed by the product $\boldsymbol{D} = \boldsymbol{A} \times \boldsymbol{T} = \{(x, t) : x \in \boldsymbol{A}, t \in \boldsymbol{T}\}$. Call $\partial \boldsymbol{D}$ the boundary of $\boldsymbol{D}$, $\overline{\boldsymbol{D}}$ its closure, and $\partial_T \boldsymbol{D}$, $\partial_S \boldsymbol{D}$, and $\partial_B \boldsymbol{D}$ the top, side, and bottom portions of $\partial \boldsymbol{D}$:

$$\partial_T \boldsymbol{D} = \{(x, t) : x \in \boldsymbol{A}, t = a\}$$
$$\partial_S \boldsymbol{D} = \{(x, t) : x \in \partial \boldsymbol{A}, t \in \boldsymbol{T}\}$$
$$\partial_B \boldsymbol{D} = \{(x, t) : x \in \boldsymbol{A}, t = b\}$$

and, for convenience, call $\partial_{SB} \boldsymbol{D}$ the side-bottom boundary:

$$\partial_{SB} \boldsymbol{D} = \partial_S \boldsymbol{D} \cup \partial_B \boldsymbol{D}.$$

The following theorems hold.

Theorem: Consider a function $f : \mathbb{R}^{n+1} \to \mathbb{R}$ that is continuous on $\overline{\boldsymbol{D}}$, and twice differentiable on $\boldsymbol{D} \cup \partial_T \boldsymbol{D}$. If f satisfies the differential inequality

$$C(x, t) f_t - c(x, t) \triangle f - \nabla c \cdot \nabla f \leq 0 \qquad (20)$$

on $\boldsymbol{D}$, with $C : \mathbb{R}^{n+1} \to \mathbb{R}^+$ continuous on $\overline{\boldsymbol{D}}$, and differentiable on $\boldsymbol{D} \cup \partial_T \boldsymbol{D}$, then it obeys the maximum principle, i.e., the maximum of f in $\overline{\boldsymbol{D}}$ is reached on the bottom-side boundary $\partial_{SB} \boldsymbol{D}$ of $\boldsymbol{D}$:

$$\max_{\overline{\boldsymbol{D}}} f = \max_{\partial_{SB} \boldsymbol{D}} f.$$

Corollary: Consider a function f satisfying the hypotheses of the previous theorem, and such that f is twice differentiable on $\partial_S \boldsymbol{D}$, and $\nabla_x f = 0$ (where ∇_x indicates the gradient operator along the x direction). Then

$$\max_{\overline{\boldsymbol{D}}} f = \max_{\partial_B \boldsymbol{D}} f.$$

The following proof is adapted from John [10].

Proof: First consider f satisfying the stricter condition

$$C(x, t) f_t - c(x, t) \Delta f - \nabla c \cdot \nabla f < 0. \qquad (21)$$

By hypothesis f is continuous on $\overline{\boldsymbol{D}}$, a compact set, hence it has a maximum in it. Call $\boldsymbol{p} = (y, \tau)$ this maximum.

Suppose that $\boldsymbol{p} \in \boldsymbol{D}$. Since f is twice continuously differentiable in $\boldsymbol{D}$ we can write the first three terms of the Taylor expansion of f about $\boldsymbol{p}$:

$$f(\boldsymbol{p} + \epsilon \boldsymbol{v}) = f(\boldsymbol{p}) + \epsilon \nabla f^T \boldsymbol{v} + \epsilon^2 \boldsymbol{v}^T \mathcal{H} f \boldsymbol{v} + O(\epsilon^3) \leq f(\boldsymbol{p}) \qquad (22)$$

where $\boldsymbol{v} \in \mathbb{R}^{n+1}$, $\epsilon \in$ some neighborhood of zero, and $\mathcal{H}f$ indicates the $n + 1 \times n + 1$ Hessian matrix of f. For the sake of compactness, unlike in the rest of the paper, ∇f in (22) indicates the gradient of f with respect to the space coordinates *and* the time coordinate. Since $\boldsymbol{p}$ is a point where f has a maximum, the gradient ∇f in the first order term of the expansion (22) is equal to zero therefore the second term cannot be positive, $\forall \boldsymbol{v} \in \mathbb{R}^{n+1} : \boldsymbol{v}^T \mathcal{H} f \boldsymbol{v} \leq 0$; the Hessian matrix is therefore negative semidefinite, which implies that the entries on its diagonal are either equal to zero or negative. The Laplacian is a sum of entries on the diagonal and therefore $\Delta f \leq 0$. This would mean that at $\boldsymbol{p}$

$$C(x, t) f_t - c(\boldsymbol{p}) \Delta f - \nabla c \cdot \Delta f \geq 0$$

contradicting the hypothesis.

Similarly, if $\boldsymbol{p} \in \partial_T \boldsymbol{D}$ the first derivative with respect to t of f could only be positive or equal to zero, while the first derivatives with respect to the x variables would have to be equal to zero, and the second derivatives with respect to the x variables could only be equal to zero or negative, giving the same inequality at $\boldsymbol{p}$ as above. This would again contradict the hypothesis. So, if f satisfies (21), then it obeys the maximum principle.

If f satisfies the weak inequality (20) the function g defined as $g = f - \lambda(t - a)$ satisfies the strict inequality (21), and therefore the maximum principle, for any $\lambda > 0$. Observe that $f = g + \lambda(t - a) \leq g + \lambda(b - a)$ on $\overline{\boldsymbol{D}}$, and because of this

$$\max_{\overline{\boldsymbol{D}}} f \leq \max_{\overline{\boldsymbol{D}}} \left(g + \lambda(b - a) \right) = \max_{\partial_{SB} \boldsymbol{D}} \left(g + \lambda(b - a) \right) \leq \max_{\partial_{SB} \boldsymbol{D}} \left(f + \lambda(b - a) \right).$$

Letting $\lambda \to 0$ we obtain the thesis. □

Notice that the maximum principle also guarantees that there are no local maxima of f in $\boldsymbol{D} \cup \partial_T \boldsymbol{D}$. The same technique used in the proof restricting $\boldsymbol{D}$ to be a cylinder contained in the neighborhood where the local maximum is a strict maximum may be used to see that the existence of one at $p \in \boldsymbol{D} \cup \partial_T \boldsymbol{D}$ would violate the differential inequality.

The corollary may be proven along the same lines: since f is, by hypothesis, differentiable on $\partial_S \boldsymbol{D}$ one can use (21), and (22) for any $\boldsymbol{p} \in \partial_S \boldsymbol{D}$, with $\boldsymbol{v}$ in an appropriate hemisphere so that $\boldsymbol{p} + \epsilon \boldsymbol{v} \in \boldsymbol{D}$.

If a function f satisfies the differential equation

$$C(x, t) f_t - c(x, t) \triangle f - \nabla c \cdot \nabla f = 0 \qquad (23)$$

with the hypotheses already stated on the functions $C(\cdot)$ and $c(\cdot)$, the arguments above can be run for f and $h = -f$ proving that both a maximum and minimum principle have to be satisfied.

The diffusion equation (3) is a special case of (23) (set $C(x, t) = 1$, and $f = I$), hence the scale-space brightness function $I(x, y, t)$ obeys the maximum principle provided that the conduction coefficient c never takes negative value (in fact the condition that c does not take negative value where f has a maximum is sufficient) and is differentiable. If adiabatic ($\nabla_x f = 0$) boundary conditions are used then the hypotheses of the corollary are satisfied too, and the maxima may only belong to the initial condition.

Solutions f of (3) have an additional property if the conduction coefficient is constant along the space axes: $c = c(t)$. In this case, all spatial derivatives of f are solutions of (3), and therefore satisfy the hypotheses of the maximum principle. So the causality criterion is satisfied by all such functions: the components of the gradient, the Laplacian, etc. It is important to notice that this is not true in general for solutions of (3) when the conduction coefficient varies in scale and space. We show in Section IV-B that in fact anisotropic diffusion can increase the contrast (i.e., the magnitude of the gradient) of edges in the image.

Acknowledgment

We are grateful to L. Semenzato, A. Casotto, P. Kube, and B. Baringer who gave very friendly assistance in setting up the software simulations, and taking the pictures. R. Brodersen kindly provided the photographic equipment. B. Hummel pointed to us the result by Niremberg.

References

[1] J. Babaud, A. Witkin, M. Baudin, and R. Duda, "Uniqueness of the gaussian kernel for scale-space filtering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, Jan. 1986.

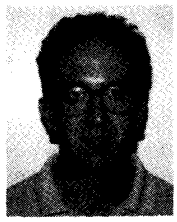
[2] A. Blake and A. Zisserman, *Visual Reconstruction*. Cambridge, MA: MIT Press, 1987.

[3] Canaletto, "View in Venice," National Gallery of Art, Washington, DC, circa 1740.

[4] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, pp. 679–698, 1986.

[5] J. Clark, "Singularity theory and phantom edges in scale space," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, no. 5, pp. 720–727, 1988.

[6] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 721-741, Nov. 1984.
[7] A. Hummel, "Representations based on zero-crossings in scale-space," in *Proc. IEEE Computer Vision and Pattern Recognition Conf.*, June 1986, pp. 204-209; reproduced in: *Readings in Computer Vision: Issues, Problems, Principles and Paradigms*, M. Fischler and O. Firschein, Eds. Los Altos, CA: Morgan Kaufmann, 1987.
[8] ——, "The scale-space formulation of pyramid data structures," in *Parallel Computer Vision*, L. Uhr, Ed. New York: Academic, 1987, pp. 187-223.
[9] A. Hummel, B. Kimia, and S. Zucker, "Deblurring Gaussian blur," *Comput. Vision, Graphics, Image Processing*, vol. 38, pp. 66-80, 1987.
[10] F. John, *Partial Differential Equations*. New York: Springer-Verlag, 1982.
[11] J. Koenderink, "The structure of images," *Biol. Cybern.*, vol. 50, pp. 363-370, 1984.
[12] J. Malik, "Interpreting line drawings of curved objects," *Int. J. Comput. Vision*, vol. 1, no. 1, pp. 73-103, 1987.
[13] D. Marr, *Vision*. San Francisco, CA: Freeman, 1982.
[14] J. Marroquin, "Probabilistic solution of inverse problems," Ph.D. dissertation, Massachusetts Inst. Technol., 1985.
[15] ——, "Probabilistic solution of inverse problems," Artificial Intell. Lab., Massachusetts Inst. Technol., Tech. Rep. AI-TR 860, 1985.
[16] D. Mumford and J. Shah, "Optimal approximation of piecewise smooth functions and associated variational problems," *Commun. Pure Appl. Math.*, vol. 42, pp. 577-685, 1989.
[17] L. Nirenbarg, "A strong maximum principle for parabolic equations," *Commun. Pure Appl. Math.*, vol. VI, pp. 167-177, 1953.
[18] P. Perona and J. Malik, "Scale space and edge detection using anisotropic diffusion," in *Proc. IEEE Comput. Soc. Workshop Computer Vision*, Miami, FL, 1987, pp. 16-27.
[19] ——, "A network for edge detection and scale space," in *Proc. IEEE Int. Symp. Circuits and Systems*, Helsinki, June 1988, pp. 2565-2568.
[20] A. Rosenfeld and M. Thurston, "Edge and curve detection for visual scene analysis," *IEEE Trans. Comput.*, vol. C-20, pp. 562-569, May 1971.
[21] A. Witkin, "Scale-space filtering," in *Int. Joint Conf. Artificial Intelligence*, Karlsruhe, West Germany, 1983, pp. 1019-1021.
[22] A. Yuille and T. Poggio, "Scaling theorems for zero crossings," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, Jan. 1986.



Pietro Perona was born in Padua, Italy, on September 3, 1961. He received the Doctor degree in electrical engineering cum laude from the University of Padua in 1985 with a thesis on dynamical systems theory.

He received the Ph.D. degree from the Department of Electrical Engineering and Computer Science of the University of California at Berkeley in 1990. His research interests are in computational and biological vision.

Jitendra Malik (A'88) was born in Mathura, India, on October 11, 1960. He received the B.Tech degree from Indian Institute of Technology, Kanpur, in 1980 where he was awarded the gold medal for the best graduating student in electrical engineering. He received the Ph.D. degree in computer science from Stanford University, Stanford, CA, in 1986.

Since January 1986, he has been an Assistant Professor in the Computer Science Division, Department of EECS, University of California at Berkeley. Since October 1988 he has also been a member of the group in Physiological Optics at UC Berkeley. His research interests are in machine vision and computational modeling of early human vision. These include work on edge detection, texture segmentation, line drawing interpretation, and 3-D object recognition.

Dr. Malik received a Presidential Young Investigator Award in 1989.