Multiscale Image Segmentation by Integrated Edge and Region Detection

Mark Tabb and Narendra Ahuja, *Fellow, IEEE*

***Abstract*—This paper is concerned with the detection of low-level structure in images. It describes an algorithm for image segmentation at multiple scales. The detected regions are homogeneous and surrounded by closed edge contours. Previous approaches to multiscale segmentation represent an image at different scales using a scale-space. However, structure is only represented implicitly in this representation, structures at coarser scales are inherently smoothed, and the problem of structure extraction is unaddressed. This paper argues that the issues of scale selection and structure detection cannot be treated separately. A new concept of scale is presented that represents image structures at different scales, and not the image itself. This scale is integrated into a nonlinear transform which makes structure explicit in the transformed domain. Structures that are stable (locally invariant) to changes in scale are identified as being perceptually relevant. The transform can be viewed as collecting spatially distributed evidence for edges and regions, and making it available at contour locations, thereby facilitating integrated detection of edges and regions without restrictive models of geometry or homogeneity. In this sense, it performs Gestalt analysis. All scale parameters of the transform are automatically determined, and structure of any arbitrary geometry can be identified without any smoothing, even at coarse scales.**

I. Introduction

THIS PAPER addresses the classical problem of detecting low-level structure in images, or image segmentation. This problem involves the identification of local areas (regions) in an image that are homogeneous and dissimilar to all spatially adjacent regions. Homogeneity may be measured in terms of color, texture, motion, depth, etc., but for the purposes of this paper it is measured by graylevel similarity. Structure identification is inherently a multiscale problem because image structure is recursive, i.e., regions may contain substructure, which themselves contain substructure, etc. Of course, digital images have finite resolution, so the number of levels of structure is limited (typically, 3–4). As an example of the multiscale nature of structure, consider Fig. 1, which consists of an image of a sailboat on a lake. The cloud mass and water are each single regions at a scale which allows for significant intensity variation. However, if sensitivity to intensity variation is increased, then individual clouds and the streaks within the water should be identified as regions.

Manuscript received December 13, 1994; revised December 14, 1995. This work was supported in part by the Advanced Research Projects Agency under Grant N00014-93-1-1167 and the NSF under Grant IRI-93:19038. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. William E. Higgins.

The authors are with the Beckman Institute, Coordinated Science Laboratory, and the Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801 USA (e-mails: mtabb@vision.ai.uiuc.edu; ahuja@vision.ai.uiuc.edu).

Publisher Item Identifier S 1057-7149(97)02461-5.

Fig. 1. Image of a sailboat on a lake. The image contains significant multiscale structure. For example, at a scale where regions have significant intensity variation, the cloud mass and water each can be considered as single regions. However, if sensitivity to intensity variation is increased, individual clouds and the streaks within the water should be identified as regions.

This paper assumes no *a priori* knowledge of either the scale or geometry of any structures present within an image. The goal is to identify all image structures without any smoothing of the structure boundaries, regardless of the scale at which the structures occur, and without incorrectly identifying any nonrelevant areas as structures. Perceptual validity is used as a subjective measure of structure relevance. A large number of algorithms for image segmentation have been proposed over the years. However, almost all of these methods completely ignore the issue of scale. As a result, they are capable of identifying only a limited variety of structures. Some examples of these approaches include: thresholding techniques [1], [2], region growing [2]–[4], split-and-merge [5], watersheds [6], rule-based systems [7], and Markov random field-based (MRF-based) models [8]. Prior attempts at multiscale segmentation [9]–[12] all involve the creation of a scale-space [9], [13]–[15] wherein an image is represented at different scales by a single scale parameter. Although scale-space produces a hierarchy which represents a signal at different levels of detail, it is not a structural decomposition. Structure is contained implicitly within the scale-space, and the representation of the image

at any given scale may actually contain multiscale structure. Hence, a multiscale segmentation algorithm is still required to extract the image structures from the scale-space decomposition. The basic problem with scale-space is that it represents an image at different scales, but not the image structure. Further, all scale-spaces inherently involve some smoothing of the image, which causes the boundaries of any extracted structure to be smoothed as well.

A new approach to segmentation is presented in this paper. Scale is defined such that it represents image structure at different scales, not the image itself. For each structure, this involves the use of a homogeneity scale which describes the amount of homogeneity variation within the structure, and a set of spatial scales which describe the size and shape of the structure. This concept of scale is integrated into a novel nonlinear transform introduced in [16]–[18]. The structure at any given scale is represented explicitly in the transformed domain. This allows easy examination of the change in structure due to a change in scale. Structures that are stable (locally invariant) with scale are identified as corresponding to perceptually valid image structure. In this manner, the processes of scale selection and structure identification are integrated and performed automatically.

The transform can be viewed as collecting spatially distributed evidence for edges and regions and making it explicitly available at contour locations, and, in this sense, it performs Gestalt analysis. Further, it does this without using restrictive models of structure geometry or intensity variation, and without the need for any user-specified parameters. This allows the identification of precise structure boundaries that are completely unsmoothed, even at coarse scales. The transform incorporates into its definition the duality of region- and edge-based descriptions of image structure, namely, that the regions have a smooth variation of some property (e.g., intensity) in the interior and a sharp discontinuity across the boundary. Thus, the transform performs integrated edge and region detection, and can be viewed as a multiscale edge and blob detector at the same time.

The rest of this paper is organized as follows: Section II reviews the general transform and presents the specific form used in this paper. Section III discusses the extraction of image structure from the transformed image, and Section IV gives experimental results for synthetic and real images. Section V presents concluding remarks.

II. The Transform

In this section, we briefly review the transform for a continuous, two-dimensional (2-D) gray-scale image. Extensions to discrete space, arbitrary dimension, and vector-valued data are straightforward. Details can be found in [16]–[18]. The general form of the transform maps an image, $I(x,y)$, into a family of attraction force fields, $\mathbf{F}(x,y;\sigma_g(x,y),\sigma_s(x,y))$, as follows:

$$\mathbf{F}(x,y;\ \sigma_g(x,y),\sigma_s(x,y)) = \iint_R d_g(\Delta I,\sigma_g(x,y))\cdot d_s(\vec{r},\sigma_s(x,y))\frac{\vec{r}}{\|\vec{r}\|}dwdv \quad (1)$$

where $R = \text{domain}(I(u,v))\backslash\{(x,y)\}$ and $\vec{r} = (v-x)\vec{i} + (w-y)\vec{j}$. For each pixel (point) in an image, the transform analyzes the intensities present in a neighborhood of the pixel, and produces a force vector. Associated with each pixel is a homogeneity scale, σ_g, which reflects the homogeneity of the region into which the pixel groups, and a spatial scale, σ_s, which controls the neighborhood from which the force on the pixel is computed. The force field encodes the region structure in a manner which allows easy extraction. The scale parameters, σ_g and σ_s, determine the scale at which the image regions get encoded within the field.

The transform computes at each pixel a vector sum of pairwise affinities between the pixel and all other pixels. The resultant vector produced by the transform at each pixel defines both the direction and magnitude of attraction experienced by the pixel from the rest of the image. The affinity between a pair of pixels is determined by the scale parameters. The spatial scale parameter, σ_s, controls the spatial distance function, $d_s(\cdot)$, and the homogeneity scale parameter, σ_g, controls the homogeneity distance function, $d_g(\cdot)$. For a grayscale image, the homogeneity between two pixels, ΔI, is given by

$$\Delta I = |I(x,y) - I(v,w)|. \quad (2)$$

With the above definition of the force field, $\mathbf{F}$, pixels are grouped together into regions whose boundaries correspond to diverging force vectors in $\mathbf{F}$ and whose skeletons correspond to converging force vectors in $\mathbf{F}$. In addition, an increase in σ_g should cause less homogeneous structures to be encoded, and an increase in σ_s should cause larger structures to be encoded. To ensure such a relationship between scale and structure, the distance functions, $d_s(\cdot)$ and $d_g(\cdot)$, should possess the following properties:

1) *Unit Range*: The transform measures the degree of attraction among pixels, not repulsions. Therefore, $0 \le d_g(\cdot), d_s(\cdot) \le 1$ is required.
2) *Decreasing Attraction (image characteristic)*: The degree of attraction between pixels should be directly proportional to their similarity, $d_g(\Delta I_1,\sigma_g) \ge d_g(\Delta I_2,\sigma_g)$ for $\Delta I_1 \le \Delta I_2,\ \forall\sigma_g$, and proximity, $d_s(\vec{r}_1,\sigma_s) \ge d_s(\vec{r}_2\sigma_s)$ for $\|\vec{r}_1\| \le \|\vec{r}_2\|,\ \forall\sigma_s$.
3) *Increasing Attraction (scale characteristic)*: Pixel similarity should be directly proportional to both homogeneity scale, $d_g(\Delta I,\sigma_g^1) \le d_g(\Delta I,\sigma_g^2)$ for $\sigma_g^1 \le \sigma_g^2$, and spatial scale, $d_s(\vec{r},\sigma_s^1) \le d_s(\vec{r},\sigma_s^2)$ for $\sigma_s^1 \le \sigma_s^2$.
4) *Isotropicity*: Structure should not be detected by the transform preferentially in any direction. Thus, we need $d_s(\vec{r},\sigma_s) = f(\|\vec{r}\|,\sigma_s)$.
5) *Locality*: The field at each pixel should depend only on a local (albeit adaptively determined) neighborhood around that pixel. Hence, let $d_s(\vec{r},\sigma_s) = 0$ for $\|\vec{r}\| > c\cdot\sigma_s$, for some constant, c.

Two possible forms for these functions satisfying the above criteria are unnormalized Gaussian

$$d_g(\Delta I,\sigma_g) \sim \sqrt{2\pi\sigma_g^2}N_{\Delta I}\left(0,\sigma_g^2\right) \quad (3)$$

$$d_s(\vec{r},\sigma_s) \sim \begin{cases}\sqrt{2\pi\sigma_s^2}N_{\|\vec{r}\|}\left(0,\sigma_s^2\right), & \|\vec{r}\| \le 2\sigma_s \\ 0, & \|\vec{r}\| > 2\sigma_s\end{cases} \quad (4)$$

and box-car window

$$d_g(\Delta I, \sigma_g) \sim B_{\Delta I}(\sigma_g) \tag{5}$$

$$d_s(\vec{r}, \sigma_s) \sim B_{||\vec{r}||}(\sigma_s) \tag{6}$$

where

$$B_x(c) = \begin{cases} 1, & |x| \le c \\ 0, & \text{else} \end{cases} \tag{7}$$

Although $\mathbf{F}$ is not invariant to the form of $d_g(\cdot)$ and $d_s(\cdot)$, different forms result in minimal change in the encoded structure information, so long as the forms satisfy the listed criteria. This is due to the property that structure is represented as converging and diverging vectors which are locally invariant to changes in scale. Experiments have been performed using different functions (box-car, Gaussian, linear, exponential) for both $d_g(\cdot)$ and $d_s(\cdot)$ on various real images. There was no change in the detected structures using different forms for $d_s(\cdot)$, and only marginal changes for different forms of $d_g(\cdot)$. A box-car function for $d_g(\cdot)$ gives the best results, a point which is explained in Section II-D, after some necessary concepts have been introduced in the interceding sections.

A. Specific Formulation of Scale Used for Image Segmentation

Two independent scale parameters per pixel is overgeneral for characterizing image structure. In this paper, a structure is modeled by a single σ_g which characterizes the degree of homogeneity of the structure, and a σ_s at each pixel which characterizes its size and shape. This number of spatial scales is necessary in order to identify a structure without any spatial smoothing. For example, identifying an octopus-like structure (large, main body with fine appendages) requires large values of σ_s within the body and small values within the appendages. In contrast, if retaining fine boundary detail is not necessary, then a (σ_g, σ_s) pair is sufficient to represent the scale of a structure. In Section II-B, it is shown that, if the value of σ_g for which a pixel belongs to a particular structure is known, then the appropriate value for σ_s can be determined automatically. This allows (1) to be simplified to

$$\mathbf{F}(x, y;\ \sigma_g(x, y), \sigma_s(x, y)) = \mathbf{F}_{\sigma_g}(x, y). \tag{8}$$

The homogeneity scale, σ_g, is made spatially invariant and is the sole independent scale parameter to the transform. This results in an image being mapped into a one-parameter family of attraction force fields, $\mathbf{F}_{\sigma_g}$, which contains all of the multiscale structure present in an image, thereby simplifying the problem of structure identification to a one-dimensional (1-D) search.

In addition to regarding the set of spatial scales associated with a structure as describing its size and shape, an alternative (but equivalent) interpretation is that the spatial scales describe the appropriate amount of spatial information necessary to identify a structure having a given degree of homogeneity. For example, if σ_s at a pixel is too small, then there may not be enough spatial information available for the transform to determine to which region the pixel should group. Similarly, if σ_s is too large, then the transform may group the pixel into a region containing very distant and unrelated pixels. Fig. 2 gives an example which demonstrates the necessity of selecting σ_s properly in order to determine whether or not a region boundary (edge) is present at a pixel. This example is independent of the method actually used to perform the region (edge) identification (the transform in this case).

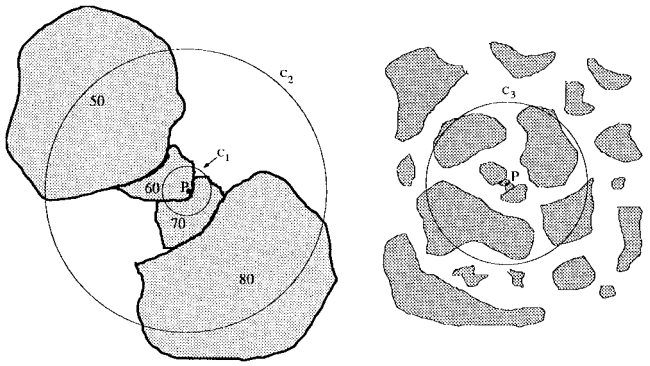
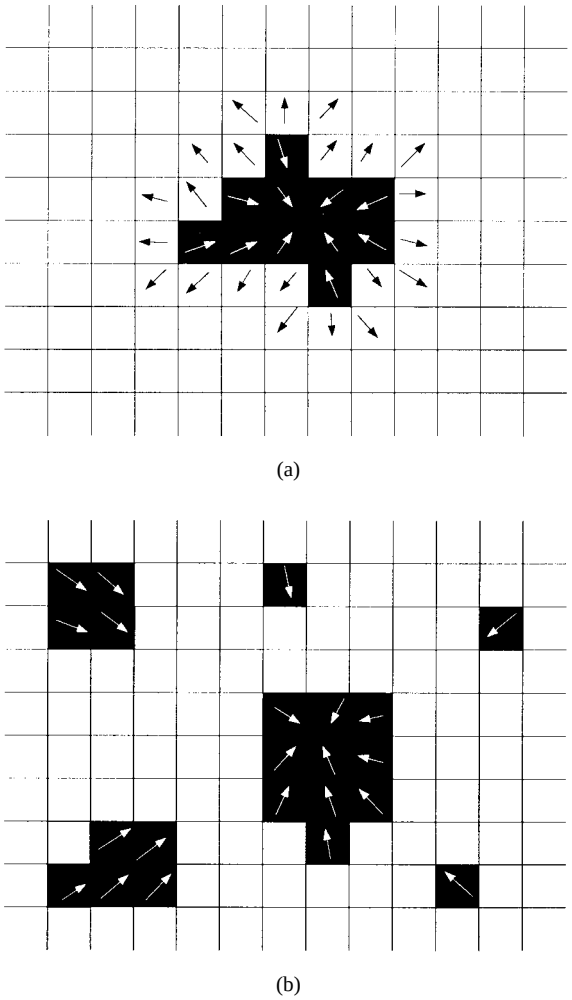


Fig. 2. (a) For $\sigma_g < 10$, the circle C_1 contains enough information to identify that an edge is present at Point P. For $10 \le \sigma_g < 30$, the edge should still be present, but C_1 does not contain enough information to determine this. However, C_2 does, (b) This illustrates the problem of overscaling.There is too much information within circle C_3 to determine whether or not an edge is present at Point P. All local structure information has been lost.

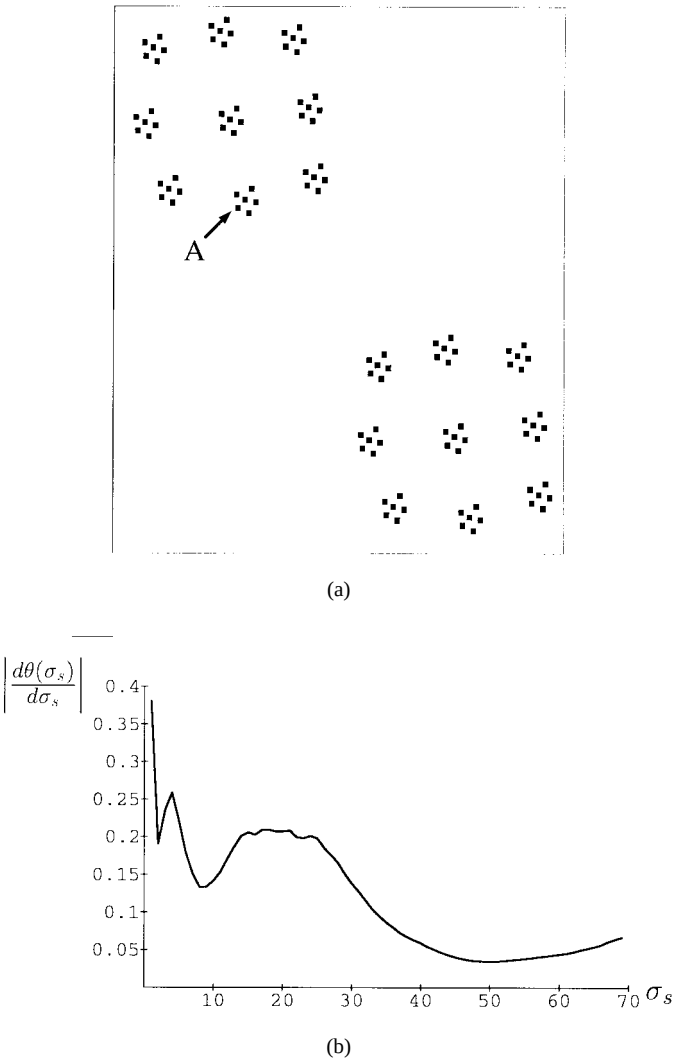
There are four regions of constant intensity in Fig. 2(a), each labeled by its intensity. The most reasonable groupings with respect to σ_g for these regions are {50, 60, 70, 80} for $0 \le \sigma_g < 10$, {50–60, 70–80} for $10 \le \sigma_g < 30$, and {50–60–70–80} for $\sigma_g \ge 30$. Consider the point P which lies on the high-frequency edge separating regions 60 and 70, and the low-frequency edge between regions 50 and 80. Let the spatial information usable for determining whether or not an edge is present at P be contained within a circle of radius σ_s centered at P, and also let r_i denote the radius of circle C_i. For $\sigma_g < 10$, $\sigma_s = r_1$ provides enough information for determining that an edge should be present at P. For $10 \le \sigma_g < 30$ an edge should still be present, but $\sigma_s = r_1$ does not provide enough information to support this conclusion; however, $\sigma_s = r_2$ does. Too much information is just as much of a problem as too little. For both these values of σ_g, choosing $\sigma_s = r_3$ as in Fig. 2(b) provides much more information than necessary. Effectively, the local structural information is lost. It should be apparent from Fig. 2 that, for each value of σ_g at each pixel, there is some appropriate range of values for σ_s which is necessary for structure identification to be possible.

Within the context of the transform, selecting an appropriate value of σ_s at a pixel, for a given σ_g, corresponds to the pixel belonging to a region of contracting flow (inward force vectors) defined by contiguous pixels. Such a region is termed a *region of attraction*. The region boundary is the source of the flow. Consider a region whose boundary is given by a closed curve V, where ∇V is the outward normal of V. Denote by $\mathbf{F}^-$ the field immediately on the interior of V and by $\mathbf{F}^+$ the field on the immediate exterior. From the property of contracting



(a) (b)

Fig. 3. (a) Proper choice of σ_s results in the black region being transformed into a zone of attraction consisting of a single region of contracting flow. (b) σ_s has been selected too large causing the zone of attraction to contain several regions of contracting flow.



(a) (b)

Fig. 4. (a) Binary image containing square black regions. (b) A plot of the change in vector direction with σ_s is given for a pixel belonging to the square pointed to by A. The three main minima correspond to the spatially stable points. The pixel belongs to a different zone of attraction at each of these points, corresponding to the three perceptual dot clusters to which the pixel belongs.

flow, V satisfies the two relations

$$\nabla V \cdot \mathbf{F}^- \leq 0,\ \nabla V \cdot \mathbf{F}^+ \geq 0 \tag{9}$$

since every point on a boundary curve separates at least two areas of contracting flow. Fig. 3(a) shows $\mathbf{F}^-$ and $\mathbf{F}^+$ vectors for a black region on a white background. Other values of (σ_g, σ_s) may result in contracting flow over a set of disconnected regions [see Fig. 3(b)]. The term *zone of attraction* is used to denote the image space characterized by contracting flow, regardless of whether or not it consists of one or more regions.

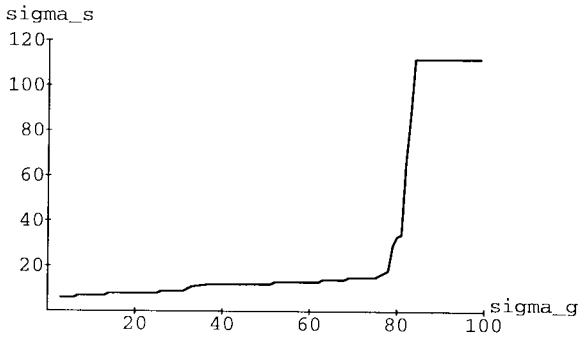
B. Determining Values for σ_s

For a given σ_g, at each pixel (x_0, y_0) within some region R, $\sigma_s(x_0, y_0)$ needs to be chosen such that R corresponds to a region of contracting flow in $\mathbf{F}$. In general, a continuous range of values of $\sigma_s(x_0, y_0)$, denoted $[\sigma_s^-, \sigma_s^+]$, yields contracting flow. This range is determined by examining the behavior of $\mathbf{F}_{\sigma_g}(x_0, y_0)$ as $\sigma_s(x_0, y_0)$ is varied. $\|\mathbf{F}_{\sigma_g}(x_0, y_0)\|$ is small if the pixel intensities are uniformly distributed where $d_s(\cdot)$ is nonzero because the pairwise vector components in the vector sum computed at (x_0, y_0) by (8) tend to cancel out. As the intensity distribution becomes more asymmetric, $\|\mathbf{F}_{\sigma_g}(x_0, y_0)\|$ becomes larger since the pairwise vector components no longer cancel. For every σ_g, a pixel (x_0, y_0) belongs to some region of attraction. If $\sigma_s(x_0, y_0)$ is small enough so that the spatial support of $d_s(\cdot)$ does not extend beyond the region, then $\|\mathbf{F}_{\sigma_g}(x_0, y_0)\|$ is small and the vector direction very sensitive to noise. If $\sigma_s(x_0, y_0)$ is increased so that the spatial support of $d_s(\cdot)$ extends beyond the region, $\|\mathbf{F}_{\sigma_g}(x_0, y_0)\|$ becomes larger and the vector direction becomes stable approximately orthogonal to the nearest part of the region boundary. When this occurs, the pixel has been properly scaled.

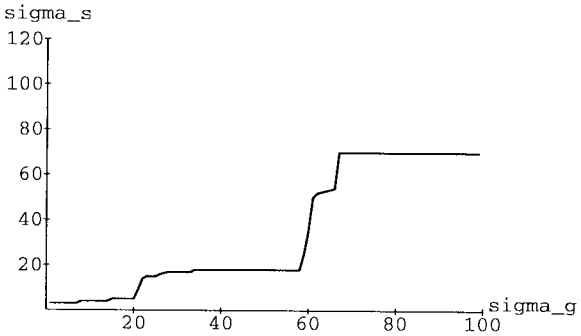
Define

$$\sigma_s^- = \min(\sigma_s) \text{ such that } \|\mathbf{F}_{\sigma_g}(x_0, y_0)\| \geq T(\sigma_g, \sigma_s). \tag{10}$$

(a)



(b)



(c)

Fig. 5. (a) Image of people at a 3-D movie. (b) and (c) show plots of σ_s versus σ_g for the two pixels located under the leftmost and rightmost, respectively, crosshairs $(+)$ overlaid onto the image. For (b), the pixel belongs to two regions of attraction: the woman's face before and after it merges into the background. For (c), the pixel belongs to three regions of attraction: the man's shirt, the shirt merged into the rest of the man, and the man merged into the background. Transitions between regions of attraction are clearly indicated by discontinuities in the value of σ_s.

The value of T determines the smallest feature size that is captured in the field structure. Explicit relationships between structure scale and size can be created by varying T with scale. For example, making T an increasing function of scale increases the effective minimum structure size as scale becomes coarser. Such issues are not addressed here, however. For the purposes of this paper, we set $T = 4$, which is small enough to allow structures of any arbitrary size and shape to be identified except those which consist of only a couple of pixels, and, hence, are heavily aliased. In general, T represents the minimum value for $||\mathbf{F}_{\sigma_g}(x_0, y_0)||$ which signifies the existence of an edge, and, when $||\mathbf{F}_{\sigma_g}(x_0, y_0)|| = T$, an edge exists roughly within a distance of σ_s^- from the pixel. If the pixel is near the region boundary, σ_s^- is small, and σ_s^- is approximately equal to the radius of the region if the pixel is near the region center.

The value of σ_s^+ corresponds to the largest $\sigma_s(x_0, y_0)$, which does not result in overscaling, i.e., the pixel becoming attracted to a disconnected zone of attraction. This situation is detected by examining the behavior of $\mathbf{F}_{\sigma_g}(x_0, y_0)$ as $\sigma_s(x_0, y_0)$ is increased beyond σ_s^-. As this occurs, the pixel initially belongs to the connected zone of attraction. The vector direction tends to change very slowly as long as this is true. However, when $\sigma_s(x_0, y_0)$ becomes large enough, a transition occurs and the pixel now belongs to a different zone of attraction. The vector direction changes much more rapidly during such a transition. As $\sigma_s(x_0, y_0)$ is further increased, the same cycle of behavior may recur: an interval of little change in the vector direction followed by rapid change during transition. We term the state of little change as characterized by *spatial stability*. For each zone of attraction to which the pixel (x_0, y_0) belongs, the value of $\sigma_s(x_0, y_0)$ for which the vector direction changes the least is called a *spatially stable point*. Fig. 4(a) shows a binary image containing squares composed of nine pixels. A plot of $\left|\frac{d}{d\sigma_s}\theta(\sigma_s)\right|$ is given in Fig. 4(b) for a pixel belonging to the square pointed to by A, where $\theta(\sigma_s)$ is defined as the angle in radians of $\mathbf{F}_{\sigma_g}(x_0, y_0)$ from some reference direction. The stability points of the graph occur at $\sigma_s = \{2, 9, 48\}$.

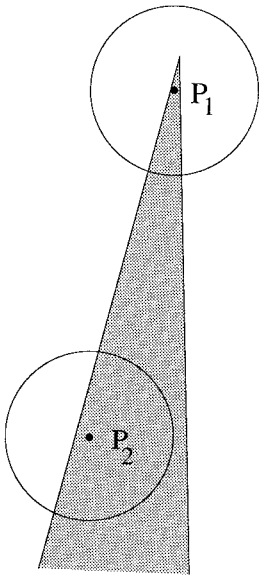


Fig. 6. Demonstration of the increase in the interval of σ_g for which structures transition due to the use of functions for $d_g(\cdot)$ other than a box-car. Consider the transform computed at points P_1 and P_2 within a uniform triangular region, and using the same value of σ_s, represented by the radius of the drawn circles. If $d_g(\cdot)$ is box-car, both points transition to the background at the same value of σ_g, whereas if $d_g(\cdot)$ is Gaussian, P_1 transitions before P_2, resulting in a finite transition interval for the region.

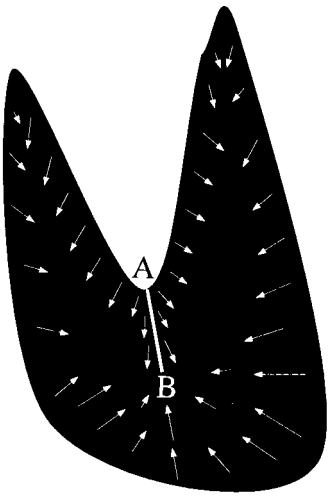


Fig. 7. The presence of a concavity causes vectors to diverge across the line segment formed by joining the concavity center on the region boundary (Point A) to the nearest point on the medial axis of the region (Point B).

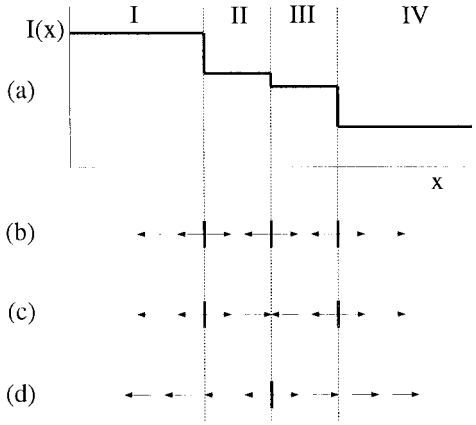


Fig. 8. (a) 1-D signal with four levels. (b)–(d) Edges and some field vectors are shown for three different scales: (b) at fine scale, four regions are detected; (c) at a coarser scale, regions II and III merge together; (d) at an even coarser scale, the edge separating regions II and III has reappeared, resulting in one region consisting of regions I and II and another consisting of regions III and IV.

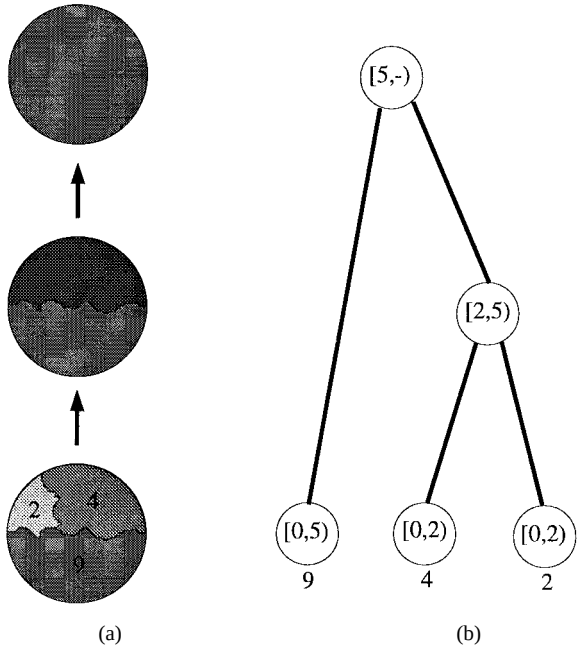


Fig. 9. (a) Segmented regions at different scales, with the intensities of the finest scale regions labeled. (b) Tree corresponding to this segmentation. Each node corresponds to a different region, and the range of σ_g for which each region is present is shown within each node.

These are the locations of the three largest local minima in the graph. Each minimum corresponds to a different zone of attraction for this pixel. These zones of attraction correspond to the three clusters in the image to which one perceives the pixel A belonging. The cluster containing all pixels in the image also has a stability point, but this point occurs for a larger σ_s than shown in Fig. 4(b).

To prevent overscaling, the upper bound, σ_s^+, is defined as the value of the first spatially stable point, i.e.,

$$\sigma_s^+ = \min(\sigma_s) \text{ such that } \frac{d}{d\sigma_s}\left|\frac{d}{d\sigma_s}\theta(\sigma_s)\right| \geq 0 \qquad (11)$$

for $\sigma_s \geq \sigma_s^-$. Selecting $\sigma_s(x_0, y_0) \in [\sigma_s^-, \sigma_s^+]$ for each pixel yields an $\mathbf{F}$ with every force vector belonging to a region of attraction.

C. Behavior of $\mathbf{F}$ with σ_g

In this section, the effect of varying σ_g on $\mathbf{F}$ is examined first at the level of individual pixels, and then for entire structures. For a given value of σ_g, each $\sigma_s(x_0, y_0)$ is chosen properly as detailed in the preceding section.

As σ_g varies, a pixel initially belongs to the same region of attraction for some range of values, but at some point it makes a transition and becomes a part of another region of attraction.

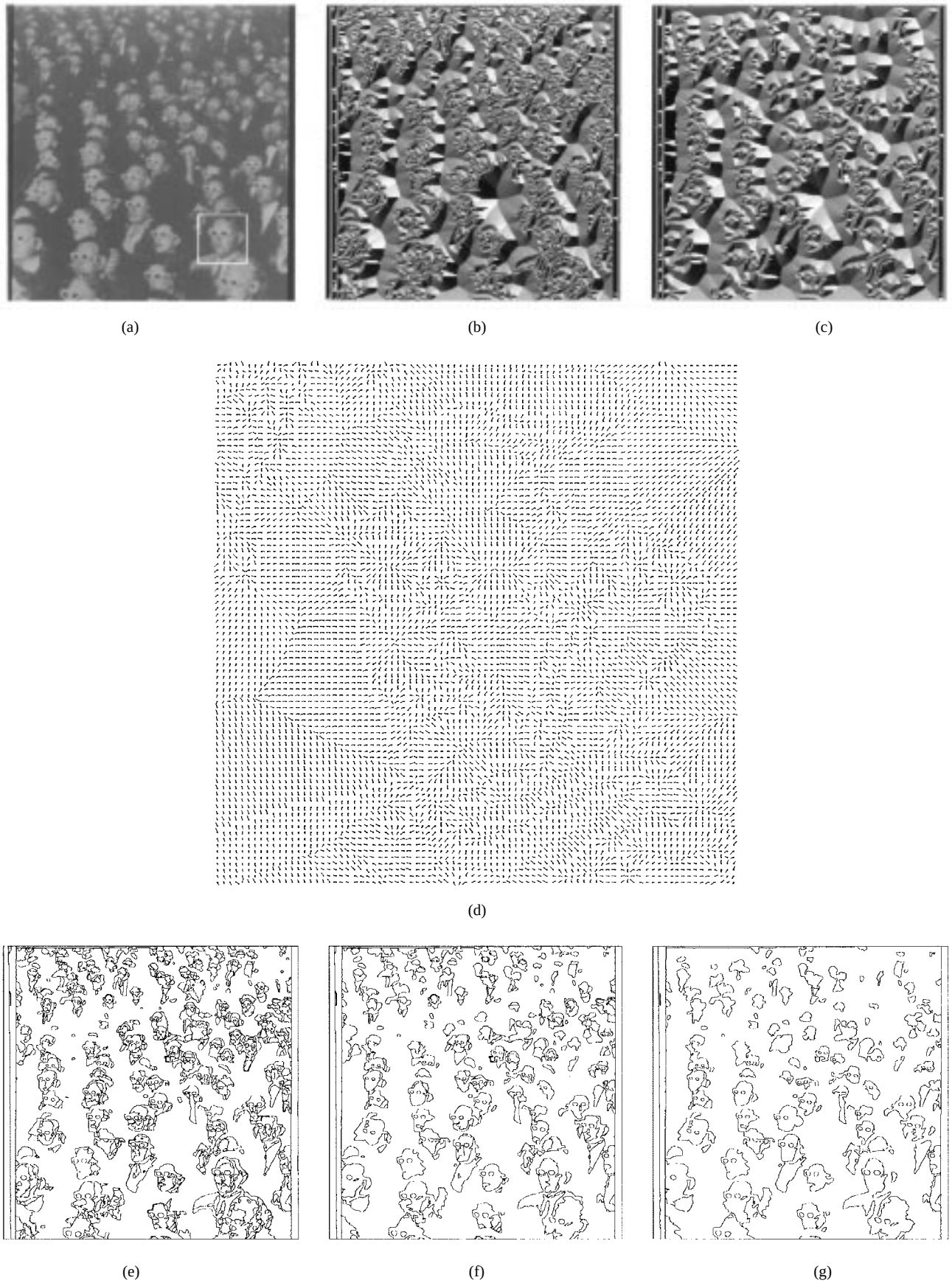


Fig. 10. (a) People at a 3-D movie. (b) Force vectors present within $\mathbf{F}_5$. The vector directions are intensity coded so that the brightness is proportional to the clockwise angle of the force vector from the positive x-axis. Intensity discontinuites correspond to region boundaries and skeletons, although some discontinuities are artifacts of intensity based (linear) coding of the cyclic direction values. (c) Same as (b), but for $\mathbf{F}_{39}$. (d) Analogous to (b), but with the force vectors shown as line segments. The length of the line segment represents the vector magnitude, and the vector tail is indicated by a small square. (e)–(g) Region boundaries present at, respectively, $\sigma_g = 5, 21, 39$.

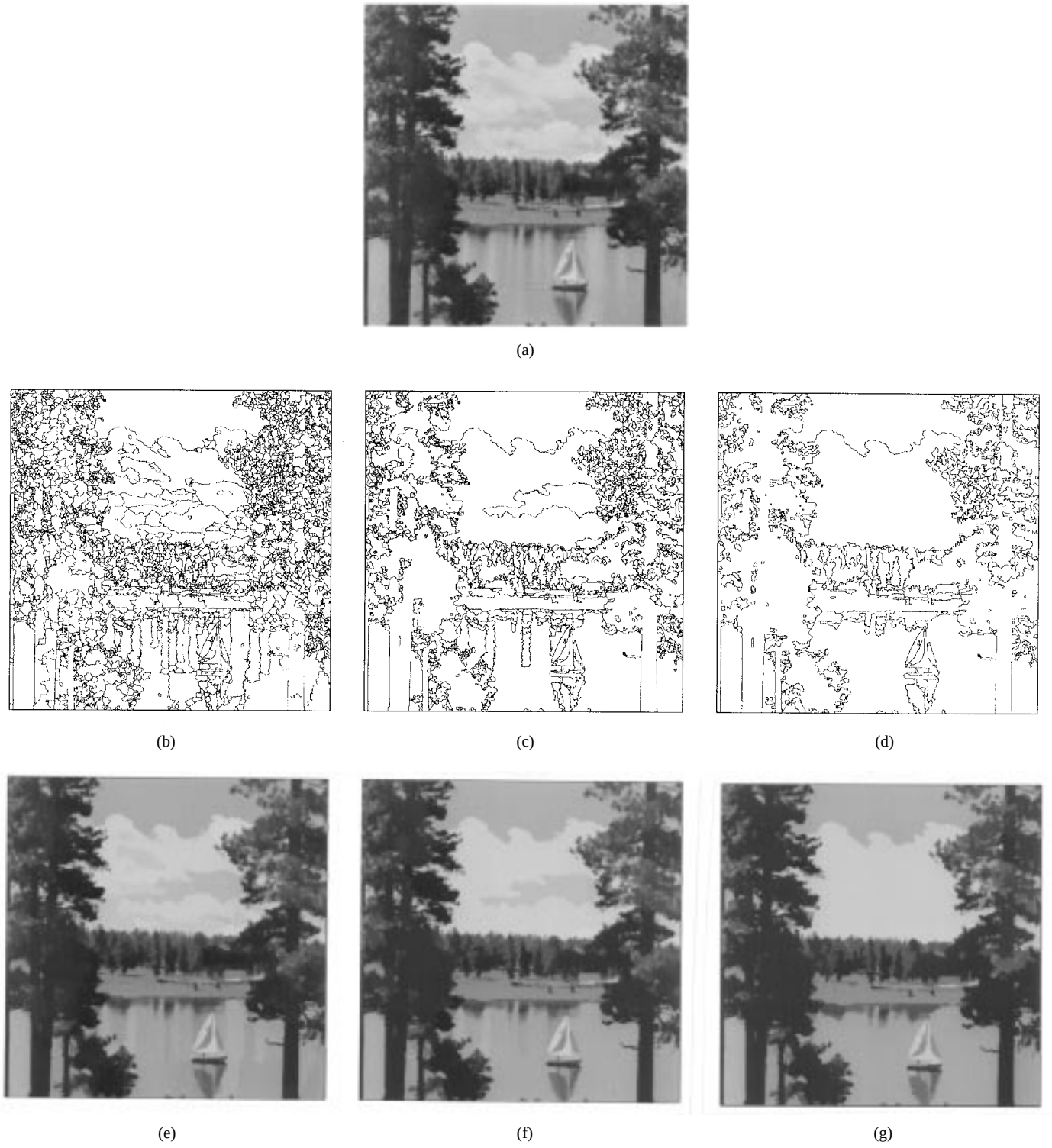


Fig. 11. (a) Sailboat on a lake. (b–d) Region boundaries present at, respectively, $\sigma_g = 5, 10, 26$. (e)–(g) Same as (b)–(d), but with regions displayed by their average gray level.

These transitions are readily discernible by examining the σ_g–σ_s plot of a pixel. As σ_g increases, the region of attraction to which a pixel belongs increases in size as pixels previously outside the region (because they were too dissimilar), gradually merge into the region. Since σ_s^- represents the approximate distance between the pixel and the nearest boundary of its region of attraction, σ_s^- is nondecreasing as σ_g increases. While the pixel belongs to the same region of attraction, σ_s^- increases slowly as σ_g increases. The nearest boundary point of the next (coarser) region of attraction of the pixel is farther away than the nearest boundary point of the present region of attraction, resulting in a discontinuity in the value of σ_s at the σ_g where the transition occurs. Examples of two typical σ_g–σ_s plot are given in Fig. 5.

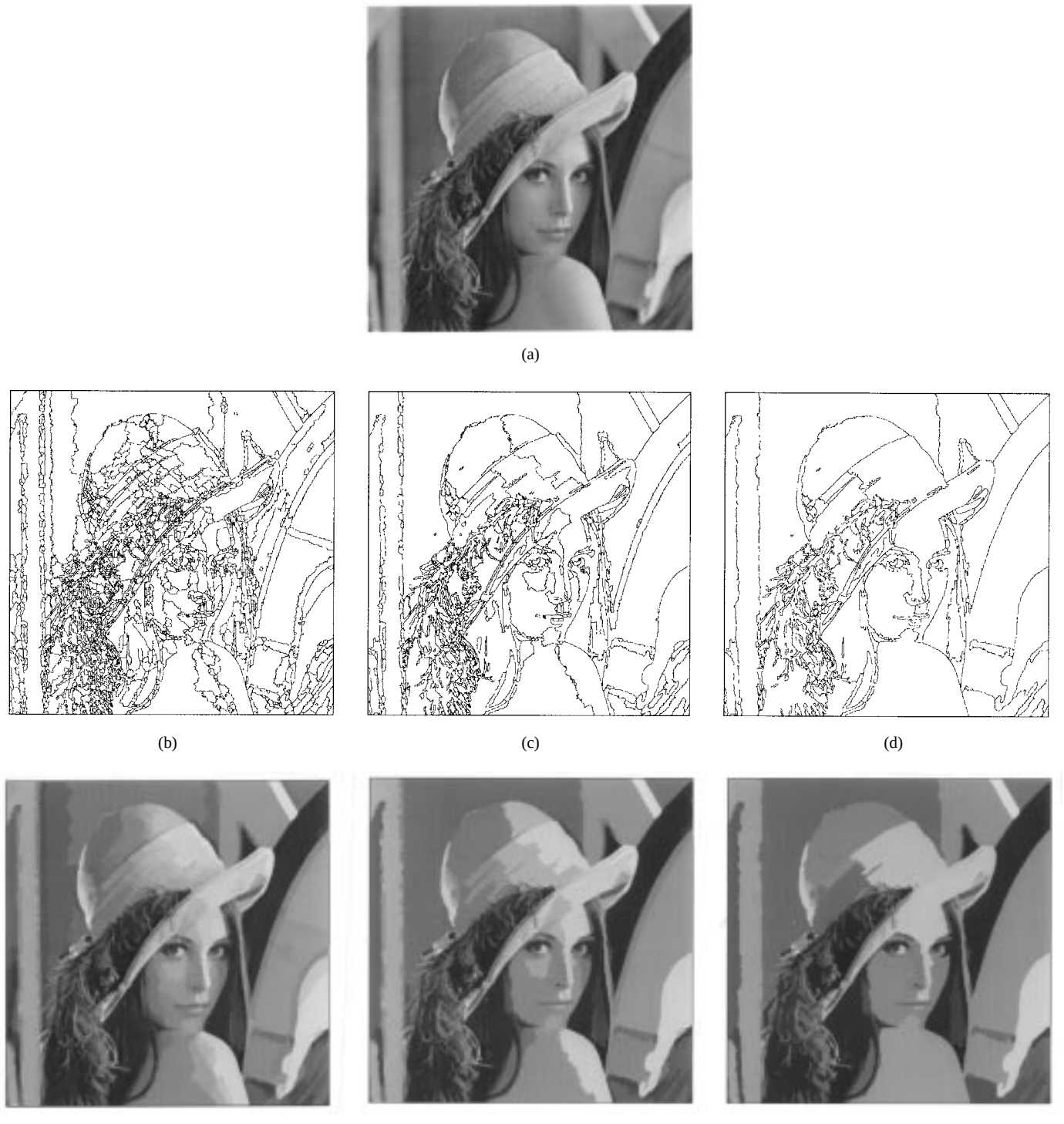


Fig. 12. (a) Lena. (b)–(d) Region boundaries present at, respectively, $\sigma_g = 7, 10, 21$. (e)–(g) Same as (b)–(d), but with regions displayed by their average gray level.

Fig. 5(a) shows an image of people at a three-dimensional (3-D) movie. Plots of σ_s vs. σ_g are shown in 5(b) and 5(c) for the pixels located at the centers of the leftmost and rightmost crosshairs $(+)$ overlaid onto the image, respectively. The pixel for Fig. 5(b) belongs to two relevant regions of attraction. The first consists of the woman's face, and the second occurs as the face merges into the background region. The pixel for Fig. 5(c) belongs to three relevant regions of attraction. The first consists of the man's shirt, the second occurs when the shirt merges with the rest of the man, and the third occurs when the man merges into the background. For both 5(b) and (c), transitions between regions of attraction are clearly indicated by discontinuities in the value of σ_s.

From examination of the σ_g–σ_s plot of a pixel, both the number of structures to which the pixel belongs as well as the range of σ_g for which the pixel belongs to each structure

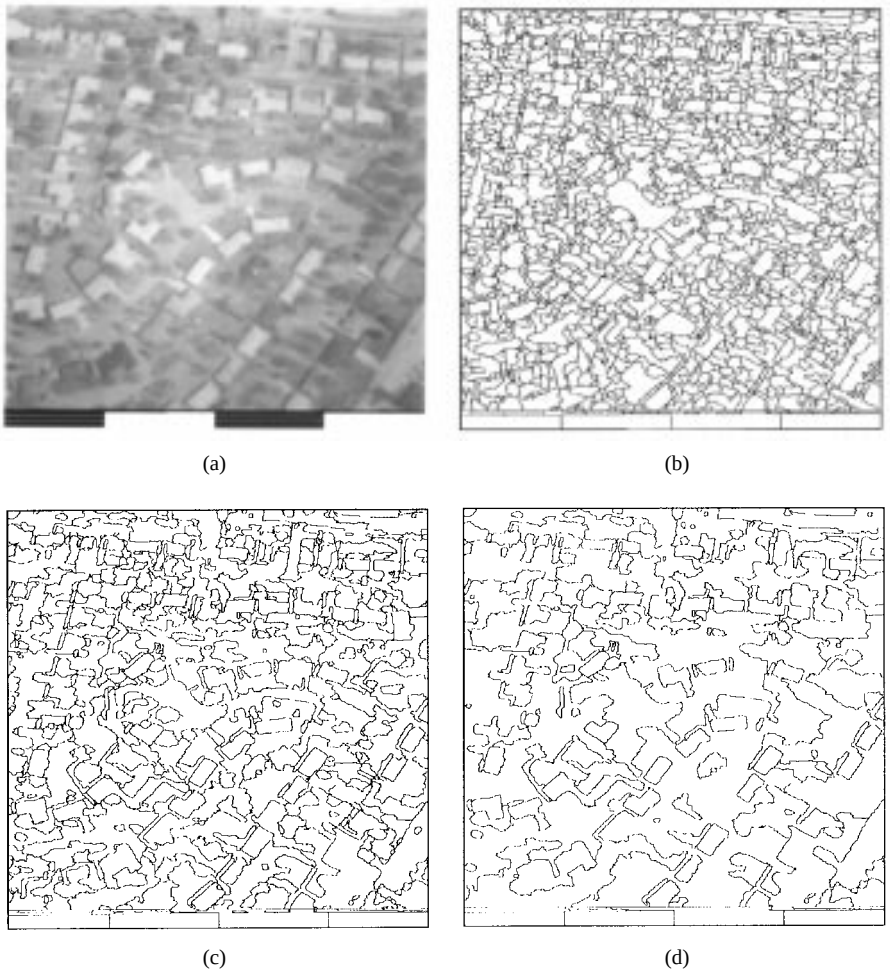


Fig. 13. (a) Aerial scene. (b)–(d) Region boundaries present at, respectively, $\sigma_g = 7, 13, 26$.

can be determined. However, in and of itself, this does not provide enough information to do structure identification. An understanding of the behavior of entire structures with respect to σ_g is also required. As σ_g varies, some structures coalesce while others disappear. Because image structures generally have nonzero variance, regions of attraction form and disappear over some range of σ_g. Hence, for any given structure, as σ_g is increased, the structure gradually forms, then remains relatively stable for some interval, and then gradually disappears as its constituent pixels form into other, coarser structures. The identification of these structures from $\mathbf{F}$ is addressed in Section III.

D. Homogeneity Distance Function Selection

The purpose of this section is to motivate briefly the use of a box-car function for $d_g(\cdot)$. Consider Fig. 6, which contains a uniform triangular region having contrast C with the background. The transform is computed at points P_1 and P_2 within the region for the same value of σ_s, represented by the radius of the drawn circles. If a box-car is used for $d_g(\cdot)$, then both points merge into the background region at $\sigma_g = C$. However, if $d_g(\cdot)$ is Gaussian, P_1 merges into the background earlier than P_2 because the proportion of area within the circle centered at P_1 occupied by the region is less than the proportion within the circle at P_2. The net effect is that the range of σ_g for which the region merges into the background is increased. This result also holds under the addition of zero-mean noise. It is easily shown that the form of $d_g(\cdot)$ which yields the shortest transition interval is the box-car function. A longer transition interval reduces the interval of σ_g for which a structure is stable, and may even eliminate it entirely. Only structures which are marginally stable (using the stable measure presented in Section III) are negatively effected by this, which explains why the use of a box-car for $d_g(\cdot)$ vis-a-vis other functions results in only a slight improvement in segmentation quality.

III. Identifying Regions of Attraction Within $\mathbf{F}$

This section presents a method for identifying the regions of attraction within $\mathbf{F}$ that correspond to relevant image structure. In order to illustrate the integration of edge- and region- based representations within $\mathbf{F}$, Section III-A describes a purely edge-based structure extraction method which gives closed contours, and, hence, regions. In Section III-B, a stability criterion is used to identify which of these structures is perceptually valid. Finally, the issue of a compact and efficient

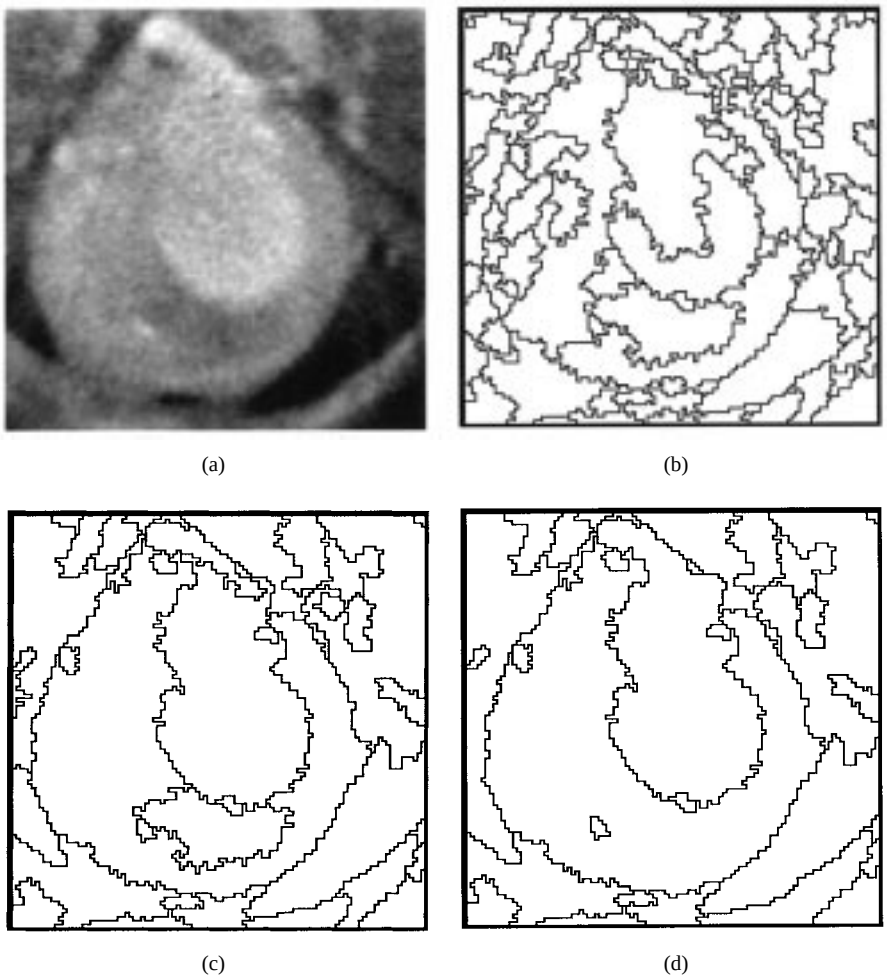


Fig. 14. (a) 2-D slice from a CT scan of a dog heart. (b)–(d) Region boundaries present at, respectively, $\sigma_g = 5, 13, 21$.

form for representing the segmentation is addressed in Section III-C.

A. Identifying Region Boundaries

Since (9) holds along each boundary curve, the vectors along each boundary curve diverge from one another. For the interval of σ_g for which a structure is fully formed within $\mathbf{F}$, these boundaries form closed contours. Thus, structure identification consists of searching $\mathbf{F}$ for all sets of boundaries that form closed contours. Toward this end, a local measure for a region boundary needs to be defined. In a continuous space, one can show that the only place within $\mathbf{F}$ where vectors diverge from one another is at region boundaries, and that this divergence is always π, regardless of the boundary geometry [16]. However, discretization may reduce this divergence somewhat, and, for regions containing concavities, causes vectors to diverge from one another across the line segment formed by joining the center of the concavity on the region boundary to the nearest point on the medial axis of the region, as illustrated in Fig. 7. In a continuous space, the angle between two vectors across segment $\overline{AB}$ approaches zero in the limit as the vectors are chosen arbitrarily close to $\overline{AB}$. For a digital image, the finite spatial resolution causes a nonzero angle of at most ϵ to exist between the vectors across the line segment. The value of ϵ decreases as σ_s, and, hence, T, increases. For $T = 1$, it is easily shown that $\epsilon = 2\arctan(1/(\sqrt{2}+1)) \approx \pi/4$, and for $T = 4$, $\epsilon \approx \pi/7$. The angular distributions of the divergence present at region boundaries and concavities never overlap, so can be used as a threshold for classifying these two cases. Region boundaries are identified by comparing each vector with its eight nearest neighbors. For example, to determine whether or not a region boundary exists between (x_0, y_0) and (x_0+1, y_0) at $\sigma_g = \sigma_{g0}$, the following test is used:

$$\begin{aligned} &\text{If } \mathbf{F}_{\sigma_{go}}(x_0,y_0)_x \le 0 \ \&\ \mathbf{F}_{\sigma_{go}}(x_0+1,y_0)_x \ge 0 \\ &\quad \&\ \frac{\mathbf{F}_{\sigma_{go}}(x_0,y_0)\cdot\mathbf{F}_{\sigma_{go}}(x_0+1,y_0)}{\|\mathbf{F}_{\sigma_{go}}(x_0,y_0)\|\|\mathbf{F}_{\sigma_{go}}(x_0+1,y_0)\|} < \cos\epsilon \\ &\text{then } (x_0,y_0)\ \&\ (x_0+1,y_0) \text{ are boundary pts.} \end{aligned} \tag{12}$$

Analogous tests are used for the other neighboring vector pairs.

B. Structure Stability

Not all of the structures identified in the previous section are perceptually valid. This is because σ_g reflects the relative homogeneity within a structure, but not the relative contrast

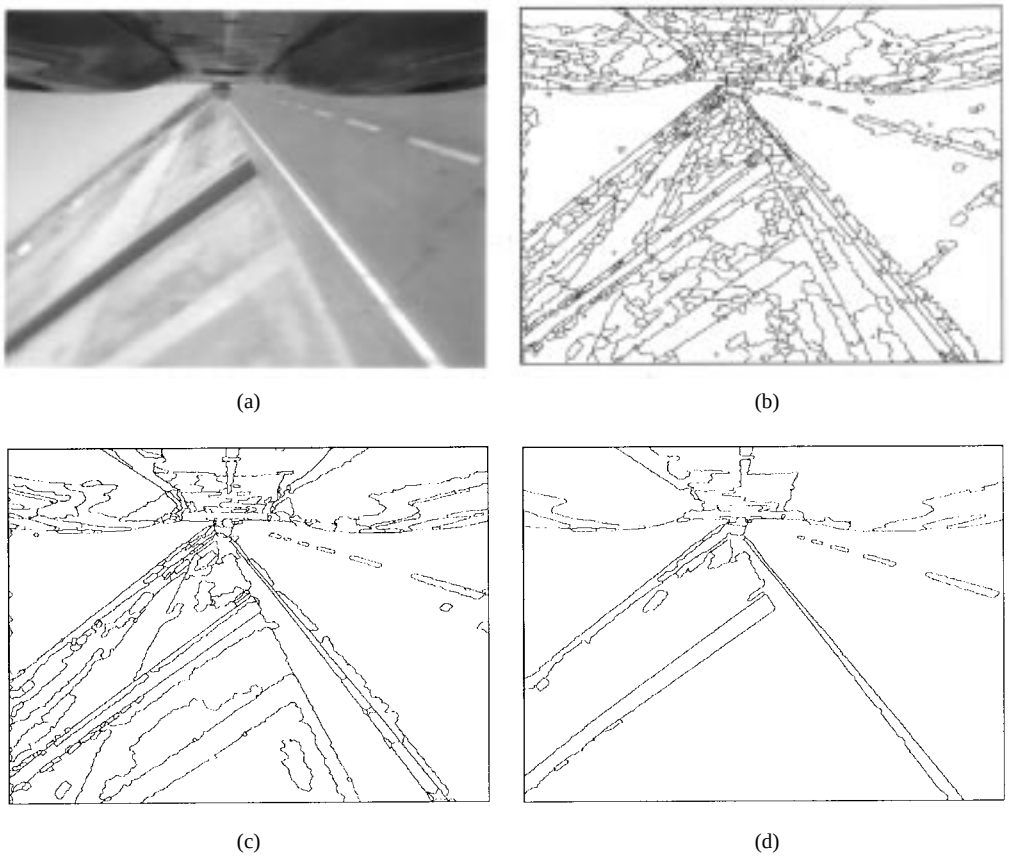


Fig. 15. (a) Aerial image taken from a camera mounted beneath an airplane. (b)–(d) Region boundaries present at, respectively, $\sigma_g = 5, 10, 26$.

of the structure with neighboring structures. The latter is reflected in the extent of a structure in $\mathbf{F}$, or σ_g—*lifetime*. This is measured from the point at which a structure is half-formed to when it has half-disappeared. This results in a pixel being assigned to a structure for every value of σ_g. Exact computation of the σ_g—lifetime of a structure requires examination of the $\sigma_g - \sigma_s$ plot of each pixel within the structure for the discontinuities in σ_s, which indicate transitions of the pixel into and out of the structure. This could be done but is unnecessarily expensive computationally. A simpler approach is to approximate the lifetime using only the structure boundary pixels, since a transition at a boundary pixel is reflected in an edge appearing or disappearing, and this edge information is already computed in Section III-A.

A structure is considered stable with σ_g, and, hence, perceptually relevant, if it obeys the following "octave" rule:

A structure with σ_g—lifetime $[g_1, g_2]$, is stable iff $g_2 > 2g_1$.

The octave rule essentially requires the homogeneity variation within a structure to be less than the relative contrast between the structure and neighboring structures. This results in structures that not only become less homogeneous as scale becomes coarser, but also have increasing contrast with neighboring regions. The octave rule yields good results in practice, but is not based on psychophysical criteria. Experiments are planned in which people subjectively evaluate segmentations produced by different rules in order to determine a perceptually valid criterion for identifying the regions of attraction which correspond to relevant structures.

C. Segmentation Representation

This section addresses the issue of how best to represent the segmentation obtained in Sections III-A and III-B. The segmentation consists of a list of relevant structures along with their σ_g—lifetimes and constituent pixels. Because this segmentation retains the geometric fidelity of structure boundaries at all scales, the segmentation often can be represented as a tree, i.e., every structure consists of a set union of structures present at any finer scale. In certain isolated situations, however, this does not hold true, as is demonstrated in Fig. 8. A 1-D signal having four levels is shown in Fig. 8(a). At some initial scale, the resulting field is shown in Fig. 8(b) along with the three edges detected. These edges separate the signal into four regions, numbered as shown. At some coarser scale, Fig. 8(c), the edge separating regions II and III disappears and these regions merge. However, at a yet coarser scale, Fig. 8(d), this edge reappears and the other two edges disappear.

Because a tree is such a convenient representation, the segmentation is post-processed to explicitly force it into this form. This is accomplished by projecting structure boundaries downward to all finer scales. This prevents certain regions, such as II and III in Fig. 8, from merging together. The segmentation is now representable as a tree, such as the

one in Fig. 9. Each node in the tree corresponds to a segmented structure. Information about a structure which can be stored at its associated node includes average intensity, texture statistics, σ_g—lifetime, boundary chain-code, moments, area, etc. The base of the tree contains the finest scale structures, and incrementally coarser structures are represented as the tree is traversed upwards. The tree representation has many useful aspects, including efficient coarse-fine access to the image structures as well as compact storage of the segmentation.

IV. Experimental Results

This section presents the results of the described image segmentation method for a variety of real images. For each image in Figs. 10–15, the tree representing the multiscale segmentation is computed. The segmentation is visualized by displaying the structure boundaries, as well as the average gray-scale of the structures for Figs. 11 and 12, for each structure present within the tree at a particular value of σ_g. For each image, three different values of σ were selected so that a wide variety of different structures can be seen. For Fig. 10, in addition to the segmentation results shown in Fig. 10(e)–(g), the directions of the force vectors that comprise $\mathbf{F}_5$ and $\mathbf{F}_{39}$ are displayed in Fig. 10(b) and (c), respectively. The vector directions are intensity coded so that the brightness is proportional to the clockwise angle the force vector makes from the positive x-axis. Since structure boundaries and skeletons are represented within $\mathbf{F}$ by diverging and converging, respectively, field vectors with phase differences of approximately π, they appear as sharp intensity discontinuities. Some intensity discontinuities, however, are artifacts due to the branch cut present along the positive x-axis, and should be ignored. Also, the force vectors of $\mathbf{F}_5$ that corresponding to the area within the white box in Fig. 10(a) are displayed as line segments in Fig. 10(d). The length of each line segment represents the vector magnitude, and the vector tail is indicated by a small square. Fig. 10(d) allows one to inspect more closely the structures encoded within $\mathbf{F}_5$.

Figs. 10–15 were selected to demonstrate the performance of this algorithm in a variety of different situations, for example, noise (Fig. 14); shading (Fig. 12); significant multiscale structure (Figs. 10, 11, 14, and 15); sharp and diffuse edges in close proximity (Figs. 12–15); and high-curvature structure boundaries (Figs. 10, 11). In all cases, the identified structures closely correspond to human perception of relevant structure, and the structure boundaries closely align with the actual boundaries in the image. Finally, the amount of computation required by this algorithm is reasonable. About 10 s is currently required to segment a 512×512 image on a Sun SPARC20 Workstation..

V. Conclusion

The application of a new nonlinear transform introduced in [16]–[18] to the problem of image segmentation has been described. The identified regions correspond to perceptually valid image structure, and the identified region boundaries align closely with the actual boundaries of the structures, regardless of the scale of the structures. All parameters of the transform are selected automatically, eliminating the need for any user-specified parameters. Automatic selection of both the homogeneity and spatial scales avoids the need to make restrictive *a priori* assumptions about either the geometric or homogeneity characteristics of the structure. A tree is constructed that represents all of the structure in a given image and the range of homogeneity scale for which each structure is present. This approach to structure detection is distinguished from previous methods in several respects. First, scale is formulated in a manner which naturally represents image structure. Second, the processes of scale selection and structure detection are integrated. In addition, a unification between region- and edge- detection is achieved in the transformed domain. Finally, structure of arbitrary geometry can be detected without any smoothing of the structure boundaries, even at coarse scales.

References

[1] R. Ohlander, K. Price, and D. Reddy, "Picture segmentation using a recursive region splitting method," *Comput. Graph. Image Processing*, vol. 8, no. 3, pp. 313–333, 1978.

[2] R. Haralick and L. Shapiro, "Survey: Image segmentation techniques," *Comput. Vis. Graph. Image Processing*, vol. 29, no. 1, pp. 100–132, 1985.

[3] S. Zucker, "Survey region growing: Childhood and adolescence," *Comput. Graph. Image Processing*, vol. 5, no. 4, pp. 382–399, 1976.

[4] O. Monga, "An optimal region growing algorithm for image segmentation," *Int. J. Pattern Recog. Artif. Intell.*, vol. 1, no. 4, pp. 351–375, 1987.

[5] S. Horowitz and T. Pavlidis, "Picture segmentation by a directed split-and-merge procedure," in *Proc. 2nd Int. Joint Conf. Pattern Recognition*, 1974, pp. 424–433.

[6] F. Meyer and S. Beucher, "Morphological segmentation," *J. Vis. Commun. Image Represent.*, vol. 1, no. 1, pp. 21–46, 1990.

[7] A. Nazif and M. Levine, "Low level image segmentation: An expert system," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, no. 5, pp. 555–577, 1984.

[8] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, no. 11, pp. 721–741, 1984.

[9] T. Lindeberg, *Scale-Space Theory in Computer Vision*. Boston, MA: Kluwer, 1994.

[10] R. Whitaker and S. Pizer, "Geometry-based image segmentation using anisotropic diffusion," in *Shape in Picture: Mathematical Descriptions of Shape in Grey-Level Images*. New York: Springer-Verlag, 1992, pp. 641–650.

[11] P. Salembier, "Morphological multiscale segmentation of images," in *Proc. SPIE Visual Communications in Image Processing'92*, vol. 1818, no. 3, pp. 620–631.

[12] D. Mumford and J. Shah, "Boundary detection by minimizing functionals, I," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1985, pp. 22–26.

[13] A. Witkin, "Scale space filtering," in *Proc. Eighth Int. Joint Conf. on Artificial Intelligence*, Karlsruhe, Germany, Aug. 1983, pp. 1019–1022.

[14] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, no. 7, pp. 629–639, 1990.

[15] R. Boomgaard and A. Smeulders, "Toward a morphological scale-space theory," in *Shape in Picture: Mathematical Descriptions of Shape in Grey-Level Images*. New York: Springer-Verlag, 1992, pp. 631–640.

[16] N. Ahuja, "A transform for detection of multiscale image structure," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition'93*, New York, pp. 780–781.

[17] N. Ahuja, "A transform for detection of multiscale image structure," in *Proc. Image Understanding Workshop*, Washington, DC, Apr. 1993, pp. 893–903.

[18] ——, "A transform for multiscale image segmentation by integrated edge and region detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 1211–1235, Dec. 1996.

Mark Tabb received the B.S. in electrical engineering from Cornell University, Ithaca, NY, in 1991. He received the M.S. and the Ph.D. degrees, in 1993 and 1996, respectively, from the University of Illinois, Urbana-Champaign, both in electrical and computer engineering.

From 1991 to 1996, he was a Research Assistant at the Beckman Institute, University of Illinois. His research interests include image processing, computer vision, and special-purpose VLSI architectures.

Narendra Ahuja (M'79–SM'85–F'92) received the B.E. degree with honors in electronics engineering from the Birla Institute of Technology and Science, Pilani, India, in 1972, the M.E. degree with distinction in electrical communication engineering from the Indian Institute of Science, Bangalore, in 1974, and the Ph.D. degree in computer science from the University of Maryland, College Park, in 1979.

From 1974 to 1975, he was Scientific Officer in the Department of Electronics, Government of India, New Delhi. From 1975 to 1979, he was at the Computer Vision Laboratory, University of Maryland. Since 1979, he has been with the University of Illinois, Urbana-Champaign, where he is currently a Professor in the Department of Electrical and Computer Engineering, the Coordinated Science Laboratory, and the Beckman Institute. His interests are in computer vision, robotics, image processing, image synthesis, sensors, and parallel algorithms. His current research emphasizes integrated use of multiple image sources of scene information to construct 3-D descriptions of scenes; the use of integrated image analysis for realistic image synthesis; parallel architectures and algorithms and special sensors for computer vision; and use of the results of image analysis for a variety of applications including visual communication, image manipulation, video retrieval, robotics, and scene navigation.

Dr. Ahuja was a Beckman Associate in the University of Illinois Center for Advanced Study for 1990–1991. He received University Scholar Award (1985), Presidential Young Investigator Award (1984), National Scholarship (1967–72), and President's Merit Award (1966). He has co-authored *Pattern Models* (New York: Wiley, 1983) and *Motion and Structure from Image Sequences* (New York: Springer-Verlag, 1992) and co-edited *Advances in Image Understanding* (IEEE Press, 1996). He is fellow of the American Association for Artificial Intelligence, the International Association for Pattern Recognition, the Association for Computing Machinery, the American Association for the Advancement of Science, and the International Society for Optical Engineering. He is a member of the Optical Society of America. He is on the editorial boards of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE; *Computer Vision, Graphics, and Image Processing*; *Journal of Mathematical Imaging and Vision*; and *Journal of Information Science and Technology*. He is a guest co-editor of the *Artificial Intelligence Journal* special issue on vision.