Mid-Level Vision and Recognition of Non-Rigid Objects

by

J. Brian Subirana-Vilanova

Master of Science in Management, Massachusetts Institute of Technology
(1993)

Llicenciat en Informàtica, Universitat Politécnica de Catalunya
(1987)

Submitted to the
Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the
Massachusetts Institute of Technology
October 29, 1993

Signature of Author______________________________

J. Brian Subirana-Vilanova
Department of Electrical Engineering and Computer Science
October 29, 1993

Certified by______________________________

Professor Shimon Ullman
Thesis Co-Supervisor

Certified by______________________________

Professor Tomaso Poggio
Thesis Co-Supervisor

Accepted by______________________________

Professor Campbell Searle
Chairman, Departmental Graduate Committee

Mid-Level Vision and Recognition of Non-Rigid Objects

by

J. Brian Subirana-Vilanova

Submitted to the Department of Electrical Engineering and Computer Science on August 13 1993, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Abstract

In this dissertation I address the problem of visual recognition of non-rigid objects. I introduce the frame alignment approach to recognition and illustrate it in two types of non-rigid objects: contour textures and elongated flexible objects. Frame alignment is based on matching stored models to images and has three stages: first, a "frame curve" and a corresponding object are computed in the image. Second, the object is brought into correspondence with the model by aligning the model axis with the object axis; if the object is not rigid it is "unbent" achieving a canonical description for recognition. Finally, object and model are matched against each other. Rigid and elongated flexible objects are matched using all contour information. Contour textures are matched using filter outputs around the frame curve.

The central contribution of this thesis is Curved Inertia Frames (C.I.F.), a scheme for computing frame curves directly on the image. C.I.F. is the first algorithm which can compute probably global and curved lines, and is based on three novel concepts: first, a definition of curved axis of inertia; second, the use of non-cartisean networks; third, a ridge detector that automatically locates the right scale of objects. The use of the ridge detector enables C.I.F. to perform mid-level tasks without the need of early vision. C.I.F. can also be used in other tasks such as early vision, perceptual organization, computation of focus of attention, and part decomposition

I present evidence against frame alignment in human perception. However, this evidence suggests that frame curves have a role in figure/ground segregation and in fuzzy boundaries, and that their outside/near/top/incoming regions are more salient. These findings agree with a model in which human perception begins by setting a frame of reference (prior to early vision), and proceeds by successive processing of convex structures (or holes).

The schemes presented for contour texture and elongated flexible objects use a comon two-level representation of shape and contour texture which may also be useful to recognize other non-rigid transformations such as holes, rigid objects, articulated objects, and symbolic objects.

Most of the schemes have been tested on the Connection Machine and compared against human perception. Some work on part segmentation and holes is also presented.

Thesis Committee: Prof. W. Eric L. Grimson
Prof. Tomas Lozano-Perez
Prof. Tomaso Poggio
Prof. Whitman Richards
Prof. Shimon Ullman

Preface

When I first came to the M.I.T. A.I. laboratory I wanted to approach "the central problem" in intelligence by developing some sort of "universal theory of recognition," something that was about to "solve A.I.." Soon I realized that this was something beyond my reach. It became clear to me that progress in A.I. would best be achieved by researching a "small intelligence problem." I then became increasingly interested in the study of the relation between human vision, the physical structure of nature, and the anatomy of computer algorithms. This stemmed, mainly, from observing that, in the past, progress in one of these areas often has contributed, or served to guide progress, in another.

The purpose of this dissertation is to elucidate the computational structure of human vision algorithms. Work in vision, before I came to M.I.T., had focussed on early vision and recognition of rigid objects. Very little work had been done in middle-level vision and in the recognition of non-rigid objects. I set myself the goal of investigating middle-level vision through an understanding of the issues involved in the recognition of non-rigid objects. I was hoping that much could be learned from such an excersise. I hope that the contributions presented today in this thesis may serve as bricks in tomorrow's "intelligent" A.I. systems.

*

This thesis deals with both computational and human aspects of vision. More progress in computer vision is essential if we want tomorrow's robots to do unpleasant, repetitive, difficult, or dangerous work in unstructured environments, where change is constant, and where there can be little or none of the careful environmental preparation and control characteristic of today's built-for-robots factory. More progress in human visual perception is essential before we can have a full understanding of the nature of our own visual system. *A priori*, there is no reason why work on these two areas should proceed either in parallel nor simultaneously. However, this thesis, as many other research projects, benefits from the synergy between the two.

Interdisciplinary work of this nature belongs to the field of computational psychology and proposes visual computations with relevant implications in both computer vision and psychology.

*

Any investigation of what is fundamentally a topic in the natural sciences must be shown to be both rigorous and relevant. It must simultaneously satisfy the seemingly incompatible standards of scientific rigor and empirical adequacy. Of the two criteria, relevance is the most important, and judging by historical example also, the most difficult to satisfy.

If an investigation is not rigorous, it may be called into question and the author perhaps asked to account for its lack of rigor. In computer vision, the critic often requires that the author provide particular runs of the algorithms proposed on examples that the critics come across; especially if he had not given enough in the first place, as is often the case. To ensure rigor, the investigation must be based upon a framework that includes a technique for performing the analysis of the claims and corresponding algorithms, and for proving them correct.

But if the investigation is not relevant, it will be ignored entirely, dismissed as an inappropriate (cute at best) hack. In order to ensure relevance, such an interdisciplinary investigation must result in useful vision machines, demonstrate a comprehensive understanding of the natural sciences involved, and provide three warrants.

The first warrant is a link between the human visual system and the task for which the studied algorithms are shown to be useful. This link can be at a psychological level, at a neuroscience level, or at both. In this thesis, the link centers around the notion of frame curves, frame alignment and inside/outside.

The second warrant is a conceptual framework for the investigation, so that the algorithms are used to answer relevant questions. The framework must include a technique for performing an analysis of the algorithms in the domain of one of the natural sciences, visual perception in this thesis. The technique must ensure that the insights of the science are preserved. It must be sufficiently general so that others can extend the investigation, should it prove fruitful to do so.

The third warrant is a contribution to one of the natural sciences, psychology in this thesis. Such a contribution might take the form of a simply-stated mathematical thesis or of a less succinct statement such as a set of new visual illusions prompted by the investigation. Whatever the form of such a contribution, it must be a guide to scientific investigation in the natural science itself. To be relevant, a thesis must make strong predictions that are easily falsified in principle, but repeatedly confirmed in practice. Only under these conditions is it possible to develop confidence in such a thesis. Some of the contributions made here include: results on the relative saliency of the inside of a contour versus that of the outside; and results on the recognition of contour textures, holes, and elongated and flexible objects.

Indeed, the relation between the study of computational models of the human visual system and that of computer vision is very fuzzy. While the first is concerned with how the human brain "sees" the second one is driven by the goal of building machines that can use visual information effectively. From a historical perspective, most relevant discoveries in one of the fields will affect the other.

*

This research builds on and corroborates two fundamental assumptions about the algorithms involved in human vision: first, the algorithms are highly dependent on particular tasks solved by the human brain; and second, the laws of nature are embodied in the algorithms (indeed, part of the intelligence of visual perception is not in the brain but in nature itself). The idea that regularities in nature play a key role in visual perception has been fostered by some. The notion that the task at hand should drive the architecture of robots has been extensively investigated recently; its implications to human perception are less well understood. In Appendix B I discuss some implications of the tasks faced by humans in the nature of the visual system.

One of the apparently-deceiving consequences of the first assumption is that we are bound to understand the computational structure of human vision as a collection of inter-related algorithms. This is not so surprising if we consider that we understand a computer as an inter-related set of subsystems such as CPU, printer, and memory to name but a few, or that we understand nature as a set of relatively independent theories such as thermodynamics, electromagnetism, or classical mechanics. It is clear to me that many interactions between the different subsystems exist. The model presented in Chapter 3 is an example. I have found some of these interactions by investigating, again, the task at hand and not by trying to speculate on very-general-purpose intermediate structures. It is as if the human visual system embodied nature betraying it at the same time (for the sake of the tasks it solves).

*

Many people have contributed to the creation of this thesis:

First and foremost I would like to acknowledge the United States for hosting me as one more of its citizens each time I have chosen to come back; a spirit I hope will last for years to come, ever increasingly.

I can not imagine a better place to research than the M.I.T. Artificial Intelligence Laboratory. I wish I could do another PhD; the thought of having another chance to approach the "universal theory of recognition" reminds me of how much fun I have had in the laboratory.

I was fortunate to meet, shortly after I arrived at the laboratory, the truly exceptional individuals who compose my thesis committee: Eric Grimson, Tomás Lozano-Pérez, Tomaso Poggio, Whitman Richards, and Shimon Ullman. Thanks for your help.

Kah Kay Sung, with whom I developed the ridge detector presented in Chapter 3 [Subirana-Vilanova and Sung 1993] was a source of continuous productive chromatic discussions when the Connection Machine is at its best (around 3am...).

During these 5 years, my work, and this thesis, have also benefited a lot from the interaction with other researchers: Tao Alter, Steve Bagley, David Braunegg, Ronen Basri, Thomas Breuel, David Clemens, Stephano Casadei, Todd Cass, Shimon Edelman, Davi Geiger, Joe Heel, Ellen Hildreth, David Jacobs, Makoto Kosugi, Pam Lipson, Tanveer Mahmood, Jim Mahoney, Jitendra Malik, David Mellinger, Sundar Narasimhan, Pietro Perona, Satyajit Rao, Yoshi Sato, Eric Saund, Amnon Shashua, Pawan Sinha.

Throughout the years I have had a number of outstanding "non-vision" officemates: David Beymer (part-time vision partner too!), Jim Hutchinson, Yangming Xu and Woodward Yang who, aside from a constant source of inspiration, have become my best friends during the many occasions in which thesis avoidance sounded like the right thing to do. (No more thesis avoidance folks! A move into ”work avoidance” appears imminent...).

I thank the Visual Perception Laboratory of the NTT Human Interface Laboratories in Yokosuka, Japan where I developed and tested some of the ideas in Appendix B; The Electronic Documents Laboratory at Xerox Palo Alto Research Center in Palo Alto, California where I developed the Connection Machine implementation of the ideas in Chapters 2, 3, and 4; The Weizmann Institute of Science in Rehovot, Israel for inviting me to work on my thesis with Prof. Shimon Ullman and his students.

The M.I.T. Sloan School of Management discovered and then fostered my interests and potential in the management of technology and beyond.

Carlos Angulo, Ramon López de Mántaras, Ignasi Juvells, Ton Sales, Josep Amat, Lluis Puigjaner, Teresa Riera, Didac Hita, Jordi Miralda, Josep Maria Prats, the Ferguson's (Isabel, Joe, Marc, Sandy, Chris, Joan, Jennifer, and Eric) and Rafael Pous provided extra encouragement to come to the U.S. when it was most needed, in the Fall of 1986, and have always been there to support me during my PhD research.

My early teachers in high-school taught me how to think and encouraged me, very early on, to pursue my studies further. I was fortunate also to have truly exceptional teachers in the Autonomous University of Barcelona, the University of Barcelona and the Polytechnic Institute of Catalonia. My thought was shaped by Pons, Castellet, Deulofeu, Gibert, Sales.

Virgina Zimmerman edited the final draft and provided useful comments on how to improve the presentation.

Mariona Artús has balanced my work/play ratio. My mother Mercedes has supported me in whatever enadavor I have chosen to pursue. My father Juan taught me, early on, that most intellectual problems can be solved if one thinks hard and long enough about them. The rest of the family have been very encouraging (often by example). Thanks.

Financial support for this work was provided through the AI Laboratory in part by the Advanced Research Projects Agency under contract number DACA 76-85-K-0685 and under Office of Naval Research contract N00014-85-K-0124; Massachusetts Institute of Technology; Balearic Islands University; Consejo Superior de Investigaciones Científicas - the Spanish Research Council; NATO; Real Colegio Complutense; The Weizmann Institute of Science; Xerox Palo Alto Research Center; Nippon Telegraph and Telephone Human Interface Laboratories.

To the memory of my grandparents:
JUAN B. SUBIRANA i SUBIRANA, PhD in Math, Architect
Prof. XAVIER VILANOVA i MONTIU, PhD in Medicine, MD

Contents

List of Figures

Visual Processing and Non-Rigid Objects

Chapter 1

1.1 Problem Statement

Automatic visual recognition systems have been successful at recognizing rigid-objects. In contrast, there has been only limited progress in the recognition of non-rigid objects.

In this thesis, we address the problems involved in recognition when there are non-rigid objects in the scene. As part of this study we develop a two-level representation for non-rigid objects based on the study of three types of non-rigid objects: elongated flexible objects, contour textures, and holes.

The bulk of the work has been devoted to the problem of finding a computational paradigm that can perform, in the presence of non-rigid objects, perceptual organization and other related tasks such as finding a focus of attention and figure/ground segregation. The computational problem that we are interested in is that of finding curves and points in images that have certain properties that are not defined locally, neither in the image nor in the scene. As we will see in Section 1.2, these tasks belong to what is called intermediate or mid-level vision.

Section 1.2 will be devoted to clarifying the notions of mid-level vision and recognition of non-rigid objects. Section 1.3 outlines the problems addressed in this thesis and Section 1.4 gives an overview of the issues involved in non-rigid vision. Section 1.5 gives an outline of the recognition approach proposed in this thesis. The following two sections give a summary of contributions and an extended abstract with pointers to the rest of the thesis.

1.2 Non-Rigid Recognition and Mid-Level Vision

Humans usually conceive the natural world as being composed of different objects such as mountains, buildings, and cars. The identity of the objects that we see does not change despite the fact that their images are constantly changing. As we move in the environment, the objects are seen from different viewpoints. The change in viewpoint results in a corresponding change in the images even if the objects have not moved. The transformation of the objects in this case is rigid. Most of the existing computational work on recognition[1] is concerned with the case in which the transformations are rigid[2].

Objects can also change, however, in a non-rigid manner and still maintain their identity, for example trees, people, cables (see Figures 1.1, 1.2, and 1.5). Here we are concerned with problems associated with the recognition of non-rigid objects.

The study of visual processing in the presence of non-rigid objects is important for three reasons. First, many objects can undergo non-rigid transformations. Second, the ability to compensate for non-rigid transformations is closely related to the problem of object classification. In many cases, two similar objects in the same class (e.g. two oak leaves) can be viewed as related by a non-rigid transformation. Third, rigid objects often appear in images together with non-rigid objects and there is a need for schemes that can process images with both rigid and non-rigid objects.

Our ability to process images of non-rigid objects is not restricted to recognition. We can effortlessly perform other tasks such as directing our focus of attention, or segmentation. These two tasks belong to what is called mid-level vision.

We define mid-level vision as the set of visual algorithms that compute global structures that support segmentation and selective attention. The term mid-level vision is used to refer to tasks that belong to neither early-vision, which computes properties uniformly throughout the image array, nor high-level vision (a.k.a. late vision) which studies recognition, navigation, and grasping[3]. Two properties characterize mid-level algorithms. First, they compute extended structures such as convex regions. In contrast, early-vision algorithms are concerned with local properties such as discontinuities or depth maps. In other words, the result of an early-level computation at any one location depends only on the input values near that location. Second, they are related to the problems of selective attention and segmentation. The result of an intermediate level computation is often a data structure that is

[1]See [Besl and Jain 1985], [Chin and Dyer 1986], [Ullman 1989], [Grimson 1990], [Perrott and Hamey 1991] for extensive reviews.

[2]The term rigid transformation in computer vision often includes scale changes which, strictly speaking, are not rigid transformations.

[3]Note that mid-level vision is usually confounded with both early level and high-level vision.

selected for latter processing by another mid-level or high-level algorithm.

Note that we have defined the terms early-level, mid-level, and high-level in terms of the tasks that are performed without referring to the order of the computations. A lot of researchers, including [Marr 1982], assume that early level computations should be done before intermediate and high level tasks. That is why they refer to them as early, mid, and late vision. We, like others, do not make this assumption. In fact, in our scheme, mid-level precedes early level. If we had to coin the terms again we would probably use local-vision (for early vision), global-vision (for intermediate vision), and symbolic vision (for late vision).

Why is visual recognition in the presence of non-rigid objects different than when only rigid objects are present? Mainly because a change in appearance of a non-rigid object can not be attributed solely to changes in viewing position or lightness conditions (see Figures 1.1, 1.2, and 1.3). This implies that rigid transformations can not be used as the only cues to recover pose. As we will see in Section 1.4, the algorithms developed for non-rigid recognition are not likely to be applicable to rigid object recognition.

In addition, mid-level vision in the presence of non-rigid objects is more complex than when only rigid objects are present. A rigid transformation can be recovered with a correspondence of only three points [Ullman 1986], [Huttenlocker and Ullman 1987]. Therefore, a segmentation which reliably computes three points belonging to an object is sufficient to recover the pose of the rigid object. This explains why most of the work on segmentation of rigid objects is both concerned with reliably locating small pieces coming from an object, and does not extend to non-rigid objects. A segmentation algorithm useful for non-rigid recognition needs to recover a more complete description, not just three points or pairs of parallel lines. Gestalt principles commonly used in rigid object segmentation research, such as parallelism, can be used also for non-rigid objects. However, to be useful in non-rigid object segmentation, Gestalt principles have to be employed to recover full object descriptions not just a small set of points such as two parallel lines. Note that this does not imply that a good segmentation scheme for non-rigid objects can not be used for rigid objects. In fact, most of the work on segmentation presented here is relevant to both rigid and non-rigid objects.

A lot of the work on rigid object recognition is based on matching pictorial descriptions of objects. Again, these schemes are based on the fact that rigid pose can be determined with the correspondence of a few features. This is a widely used approach and differences between schemes are based on the search mechanism and the features used.

1.3 Curved Inertia Frames

What are the problems involved in non-rigid vision? Is there a general technique that will work in all cases? What tasks should be performed at "mid-level"? In other words, what is a valid research agenda for non-rigid objects? Ten years from now, what will the main session titles be at a conference on non-rigid recognition?

These are essential questions to pose before one can give structure to non-rigid vision research. Section 1.4 will give an overview of the problems involved in non-rigid vision as we understand them and a possible research methodology. This will include a description of different types of non-rigid objects. Each of these types of non-rigid objects requires different techniques and therefore can be studied separately.

One type of non-rigid objects are elongated shapes that can bend along their main axis, such as animal limbs, snakes, and cables. Among the issues involved in non-rigid vision we have focussed on segmentation and recognition of elongated flexible objects (see Figure 1.4).

We have developed a scheme, Curved Inertia Frames (C.I.F.), that finds object axes in arbitrary scenes. Such a scheme is useful for three purposes. First, it can be used for the recognition of elongated and flexible objects using a novel recognition scheme, *frame alignment*, introduced in Section 1.5 and Chapter 2. Second, it can be used in segmentation of rigid and non-rigid objects, including those that are elongated and flexible. Third, Curved Inertia Frames solve a computational problem related to that of finding optimal curves in several situations including:

1. Discontinuities (in stereo, texture, motion, color, brightness).

2. Frame curves as defined in Chapter 4.

3. Skeletons in binary images as shown in Chapter 2.

4. Skeletons in brightness images as shown in Chapter 3.

5. Statistical data interpolation.

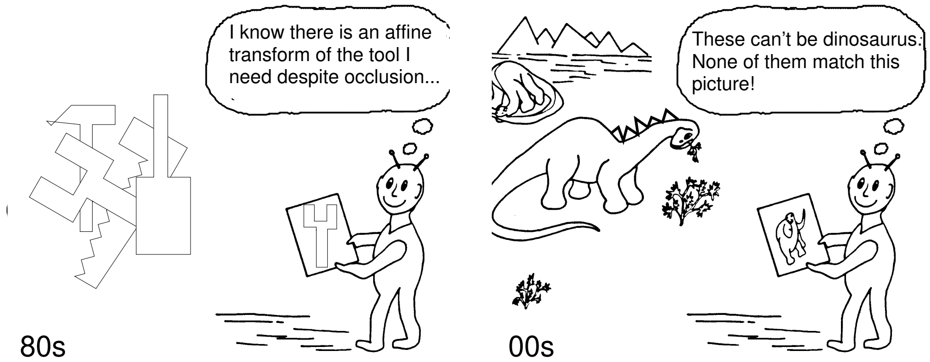


Figure 1.1: The robot of the 80s can recognize a rotated/scaled rigid object using noisy and spurious image contours. However, other transformations such as the bending of a dinosaur's neck could not be accounted for. In other words, pictorial matching (even when compensating for rotation and scale) can not account for non-rigid transformations and new techniques must be developed. Figure adapted from [Richards 1982].

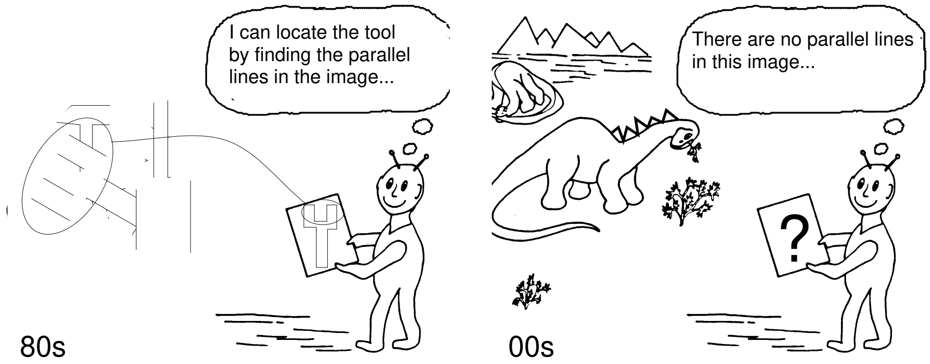


Figure 1.2: Mid-level algorithms that work for rigid objects do not work for non-rigid objects.

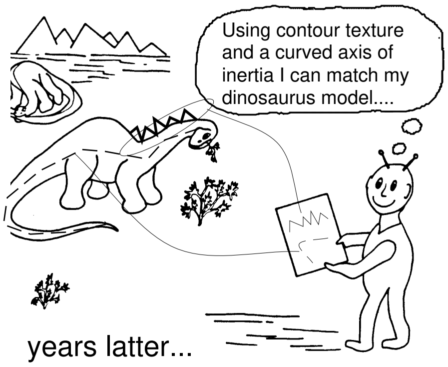


Figure 1.3: For perceptual organization and recognition, the dinosaur can be represented using Curved Inertia Frames and contour texture.

1.4 Five Key Issues in the Study of Non-Rigid Objects

There are several issues that need to be tackled to make progress in the study of vision in the presence of non-rigid objects. Five key issues are:

1. Are there different types of non-rigid objects?

2. How might a non-rigid object be represented?

3. Non-rigid objects have fuzzy boundaries.

4. What is the nature of mid-level algorithms in the presence of non-rigid objects?

5. How might non-rigid transformations be handled? How can it be verified that a hypothesized transformation specifies a correct match of a model to an instance?

The next five Subsections describe each of these in turn.

1.4.1 *Are there different types of non-rigid objects?*

Visual recognition is the process of finding a correspondence between images and stored representations of objects in the world. One is assumed to have a library of models described in a suitable way. In addition, each model in the library can undergo a certain number of transformations (e.g. rotation, scaling, stretching) which describe its physical location in space (with respect to the imaging system). More precisely:

Definition 1 (Recognition problem): *Given a library of models, a set of transformations that these models can undergo, and an image of a scene, determine whether an/some object/s in the library is/are present in the scene and what is/are the transformations that defines its/their physical location in space*[4].

Together with autonomous navigation and manipulation, recognition is one of the abilities desired in automatic visual systems. Recognition is essential in tasks such as recognition of familiar objects and one's environment, finger-print recognition, character recognition (printed and handwritten), license plate recognition, and face recognition. Recognition can be of aid to geographical information systems, traffic control systems, visual inspection, navigation, and grasping.

As mentioned above, existing recognition schemes impose restrictions on the type of models and the transformations that they can handle. Occasionally, restrictions will also be imposed on the imaging geometry, such as fixed distance (to avoid scaling) and orthographic projection (to avoid perspective). The domain in which most work exists is that in which the library of models is assumed to be composed of rigid objects. In fact, the ability to cope with both sources of variability, changes in the imaging geometry, and non-rigid transformations, under the presence of noise and occlusion, is the essence of any non-rigid recognition scheme.

[4]This definition could be stated more precisely in terms of matching specific mathematical shapes but for the purposes of the present discussion this is not necessary.
Sometimes, solving for the transformation that defines the model is considered a separate problem and is termed the pose problem.

A natural question is whether one should recognize non-rigid objects with a different scheme than the one used for rigid objects. In fact, there is the possibility that non-rigid objects, in turn, are divided into several categories based on appliable recognition strategies.

This is an important point and not a new one to vision. For example, several techniques exist that recover depth from images: shape from stereo, shape from motion, shape from shading.

Thus, when considering non-rigid object recognition one is naturally taken to the question of whether it is just one "problem" or several. In other words, is there a small set of different categories of non-rigid objects such that objects in the same category can be recognized with similar recognition schemes? In fact, the difference between the categories may be so large that maybe it is not only matching that is different but also the other stages of the recognition process. A possible way of performing this characterization is by looking at the process that caused the non-rigidity[5] of the shape. This results in at least the following categories (See Figures 1.5, and 1.6):

- **Parameterized Objects:** These are objects which can be defined by a finite number of parameters (real numbers) and a set of rigid parts. Classical examples include articulated objects (where the parameters are the angles of the articulations) or objects stretched along particular orientations (where the parameters are the amount of stretching along the given orientations). This domain has been studied before, see [Grimson 1990], [Goddard 1992] for a list of references. Some rigid-based schemes can be modified to handle parameterized objects when the parameters that define the transformation can be recovered by knowing the correspondence of a few points (e.g. [Grimson 1990]).

- **One-dimensional Flexible Objects:** These are objects defined by a rigid object and a mapping between two curves. The simplest case is that of elongated objects that can bend such as a snake, the tail of an alligator, and the neck of a giraffe (see Figures 4.1, 2.1, and 2.3). In these examples, the two curves correspond to the spinal cord in both model and image [Subirana-Vilanova

[5]Other alternatives exist. For instance, one could use non-rigid motion research as a basis for a characterization [Kambhamettu, Goldgof, Terzopoulos and Huang 1993]

1990].

- **Elastic Objects:** For objects in this category, the transformation between the model and the image is given by a tensor field. Examples include hearts, facial expressions, and clothing. Schemes to recognize certain elastic objects in the 2D plane exist [Broit 1981], [Bajcsy and Kovacic 1987], [Feynman 1988], [Moshfeghi 1991]. Most work on non-rigid motion can be adapted to recognition of elastic objects.

- **Contour Textures:** Models in this category are defined by an abstract description along a curve which we call the *frame curve* of the contour texture. Abstract descriptions include the frequency of its protrusions or the variance in their hight. Examples in this category include clouds, trees, and oak leaves. A scheme to perform Contour Texture segmentation is presented in Chapter 4 [Subirana-Vilanova 1991].

- **Crumpling/Texture:** Certain objects such as aluminum cans or cloth can be recognized even if they are crumpled (see Figure 1.6).

- **Handwritten Objects:** Cartoons and script are non-rigid objects which humans can recognize effortlessly and which have received a lot of attention in the Computer Vision literature.

- **Symbolic Descriptors:** Models in this category are defined by abstract or functional symbolic descriptors. They include part descriptions and holes [Subirana-Vilanova and Richards 1991]. Holes are objects contained one inside another with a non-rigid arrangement such as a door in an unknown wall, see Appendix B.

Note that these categories define possible transformations for non-rigid objects. In many cases, a given object may undergo several of these transformations or have a "part" description [Pentland 1988]. For example, the tail of an alligator can bend as an elongated flexible object but its crest has a distinctive contour texture. A collection of examples of non-rigid objects can be seen in [Snodgrass and Vanderwart 1980] (note that each of the 300 shapes shown in [Snodgrass and Vanderwart 1980] can be classified into one of the above groups).

Note also that not all existing recognition schemes are designed for one particular category (from the ones above). In fact, there is a lot of work on non-rigid motion and recognition that tries to match "somewhat" distorted patterns [Burr 1981], [Segen 1989], [Solina 1987], [Pentland and Horowitz 1991] (see [Kambhamettu, Goldgof, Terzopoulos and Huang 1991] for an overview to non-rigid motion). Thus, some of these schemes are hard to characterize as belonging to one particular category.

In this thesis we will present research in two of the above categories:

- *Elongated Flexible Objects*
- *Contour Textures.*

Appendix B presents work on symbolic objects (holes and part arrangements) which is also reported elsewhere [Subirana-Vilanova and Richards 1991].

Recognition Versus Classification

One question arises as to whether it is the same physical object or a copy of it that appears in an image. In some cases, it is useful not to restrict the object in the scene by considering it to be the very same model. This is sometimes necessary as when trying to distinguish different industrial parts some of which are copies of each other - too similar to be worth distinguishing in a noisy image. Another example is a model of an arbitrary triangle which defines the model only partially and includes not one but a set of possible objects (all triangles) [Rosch, Mervis, Gray, Johnson and Boyes-Braem 1976], [Mervis and Rosh 1981]. In this last example, the recognition problem is often termed "classification" because the models do not specify just one shape but a group or class of shapes [Mervis and Rosch 1981], [Murphy and Wisniewski 1989].

Hence, there is not a clear computational difference between the recognition and classification problems. In non-rigid objects, such a distinction is even more fuzzy and less useful than in the realm of rigid objects. Indeed, whether we are looking at the same wave a few feet closer to the beach or at another wave will not make much of a difference if we are recognizing static images. In other words, often when recognizing non-rigid objects from a computational point of view it is not necessary to distinguish between the classification and the recognition problem. Saying that two leaves are oak leaves and that two clouds are of the same type can be seen as a classification problem. However, the *transformation* between the shapes in both examples can be cast in a similar way to what we would use to track a cloud in the sky. However, this latter case is an example of contour texture recognition.

For the purposes of this thesis we will concentrate on the computational problem of recovering the transformations between non-rigid objects. Whether we are interested in recognition or classification is orthogonal to our research. The facts that contour textures can often be used in classification (but also recognition) and that elongated flexible objects can often be used in recognition (but also classification) reinforces our point that the distinction between classification and recognition is fuzzy.

1.4.2 *How might we represent a non-rigid object?*

Before a recognition system can be put to work we must, somehow, instill in it knowledge about the objects that it is to locate in the image. No matter what non-rigid transformation we are addressing (see list given in previous subsection), one is left with the question of how objects will be represented. There is no reason why such knowledge can not be represented in different ways for different classes of objects. In fact, the schemes presented in this thesis use different types of representations for different non-rigid transformations.

In particular, a novel two-stage representation will be suggested for non-rigid objects and other complex shapes (see Figure 1.3). The findings presented in this thesis argue in favor of such two-level representation in which one level, which we call the "frame curve," embodies the "overall shape" of the object and the other, which we call the "contour texture," embodies more detailed information about the boundary's shape. See Chapter 4 for a more detailed description and examples.

1.4.3 *Non-rigid objects have fuzzy boundaries*

The third issue is the way in which non-rigid boundaries are represented and is therefore related to the representation one (the second in the list). Rigid object boundaries can be divided into two classes:

- The first class, sharp edge boundaries, is that of boundaries with edges corresponding to rigid wireframes such as polygons and most industrial parts. In these objects, the edges from different viewpoints are related by a rigid transformation.

- The second class, objects with smooth surfaces, is that of boundaries without this property because of a curved surface such as the nose of a face or the body of most airplanes. In this case, the brightness edges from different view points are not related by a rigid transformation, e.g. when a vertical cylinder is rotated the visible edges do not change.

This division can be used to classify rigid object recognition schemes into two classes depending on the type of rigid objects they can deal with.

This thesis argues that non-rigid objects, in addition, have a third type of boundary which we call fuzzy boundaries. Fuzzy boundaries include the boundary of a tree (where is it?) and that of a cloud (where is it?). As we will see in Chapter 4 and in Appendix B, such boundaries are also useful for representing certain complex rigid objects (such as a city skyline).

1.4.4 *What is the nature of mid-level algorithms in the presence of non-rigid objects?*

As mentioned in Section 1.2, middle level vision is concerned with the study of algorithms which compute global properties that support segmentation and selective attention. Finding a small set of features likely to come from the same object is useful for rigid object recognition but not for the recognition of non-rigid objects.

Recent work in computer vision has emphasized the role of edge detection and discontinuities in mid-level vision, and has focussed on schemes that find small subsets of features likely to come from a single object; we are interested in robust perceptual organization schemes that can compute complete regions coming from a single object without relying on existing edge detectors or other early vision modules.

This thesis suggests a scheme in which *perceptual organization precedes early vision modules* such as edge-detection or motion (see Figure 1.7). We also present evidence that agrees with a model of human visual perception in line with this suggestion.

1.4.5 *How might non-rigid transformations be handled? how can it be verified that a hypothesized transformation specifies a correct match of a model to an instance?*

The final output of a recognition system is a transformation in the representation language of the designer's choice. Such a transformation is established between the model and the output of the intermediate level processes. Since it is important to be able to recognize objects from a relatively novel view as when we recognize the outline of a familiar town from a friend's house, common studied transformations are changes in scale and viewing geometry.

Three critical issues in determining the transformation are hypothesizing (what transformations are possible), indexing (what transformations/models are likely), and verification (which transformations are accurate). Verification refers to the problem of establishing a fitness measure between stored representations of objects and hypothesized locations of the object in the pre-processed sensor output. Such a fitness measure must be such that it gives high scores if, and only if, there is an object in the hypothesized location similar to the stored model.

Confounded with recovering transformations is the tolerance to noisy and spurious data. For example, it is important that systems be able to recognize objects even if they are partially occluded as when a cloud is hidden behind a building. The scheme that we will present in the next Chapters is remarkably stable under noisy and spurious data.

Figure 1.4: The objects in these Figures (their parts or the category they belong to) are elongated and flexible.

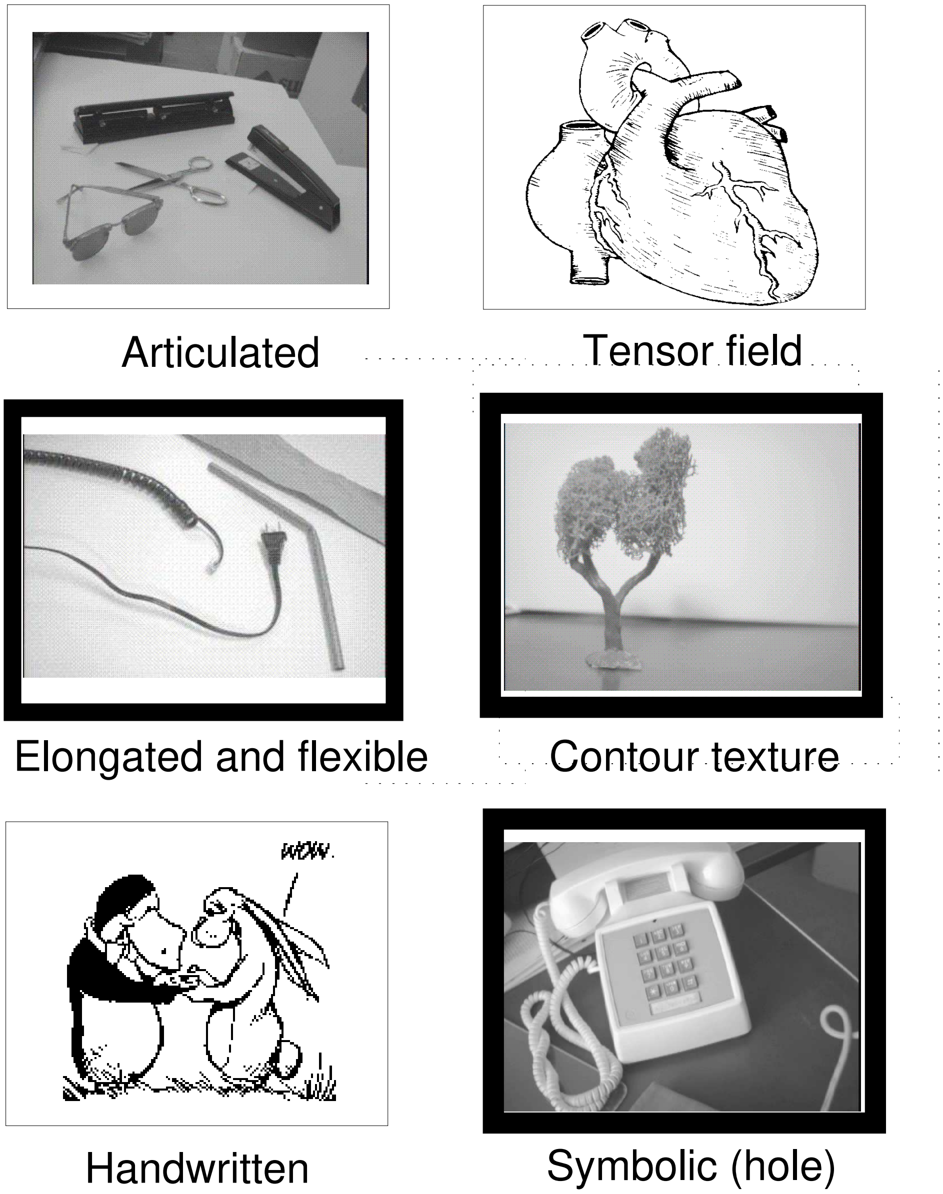


Figure 1.5: Examples of different types of non-rigid objects (see text for details and Figure 1.6). Does each of these types require a different recognition scheme?

Figure 1.6: Is crumpling an instance of 2D texture recognition?

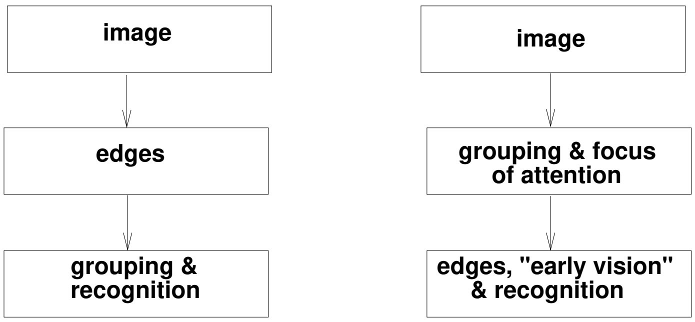


Figure 1.7: Two different views on the role of perceptual organization. *Left:* Discontinuities are the first computational step - a model widely used in Computer Vision. *Right:* We (like others) suggest a model in which perceptual organization precedes "early vision."

NON-RIGID TRANSFOR-MATION	APPLICATION	MODELING	ALGORITHM	SIGNAL PROCESSING
Elongated and flexible	Medical imag. People pose estimation as interface Handwriten character rec..	Curved axis of inertia	Dynamic pro-gramming Random networks	Ridge detector Color/brightness

Figure 1.8: This Figure illustrates the methodology proposed in this Thesis for investigating the recognition of non-rigid objects (see text for details).

1.5 The Frame Alignment Approach to Object Recognition

In Section 1.4.1 we suggested the use of different techniques for each type of non-rigid transformation. In general, an object can have different transformations. For example, the tail of an alligator is elongated and flexible and has a distinctive contour texture.

Despite the fact that the we propose different approaches for the recognition of elongated flexible objects and contour textures, there is one similarity between the approaches which is worth mentioning: in both cases a curve is used to guide the recognition process. In the first case this curved is called the skeleton of the object and in the latter the frame curve. In both cases such curve, or frame curve, is employed to align the model with the object. Thus, we call our approach "frame alignment".

Frame alignment is the approach that we propose for the recognition of elongated flexible objects and contour textures. It is closely related to the alignment method to recognition. The basic idea behing the alignment approach is to divide the search for candidate objects into two stages. First, for all candidate models determine the transformation between the viewed object and the object model. Second, determine the object-model that best matches the viewed model. This idea is old, in fact, the axis of inertia of a shape has been used before to align an object to an image [Hu 1962], [Weiser 1981], [Marola 1989a, 1989b].. The axis of inertia is a common feature provided in commertial computer vision packages since it can be computed easily and can greatly speed matching of two dimensional shapes. A summary of the philosophy behind the alignment approach can be found in [Ullman 1986].

There is a large number of existing alignment schemes. One thing that distinguishes one from another is the type of features that are used in the first stage. The set of such features is called the anchor structure and it is used to determine the transformation between the viewed object and the object model. Other proposals exist. For example, people have used different arrangements of points or lines [Lowe 1986], the center of mass [Neisser 1967], the local maxima in a distance transform

[Orbert 1989] and the axis of inertia or symmetry [Hu 1962], [Weiser 1981], [Marola 1989a, 1989b].

As another example, [Huttenlocker and Ullman 1987] present an alignment scheme for rigid objects in which three points are shown to be sufficient to recover the transformation. For each triple, the anchor structure in this example, the transformation between model and object is computed and a best one-to-one match is sought among all models. The scheme performs an exhaustive search over all possible triples of points and a record of the best match is kept. Thus, the number of points is important because it is directly related with the complexity of the search space.

The basic idea behind our recognition approach to elongated and flexible object recognition is to use a curve, the object's skeleton, as an anchor structure. The skeleton can be used to unbend the object so that it can be match against a canonical view of the model (see Figures 1.9 and 2.3). In Chapters 2 and 3 we will show how Curved Inertia Frames can be used to compute the skeleton of an elongated and flexible object.

Similarly, the basic idea behind our recognition approach to contour texture is to use the contour's central curve (or frame curve) to determine the transformation between the viewed object and the object model. Chapter 4 presents the second stage in which a set of filters are used to match the "unbent" contour textures. Again, we suggest the use of Curved Inertia Frames to compute the frame curve. [Ullman 1986] suggested that abstract descriptors be used in alignment. In such context, contour texture can be seen as a formalization of a certain class of abstract descriptors[6].

We call our approach frame alignment since it uses a curve computed by Curved Inertia Frames as an anchor structure to recognize contour textures and elongated flexible objects. Schemes that use the axis of inertia (or other frames) to align the object also belong to the frame alignment approach. Note that in both cases the computation of the curve is confouned with perceptual organization[7].

[6]Contour texture and abstract descriptors are different notions. In fact the notion of abstract descriptor is more general and need not be restricted to contour texture. In addition, and as we will see in Chapter 4, we suggest that contour texture be used also in other tasks such as perceptual organization, indexing, depth perception, and completion.

[7]Note that we have only implemented the frame computation stage for elongated and flexible objects and not for contour textures.

1.6 Major Contributions of this Thesis

Since the recognition of non-rigid objects is a relatively new domain, We have had to address a diverse set of issues. Nevertheless, this thesis has four major contributions which can be divided into two groups (following the dissertation's title):

- **Non-Rigid Object Recognition:** We have identified different relevant problem domains in the area of non-rigid visual recognition. This includes a list of different types of non-rigid objects and a two-level representation for non-rigid objects based on the two domains in which we have concentrated our work (see Figure 1.3):

  - **Elongated and Flexible Objects:** We have presented a scheme to recognize elongated flexible objects using frame alignment and evidence against its use in human perception. We have used such evidence to suggest that, in human perception, holes are independent of the "whole".
  - **Contour Texture and Fuzzy Boundaries:** We have proposed a filter-based scheme for segmentation and recognition of contour textures. In addition, we have described several peculiarities of non-rigid boundaries in human perception. Most notably the notion that fuzzy boundaries exist and that their inside/top/near/incoming regions are more salient.

  Both approaches to recognition can be casted within the frame alignment approach to object recognition.

- **Curved Inertia Frames (C.I.F.):** C.I.F. is a parallel architecture for mid-level non-rigid vision based on ridge-detection and random networks. C.I.F. is the first scheme that can find provably global curved structures. The scheme is based on a novel architecture to vision ("random networks") and on finding perceptual groups (and other mid-level structures) directly on the image (i.e. without edges) using a novel ridge detector.

In the following four subsections we discuss in more detail some of these contributions.

1.6.1 *Classification of non-rigid objects: Frame Alignment and two level-representation*

This thesis proposes the use of different schemes for recognizing contour texture and elongated structures. Identifying different domains is a hard problem and a contribution in itself, especially in new areas such as the recognition of non-rigid objects. In other vision domains it is well accepted that different techniques should be used to solve a given taks. For example, diverse techniques exist for depth perception, stereo, shading, motion, texture, focus, range data (psychophysical evidence suggests the human visual system uses several techniques as well). Similarly, we propose the use of different techniques for recognition of non-rigid objects.

The schemes that are proposed for the recognition of contour textures and elongated flexible objects are instances of frame alignment and lead to a novel two-level representation of objects which can support other non-rigid transformations.

The methodology for non-rigid vision used in this thesis is outlined in Figure 1.8. We believe it can be extended to non-rigid transformations not covered in this thesis (see Figure 5.1).

1.6.2 *Curved axis of inertia and center of mass for mid-level vision*

The study of reference frames has received considerable attention in the computer vision literature. Reference frames have been used for different purposes and given different names (e.g. skeletons, voronoi diagrams, symmetry transforms). Previous schemes for computing skeletons fall usually into one of two classes. The first class looks for a straight axis, such as the axis of inertia. These methods are global (the axis is determined by all the contour points) and produce a single straight axis. The second class can find a curved axis along the figure, but the computation is based on local information. That is, the axis at a given location is determined by small pieces of contours surrounding this location. In addition, recently schemes based on snakes [Kass, Witkin, and Terzopoulos 1988] have been presented; these schemes require an initial good estimate or are not provably optimal.

We have developed a scheme for computing skeletons which is curved and global. Such a novel definition of curved-axis of inertia can be used for several mid-level tasks. However, our motivation comes from its use in perceptual organization and computation of a focus of attention in the recognition of elongated and flexible objects. Other possible uses are described in Chapter 2.

The definition of curved axis of inertia used by Curved Inertia Frames is such that it can be applied directly on the image and for different tasks such as computing frame curves, ridge detection, and early vision.

1.6.3 *Dynamic programming and random networks*

The problem of finding curves in images is an old one. It appears in many aspects of computer vision and data processing in general. The problem solved here is closely related to that of finding discontinuities in stereo, motion, brightness, and texture.

Two things make the approach presented here relevant. First, the definition of curved axis of inertia stated above. Second, we present a dynamic programming scheme which is provably global and is guaranteed to find the optimal axes (it simply can not go wrong!).

Dynamic programming has been used before but only on bitmap images and without provably global results (due to the need for smoothing or curvature computation) nor incorporating Gestalt notions such as symmetry and convexity [Ullman 1976], [Shashua and Ullman 1988], [Subirana-Vilanova 1990], [Subirana-Vilanova and Sung 1992], [Spoerri 1992], [Freeman 1992].

We present a remarkably efficient dynamic programming network that can compute globally salient curves that are smooth (unlike the approaches of [Shashua & Ullman 1988], [Subirana-Vilanova 1990], [Spoerri 1992], and [Freeman 1992] which use non-global approximations of smoothness based on energy minimization or curvature propagation).

We present a proof that such dynamic programming networks can use one and only one state variable; this constrains their descriptive power, yet we show that one can combine curvature, length, and symmetry to find curves in an arbitrary surface (not just on bitmap representations).

There are three other things that make this approach relevant. First, it uses a non-cartesian network on the image plane. In particular, we introduce the notion of random networks which are built by randomly throwing line arrangements of processors in the image plane. The approach lends itself to other types of arrangements such as those resulting from lines uniformly sampled in polar coordinates.

Second, all processors represent (about) the same length unlike previous approaches in which processors at some orientations (such as "diagonals") have different length. This is important because our approach does not have a small subset of orientations (such as the horizontal, vertical, and diagonal) that are different than the rest.

Third, processing can be selectively targeted at certain regions of the image array. This is different from previous approaches which use cartesian networks and therefore have the same density of processors in all areas of the image array.

Finally, we should mention that C.I.F. can be used in several tasks. We demonstrate its use in two tasks (finding skeletons in bitmaps and in color images). There are other domains in which they could be used such as finding discontinuities (in stereo, texture, motion, color, brightness), finding frame curves, and statistical data interpolation.

1.6.4 *Processing directly in the image: A ridge detector*

Recent work in computer vision has emphasized the role of edge detection and discontinuities in segmentation and recognition. This line of research stresses that edge detection should be done both at an early stage and on a brightness representation of the image; according to this view, segmentation and other early vision modules operate later on (see Figure 1.7 left). We (like some others) argue against such an approach and present a scheme that segments an image without finding brightness, texture, or color edges (see Figure 1.7 right and Section 3.2). In our scheme, discontinuities and a potential focus of attention for subsequent processing are found as a byproduct of the perceptual organization process which is based on a novel ridge detector.

Our scheme for perceptual organization is new because it uses image information and incorporates Gestalt notions such as symmetry, convexity, and elongation. It is novel because it breaks the traditional way of thinking about mid-level vision and early vision as separate processes. Our suggestion implies that they are interrelated and should be thought of as the same.

In particular, we present a novel ridge detector which is designed to automatically find the right scale of a ridge even in the presence of noise, multiple steps, and narrow valleys. One of the key features of the ridge detector is that it has a zero response at discontinuities. The ridge detector can be applied both to scalar and vector quantities such as color.

We also present psychophysical evidence that agrees with a model in which a frame is set in the image prior to an explicit computation of discontinuities. This contrasts with a common view that early vision is computed prior to mid-level (or global vision).

1.7 Road Map: Extended Abstract With Pointers

In this dissertation we address the problem of visual recognition for images containing non-rigid objects. We introduce the frame alignment approach to recognition and illustrate it in two types of non-rigid objects: contour textures (Chapter 4) and elongated flexible objects (Chapters 2 and 3 and appendices). Frame alignment is based on matching stored models to images and has three stages: first, a "frame curve" and a corresponding object are computed in the image. Second, the object is brought into correspondence with the model by aligning the model axis with the object axis; if the object is not rigid it is "unbent" achieving a canonical description for recognition. Finally, object and model are matched against each other. Rigid and elongated flexible objects are matched using all contour information. Contour textures are matched using filter outputs around the frame curve.

The central contribution of this thesis is Curved Inertia Frames (C.I.F.), a scheme for computing frame curves directly on the image (Chapters 2 and 3). C.I.F. can also be used in other tasks such as early vision, perceptual organization, computation of focus of attention, and part decomposition. C.I.F. is the first algorithm which can compute probably global and curved lines. C.I.F. is based on computing reference frames using a novel definition of "curved axis of inertia" (defined in Sections 2.5 to 2.7). Unlike previous schemes, C.I.F. can extract curved symmetry axes and yet use global information. Another remarkable feature of C.I.F. is its stability to input noise and its tolerance to spurious data.

Two schemes to compute C.I.F. are presented (Chapters 2 and 3). The first uses a discontinuity map as input and has two stages. In the first stage (Section 2.5), "tolerated length" and "inertia values" are computed throughout the image at different orientations. Tolerated length values are local estimates of frame curvature and are based on the object's width, so that more curvature is tolerated in narrow portions of the shape. Inertia values are similar to distance transforms and are local estimates of symmetry and convexity, so that the frame lies in central locations. In the second stage (Section 2.6), a parallel dynamic programming scheme finds long and smooth curves optimizing a measure of total curved inertia. A theorem proving computational limitations of the dynamic programming approach is presented in Section 2.9. The study of the theorem leads to several variations of the scheme which are also presented, including a 3D version and one using random networks in Section 2.10 (random networks had never been used in vision before and provide speed ups of over 100).

The second scheme, presented in Chapter 3, differs from the above one in its first stage. In the second version, tolerated length and inertia values are computed directly in the image. This enables the computation of C.I.F. using color, texture, and brightness without the need to pre-compute discontinuities. Exploring schemes that work without edges is important because often discontinuity detectors are not reliable enough, and because discontinuity maps do not contain all useful image information. The second scheme is based on a novel non-linear vector ridge detector presented in Section 3.7 which computes tolerated length and inertia values directly on the image (Section 3.6 describes problems involved in the definition of ridges). The ridge detector does not respond to edges and selects the appropriate scale at every image location. Section 3.9 present experiments on color images and analytical results presented in Section 3.8 show that, under varying scale and noise, the scheme performs remarkably better than well known linear filters with optimum SNR.

The computation of C.I.F. is useful for middle level vision because it results in both a set of central points and a perceptual organization. Central points of the object, at which further processing can be directed, are computed using a novel definition of "curved center of mass" (Section 3.9). C.I.F. also computes a set of "largest curved axes" which leads to a perceptual organization of the image resembling a part-description for recognition. Most perceptual organization schemes designed for rigid-object recognition can not be used with non-rigid objects. This is due to the fact that they compute only a small set of localized features, such as two parallel lines, which are sufficient to recover the pose of a rigid object but not that of a non-rigid one. By using the notion of curved axis of inertia, C.I.F. is able to recover large, elongated, symmetric, and convex regions likely to come from a single elongated flexible object.

A scheme to recognize elongated flexible objects using frame alignment and the perceptual organization obtained by C.I.F. is presented in Chapter 2. The scheme uses C.I.F. as an anchor structure to align an object to a canonical "unbent" version of itself. The scheme, including the computation of C.I.F., has been implemented on the Connection Machine and a few results are shown throughout this thesis. Evidence against frame alignment is presented in Appendix B and in [Subirana-Vilanova and Richards 1991].

C.I.F. leads to a two-level representation for the recognition of non-rigid objects (see Figure 1.3). Chapters 2 and 3 introduce the first level, which is a part description of the shape computed by C.I.F.. The second, which we call contour texture, is a representation of complex and non-rigid boundaries, and is described in Chapter 4.

In Chapter 4, we show that frame alignment can also handle fuzzy non-rigid boundaries such as hair, clouds, and leaves. We call such boundaries contour textures and present (in Section 4.6) a novel filter-based scheme for their recognition. The scheme is similar to existing filter-based approaches to two-dimensional texture, with two differences: the emphasis is on a description along the main boundary contour, not along a two-dimensional region; and inside/outside relations are taken into account. We present psychophysical evidence (similar to that mentioned above) that a pictorial matching measure based solely on distance is not sufficient unless inside/outside relations are taken into account. The scheme can handle spurious data and works directly on the image (without the need for discontinuities).

In summary, the first four chapters of this thesis present a computational approach to non-rigid recognition based on a shape representation with two-levels: a part description capturing the large scale of the shape, and a complementary boundary description capturing the small scale (see Chapter 4 and Section 4.5). The former is computed by C.I.F. and the latter by contour texture filters.

In Chapter 5 we review the work presented and give suggestions for future research. In Appendix A we discuss how Curve Inertia Frames can incorporate several perculiarities of human perception and present several unresolved questions.

In Appendix B we present evidence against frame alignment in human perception. However, this evidence suggests that frame curves have a role in figure/ground segregation and in fuzzy boundaries, and that their outside/near/top/incoming regions are more salient. These findings agree with a model in which human perception begins by setting a frame of reference (prior to early vision), and proceeds by successive processing of convex structures (or holes).

Thus, in this Thesis we present work on non-rigid recognition (Chapter 1 and Sections 2.1, 2.8, 3.1, 3.2, 3.5, 3.6, 3.9, and Appendices), algorithms to compute frame curves (Chapters 1, 2, and 3), and human perception (Chapter 4, and Appendices A and B).

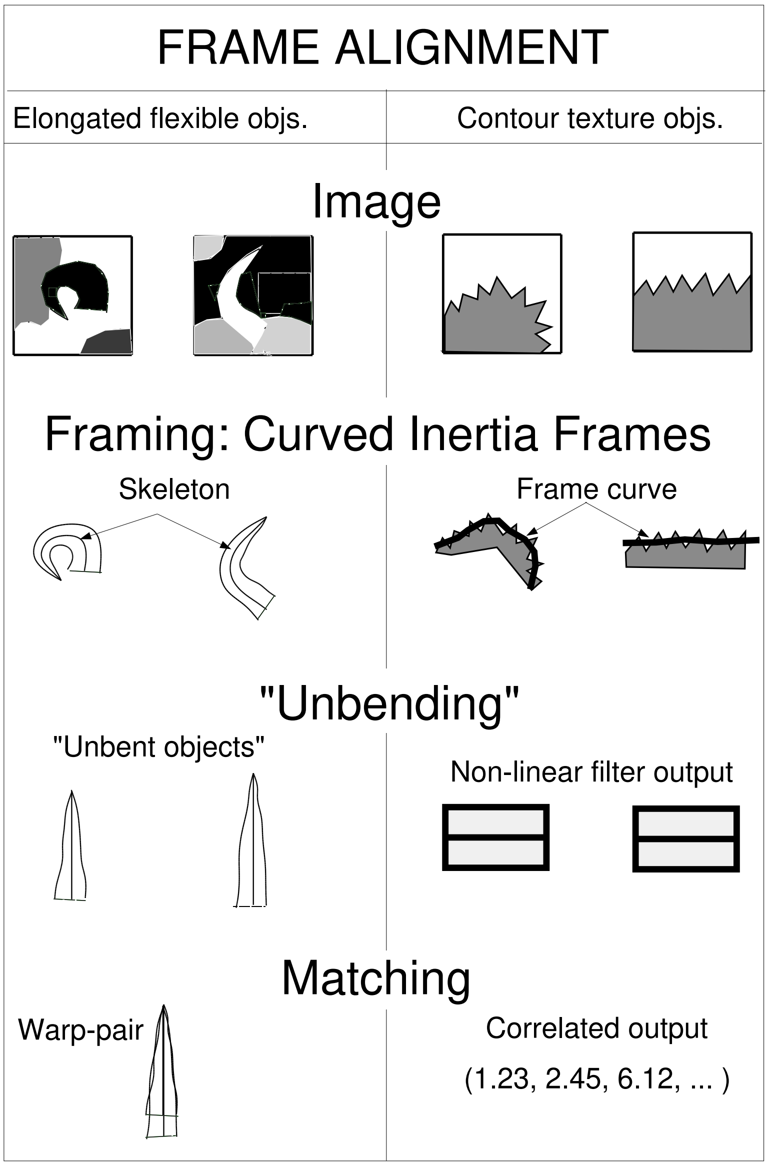


Figure 1.9: This figure illustrates how frame alignment works on two types of non-rigid objects. In both cases, perceptual organization is done by Curved Inertia Frames *Left:* Frame alignment applied to elongated and flexible objects. Curved Inertia Frames computes an axis which is used to unbend the objects and match them to a model. *Right:* Frame alignment applied to contour textures. Curved Inertia Frames computes the center line of the contour texture, the frame curve, and a filter-based scheme is applied around the frame curve to match the objects. See text for details. Two other examples are shown in Figure 5.11.

Elongated Flexible Objects

Chapter 2

2.1 Skeletons for Image Warping and the Recognition of Elongated Flexible Objects

Elongated and flexible objects are quite common (see Figures 2.1, 2.6, 2.4, 5.2, 2.2, and 2.3). For example body limbs, tree branches, and cables are elongated or have parts that are elongated and flexible. Gestalt psychologists already knew that elongation was useful in perceptual organization. Elongation has been used as acue for segmentation by many. However, there is not much work on the recognition of this type of objects [Milller 1988], [Kender and Kjeldsen 1991].

In Section 1.4 we discussed five of the issues involved in the recognition of non-rigid objects. Frame alignment, our approach to recognizing elongated and flexible objects is best understood by first considering the second issue in the list: what is the shape representation used.

A shape representation is an encoding of a shape. A common approach is to describe the points of the shape in a cartesian coordinate reference frame fixed in the image (see Figure 2.7). An alternative is to center the frame on the shape so that a canonical independent description can be achieved. For some shapes this can be obtained by orienting the frame of reference along the inertia axis of the shape (see Figure 2.7).

If the objects are elongated and flexible, we suggest another alternative that might be more appropriate, the use of a curved frame of reference (see Figure 2.4). Recognition can be achieved using a canonical description of the shape obtained by rotating or "unbending" the shape using the frame as an anchor structure (see Figure 2.4). This approach belongs to the frame alignment approach to recognition as described in Section 1.5.

In this Chapter, we address the problem of finding reference frames in bitmap images (a.k.a. skeletons, symmetry transforms, distance transforms, voronoi diagrams etc.). Our approach is called Curved Inertia Frames (C.I.F.) and is based on a novel definition of "curved axis of inertia" and the use of non-cartesian networks. In Chapter 3 we extend C.I.F. to work directly on the image.

2.2 Curved Inertia Frames and Mid-Level Vision

The relevance of our work on C.I.F. extends beyond the recognition of elongated and flexible objects. In fact, the problem of recovering a curved axis for recognition is related to many problems in intermediate level vision. Frame axis can be used for a variety of tasks such as recognition, attention, figure-ground, perceptual organization and part segmentation. For example, for complex shapes a part decomposition for recognition can be obtained with a skeleton[1]-like frame[2] (e.g. [Connell and Brady 1987], see Figure 2.8). Curved Inertia Frames is capable of performing several intermediate level tasks such as perceptual organization and locating salient points at which to direct further processing.

Little is known about middle-level vision in humans. Are middle-level visual computations performed on top (and independently of) early vision? How many different computations are performed? What is the output of middle-level vision?

Chapters 2, 3, 4 and the Appendices of this thesis address middle-level vision - with special emphasis on domains where non-rigid objects are present. We are

[1]A more detailed definition of skeletons will be given later in the Chapter

[2]There is some evidence that these types of representations may be easier to recognize than images of the objects [Rhodes, Brennan, and Carley 1987]

interested in exploiting regularities in nature to design useful algorithms which work directly in the image (without knowing the detailed contents of the image). The words "useful," "directly," and "regularities" require further clarification. By useful we mean several things: if the algorithms can help existing computer vision systems, they are useful; if the algorithms can help us gain insight into human perception, they are useful; if the algorithms model certain properties of the visual system, they are useful. By directly we mean that we are interested in solving an intermediate level task working directly in the image array. In other words, we would like to disregard early vision modules so that it is not necessary to compute discontinuities before finding the frames[3] By regularities we mean general properties about the scene being imaged which can be effectively used to solve middle-level vision problems. One of the most widely used regularities is parallelism: parallel structures normally come from the same object so that, when present in images, they can be used to segment the image without the need for explicitly knowing the object. The scheme that we present in the next Chapter, Curved Inertia Frames, is designed to make use of parallelism, convexity, elongation, and size. What is new about it is that it uses curved and global measures.

Outline of Curved Inertia Frames

We will present two versions of C.I.F.. In this Chapter we will present one that uses discontinuity maps as input. In the next Chapter we present a second version which extends the scheme to work directly on the image (without edges).

C.I.F. is divided into two successive stages. In Section 2.5, we present the first stage, in which we obtain two local measures at every point, the "inertia value" and the "tolerated length", which will provide a local symmetry measure at every point and for every orientation. This measure is high if locally the point in question appears to be a part of a symmetry axis. This simply means that, at the given orientation, the point is equally distant from two image contours. The symmetry measure therefore produces a map of potential fragments of symmetry curves which we call the inertia surfaces. In Sections 2.6 and 2.7, we present the second stage in which we find

[3]This Chapter will present a scheme which uses early vision modules. In Chapter 3 we will extend the scheme so that it works directly on the image.

long and smooth axes going through points of high inertia values and tolerated length. In Section 2.8, we introduce a novel data structure, the skeleton sketch, and show some results and applications of the scheme. In Section 2.9, we prove a theorem that shows some strong limitations on the class of measures computable by the computation described in Sections 2.6 and 2.7. In fact, it proves that the version of C.I.F. presented in Sections 2.6 and 2.7 is not global. This leads to a new computation based on non-cartesian networks in Section 2.10. This is a central Section because it introduces a truly global and efficient computation.

In Appendices B and A we discuss the relation between C.I.F. and human perception. In Section 2.11, we present some limitations of our scheme and a number of topics for future research. In Chapter 3, we extend our scheme to grey-level, color and vector images by presenting a scale-independent ridge-detector that computes tolerated length and inertia values directly on the images.

2.3 Why Is Finding Reference Frames Not Trivial?

Finding reference frames is a straightforward problem for simple geometric shapes such as a square or a rectangle. The problem becomes difficult for shapes that do not have a clear symmetry axis such as a notched rectangle (for some more examples see Figures 2.4, 2.9, 2.11, and 2.18), and none of the schemes presented previously can handle them successfully. Ultimately, we would like to achieve human-like performance. This is difficult partly because what humans consider to be a good skeleton can be influenced by high-level knowledge (see Figures 2.8 and 2.9).

2.4 Previous Work

The study of reference frames has received considerable attention in the computer vision literature. Reference frames have been used for different purposes (as discussed above) and given different names (e.g. skeletons, voronoi diagrams, symmetry transforms). Previous schemes for computing skeletons usually fall into one of two classes. The first class looks for a straight axis, such as the axis of inertia. These methods are global (the axis is determined by all the contour points) and produce a single straight axis. The second class can find a curved axis along the figure, but the computation is based on local information. That is, the axis at a given location is determined by small pieces of contours surrounding the location. Examples of such schemes are, to name but a few, Morphological Filters (see [Serra 1982], [Tsao and Fu 1984], [Meyer and Beucher 1990], [Maragos and Schafer 1986] for an overview), Distance Transforms [Rosenfeld and Pfaltz 1968], [Borgefors 1986], [Arcelli, Cordella, and Levialdi 1981], [Hansen 1992] Symmetric Axis Transforms [Blum 1967], [Blum and Nagel 1978] and Smoothed Local Symmetries [Brady and Asada 1984], [Connell and Brady 1987], [Rom and Medioni 1991]. Recently, computations based on physical models have been proposed by [Brady and Scott 1988] and [Scott, Turner, and Zisserman 1989]. In contrast, the novel scheme presented in this Chapter, which we call Curved Inertia Frames (C.I.F.), can extract curved symmetry axes, and yet use global information.

In most approaches, the compact and abstract description given by the reference frame is obtained after computing discontinuities. The version of C.I.F. presented in this Chapter also assumes a pre-computation of discontinuities. In other approaches, the abstract description given by the frame is computed simultaneously to find the frame or the description of the shape. Examples of such approaches include generalized cylinders [Binford 1971], [Nevatia and Binford 1977], [Marr and Nishihara 1978], [Brooks, Russell and Binford 1979], [Biederman 1985], [Rao 1988], [Rao and Nevatia 1988] superquadrics [Pentland 1988], [Terzopoulos and Metaxas 1986], [Metaxas 1992], extrema of curvature [Duda and Hart 1973], [Hollerbach 1975], [Marr 1977], [Binford 1981], [Hoffman and Richards 1984] to name a few. Similar computations have been used outside vision such as voronoi diagrams in robotics [Canny 1988] [Arcelli 1987]. An extension of C.I.F. to working without discontinuities is presented in Chapter 3.

2.4.1 Five problems with previous approaches

Previously presented computations for finding a curved axis generally suffer from one or more of the following problems: first, they produce disconnected skeletons for shapes that deviate from perfect symmetry or that have fragmented boundaries (see Figure 2.11); second, the obtained skeleton can change drastically due to a small change in the shape (e.g. a notched rectangle vs a rectangle as in Figure 2.11) making these schemes unstable; third, they do not assign any measure to the different components of the skeleton that indicates the "relative" relevance of the different components of the shape; fourth, a large number of computations depend on scale, introducing the problem of determining the correct scale; and fifth, it is unclear what to do with curved or somewhat-circular shapes because these shapes do not have a clear symmetry axis.

Consider for example, the Symmetric Axis Transform [Blum 1967]. The SAT of a shape is the set of points such that there is a circle centered at the point that is tangent to the contour of the shape at two points but does not contain any portion of the boundary of the shape, (see [Blum 1967] for details). An elegant way of computing the SAT is by using the brushfire algorithm which can be thought of as follows: A fire is lit at the contour of the shape and propagated towards the inside of the shape. The SAT will be the set of points where two fronts of fire meet. The Smoothed Local Symmetries [Brady and Asada 1984] are defined in a similar way but, instead of taking the center point of the circle, the point that lies at the center of the segment between the two tangent points is the one that belongs to the SLS and the circle need not be inside the shape. In order to compute the SAT or SLS of a shape we need to know the tangent along the contours of the shape. Since the tangent is a scale dependent measure so are the SLS and the SAT.

One of the most common problems (the first problem above) in skeleton finding computations is the failure to tolerate noisy or circular; this often results in disconnected and distorted frames. A notched rectangle is generally used to illustrate this point, see [Serra 1982], [Brady and Connell 1987] or [Bagley 1985] for some more examples. [Heide 1984], [Bagley 1985], [Brady and Connell 1987], [Fleck 1985, 1986, 1988], [Fleck 1989] suggest solvign this stability problem by working on the obtained SLS: eliminating the portions of it that are due to noise, connecting segments that come from adjacent parts of the shape and by smoothing the contours at different scales. In our scheme, symmetry gaps are closed automatically since we look for the largest scale available in the image, and the frame depends on all the contour, not just a small portion - making the scheme robust to small changes in the shape.

SAT and SLS produce descriptions for circular shapes which are not useful in general because they are simple and often fragemented. [Fleck 1986] addressed this problem by designing a separate computation to handle circular shapes, the Local Rotational Symmetries. C.I.F. can incorporate a preference for the vertical that will bias the frame towards a vertical line in circular shapes. When the shape is composed of a long straight body attached to a circular one (e.g. a spoon) then the bias will be towards having only one long axis in the direction of the body.

Figure 2.1: The objects in these Figures (their parts or the category to which they belong) are elongated and flexible.

Figure 2.2: Several objects which contain parts that are elongated and flexible.

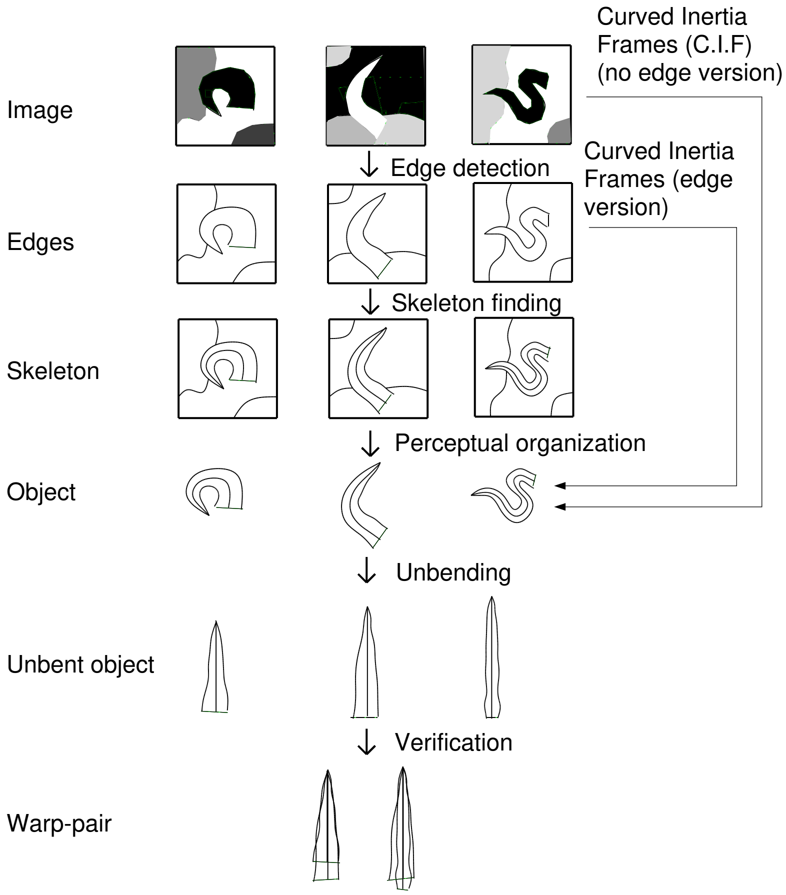


Figure 2.3: Summary of scheme proposed for the recognition of elongated flexible objects.

Figure 2.4: Which two of the three shapes on the left are more similar? One way of answering this question is by "unbending" the shapes using their skeleton as a reference frame, which results in the three shapes on the right. Once the shapes have been unbent, it can be concluded using simple matching procedures that two of them have similar "shapes" and that two others have similar length. We suggest that the recognition of elongated flexible objects can be performed by transforming the shape to a canonical form and that this transformation can be achieved by unbending the shape using its skeleton as an anchor structure. The unbending presented in this figure was obtained using an implemented lisp program (see also Figure 2.3).

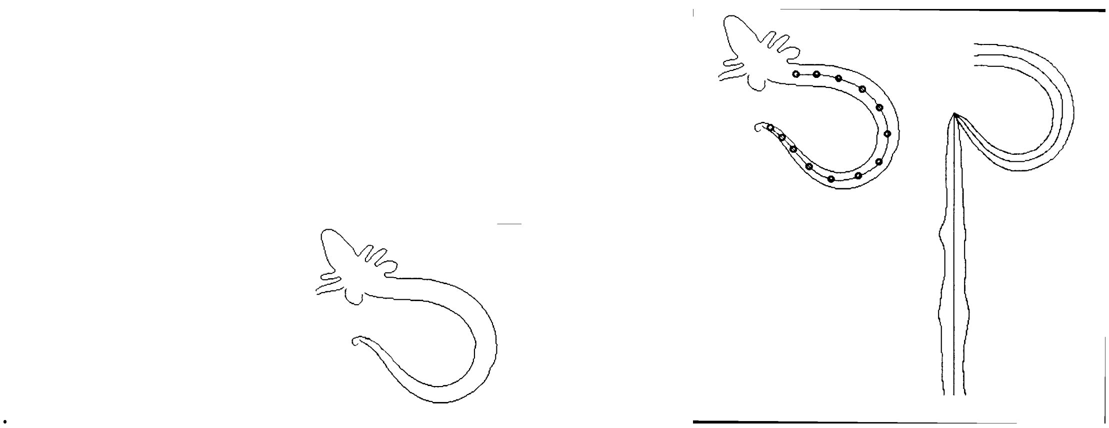


Figure 2.5: Example of the transformation suggested illustrated on a mudpupy. This figure shows the skeleton of the mudpupy and an "unbent" version of the mudpupy. The unbending has been done using an implemented lisp program.

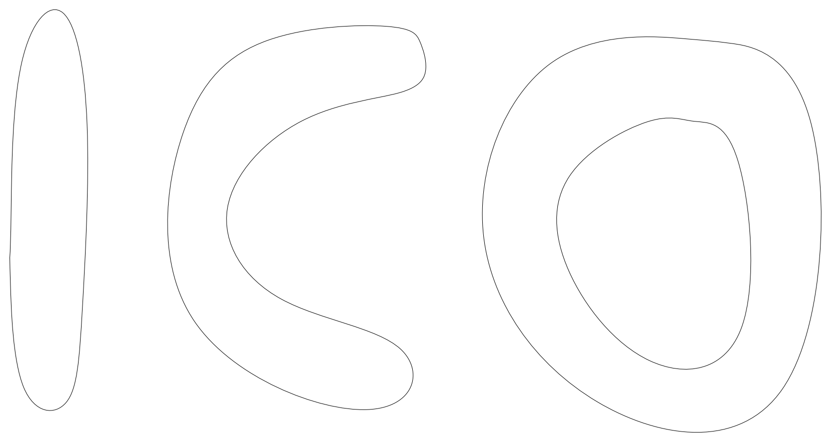


Figure 2.6: An elongated shape, when bent, gradualy develops a hole. This observation will lead us, in Appendix B to analyze the relation between holes, elongated and flexible objects, and perceptual organization.

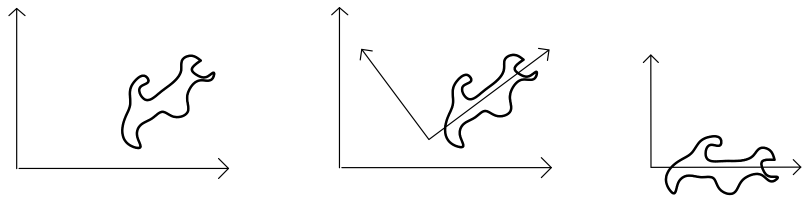


Figure 2.7: *Left:* a shape described in an image or viewer centered reference frame. *Center:* the same shape with an object centered reference frame superimposed on it. *Right:* a "canonical" description of the shape with horizontal alignment axis.

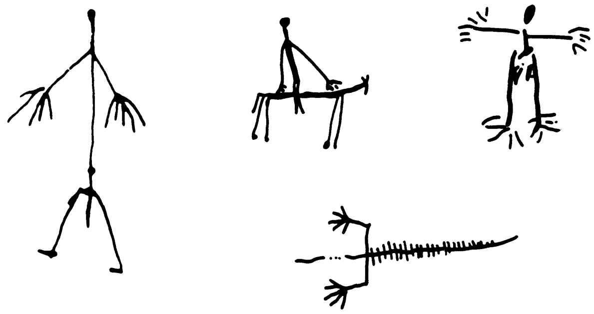


Figure 2.8: Drawings of a woman, a horse and a rider, a man, and a crocodile made by Tallensi tribes (adapted from [Deregowski 1989]). See also [Marr and Nishihara 1978].

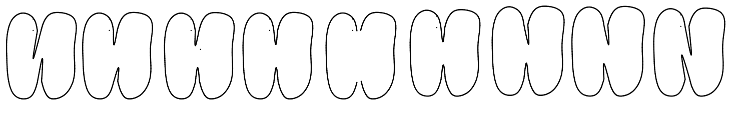


Figure 2.9: All the shapes in this Figure have been drawn by adding a small segment to the shape in the middle. At a first glance, all of these shapes would be interpreted as two blobs. But if we are told that they are letters, then finer distinctions are made between them. When we use such high level knowledge, we perceive these shapes as being different and therefore their associated skeletons would differ dramatically.

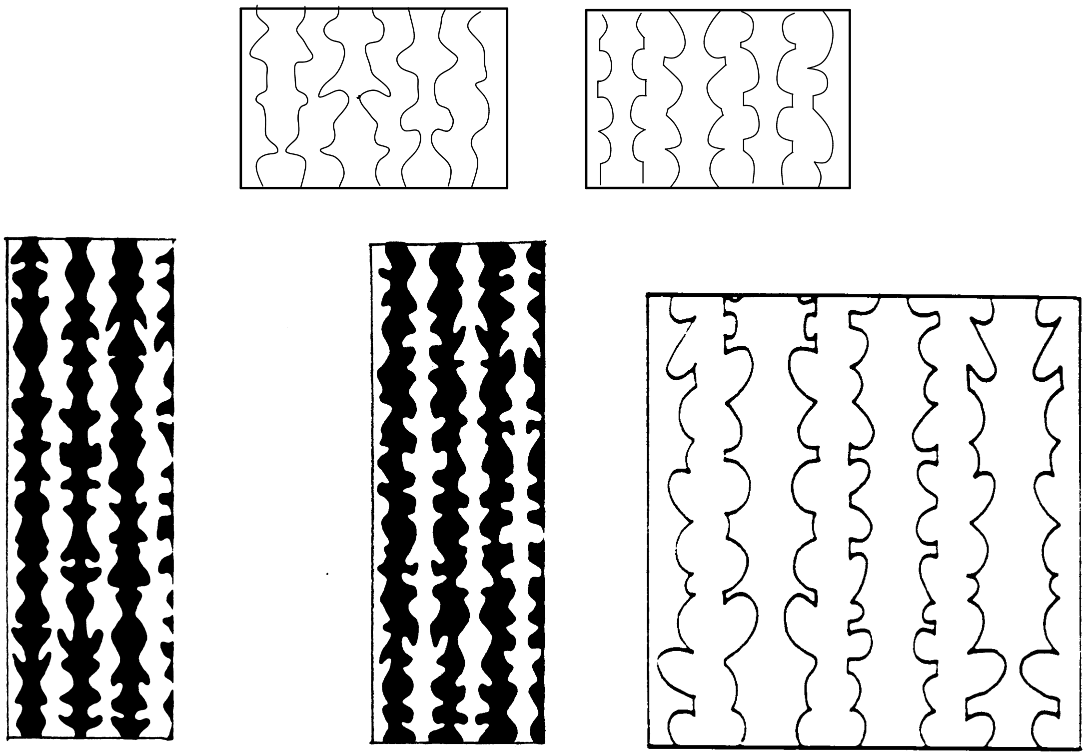


Figure 2.10: This Figure illustrates the importance of symmetry and convexity in grouping. The curves in the left image are grouped together based on symmetry. In the right image, convexity overrides symmetry, after [Kanizsa and Gerbino 76]. This grouping can be performed with the network presented in this Chapter by looking for the salient axes in the image.

2.5 Inertia Surfaces and Tolerated Length

If we are willing to restrict the frame to a single straight line, then the axis of least inertia is a good reference frame because it provides a connected skeleton and it can handle nonsymmetric connected shapes. The inertia $\mathbf{In}(SL, A)$ of a shape A with respect to a straight line SL is defined as (See Figure 2.12):

$$\mathbf{In}(SL, A) = \int_A \mathcal{D}(a, SL)^2 da \tag{2.1}$$

The integral is extended over all the area of the shape, and $\mathcal{D}(a, SL)$ denotes the distance from a point a of the shape to the line SL. The axis of least inertia of a shape A is defined as the straight line SL minimizing $\mathbf{In}(SL, A)$.

A naive way of extending the definition of axis of least inertia to handle bent curves would be to use Equation 2.1, so that the skeleton is defined as the curve C minimizing $\mathbf{In}(C, A)$. This definition is not useful if C can be any arbitrary curve because a highly bent curve, such as a space-filling-curve, that goes through all points inside the shape would have zero inertia (see Figure 2.13). There are two possible ways to avoid this problem: either we define a new measure that penalizes such curves or we restrict the set of permissible curves.

We chose the former approach and we call the new measure defined in this Chapter (see equation 4) the global curved inertia (inertia or curved inertia for short), or skeleton saliency of the curve. The curved inertia of a curve will depend on two local measures: the local inertia value $\mathcal{I}$ (inertia value for short) that will play a role similar to that of $\mathcal{D}(p, a)$ in equation 2.1 and the tolerated length $\mathcal{T}$ that will prevent non-smooth curves from receiving optimal values. The curved inertia of a curve will be defined for any curve C of length L starting at a given point p in the image. We define the problem as a maximization problem so that the "best" skeleton will be the curve that has the highest curved inertia. By "best" we mean that the skeleton corresponds to the "most central curve" in the "most interesting (i.e. symmetric, convex, large)" portion of the image.

Once a definition of inertia has been given, one needs an algoritm to compute the optimum curve. Unfortunately, such algorithm will be exponential in general because the number of possible curves is exponential in the size of the image. Therefore, one has to insure that the inertia function that one defines leads to a tractable algorithm to find the optimum curve.

2.5.1 *The inertia value*

The inertia measure $\mathcal{I}$ for a point p and an orientation α is defined as (see Figure 2.14):

$$\mathcal{I}(p, \alpha) = 2R\frac{(R-r)^s}{R^s},$$

Figure 2.14 shows how r, R, and the inertia surfaces are defined for a given orientation α. $R = d(p_l, p_r)/2$ and $r = d(p, p_c)$, where p_l and p_r are the closest points of the contour that intersect with a straight line perpendicular to α (i.e. with orientation $\alpha + \pi/2$) that goes through p in opposite directions and p_c is the midpoint of the interval between these two points. For a given orientation, the inertia values of the points in the image form a surface that we call the *inertia surface* for that orientation. Figure 2.13 illustrates why the inertia values should depend on the orientation of the skeleton, and Figure 2.15 shows the inertia surfaces for a square at eight orientations.

Local maxima on the inertia values for one orientation indicate that the point is centered in the shape at that orientation[4]. The value of the local maximum (always positive) indicates how large the section of the body is at that point for the given orientation, so that points in large sections of the body receive higher inertia values. The constant s or *symmetry constant*, 2 in the actual implementation, controls the decrease in the inertia values for points away from the center of the corresponding section, the larger s is the larger the decrease. If s is large only center points obtain high values, and if $s = 0$, all points of a section receive the same value.

[4]The inertia value can be seen as an "attractive potential" similar to artificial potentials for obstacle avoidance [Khatib 1986], [Khosla and Volpe 1988], [Hwang and Ahuja 1988].

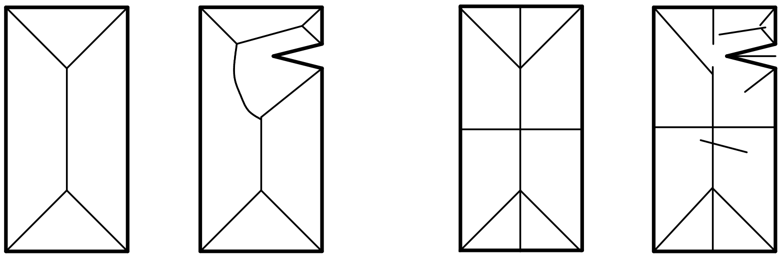


Figure 2.11: Skeletons found for a rectangle and a notched rectangle by SAT left and SLS right. Observe that the skeleton is highly distorted by the presence of the notch.

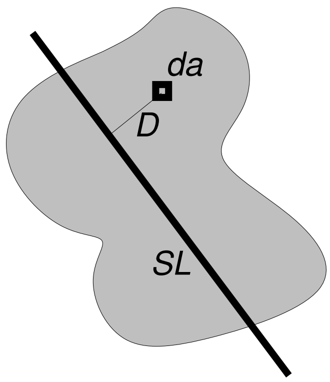


Figure 2.12: Definition of axis of inertia (see text for details).

2.5.2 *The tolerated length*

Figure 2.13 provides evidence that the curvature on a skeleton should depend on the width of the shape. As mentioned above, the tolerated length $\mathcal{T}$ will be used to evaluate the smoothness of a frame so that the curvature that is "tolerated" depends on the width of the section allowing high curvature only on thin sections of the shape. The skeleton inertia of a curve will be the sum of the inertia values "up to" the tolerated length so that for a high tolerated length, i.e. low curvature, the sum will include more terms and will be higher. The objective is that a curve that bends into itself within a section of the shape has a point within the curve with 0 tolerated length so that the inertia of the curve will not depend on the shape of the curve

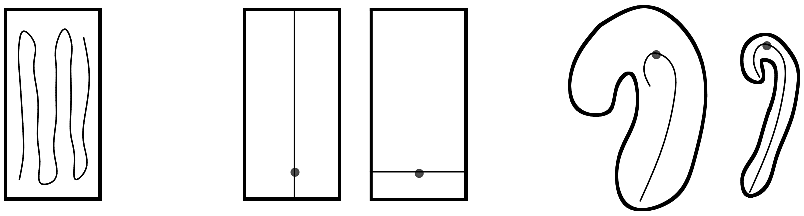


Figure 2.13: *Left:* A rectangle and a curve that would receive low inertia according to Equation 1. *Center:* Evidence that the inertia value of a point should depend on orientation. *Right:* Evidence that the tolerated curvature on a skeleton should depend on the width of the shape.

beyond that point. In other words, $\mathcal{T}$ should be 0 when the radius of curvature of the "potential" skeleton is smaller than the width of the shape at that point and a positive value otherwise (with an increasing magnitude the smoother the curve is).

We define the *tolerated length* $\mathcal{T}$ for a point with curvature of radius "r_c" as:

$$\mathcal{T}(p, \alpha, r_c) = \begin{cases} 0 & \text{if } r_c < R + r \\ r_c(\pi - arccos(\frac{r_c-(R+r)}{r_c})) & \text{otherwise} \end{cases}$$

If a curve has a point with a radius of curvature r_c smaller than the width of the shape its tolerated length will be 0 and this, as we will see, results in a non-optimal curve[5].

In this section we have introduced the inertia surfaces and the tolerated length. In Section 2.7 we give a formal definition of inertia. Intuitively, the definition is such that a curved frame of reference is a high, smooth, and long curve in the inertia surfaces (where smoothness is defined based on the tolerated length). Our approach is to associate a measure to any curve in the plane and to find the one that yields the highest possible value. The inertia value will be used to ensure that curves close to the center of large portions of the shape receive high values. The tolerated length will be used to ensure that curves bending beyond the width of the shape receive low

[5]Because of this, if a simply connected closed curve has a radius of curvature lying fully inside the curve then it will not be optimal. Unfortunately we have not been able to prove that any simply connected closed curve has such a point nor that there is a curve with such a point.

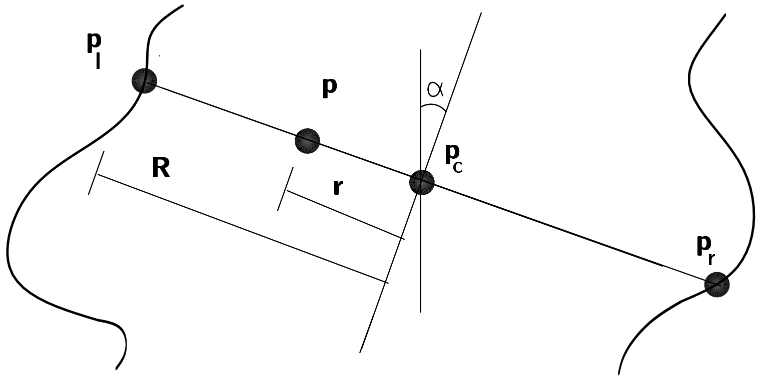


Figure 2.14: This Figure shows how the inertia surfaces are defined for a given orientation α. The value for the surface at a point p is $\mathcal{I}(R, r)$. The function $\mathcal{I}$ or *inertia function* is defined in the text. $R = d(p_l, p_r)/2$ and $r = d(p, p_c)$, where p_l and p_r are the points of the contour that intersect with a straight line perpendicular to α that goes through p at opposite directions and p_c is the midpoint of the interval between these two points. If there is more than one intersection along one direction, then we use the nearest one. If there is no intersection at all, then we give a preassigned value to the surface, 0 in the current implementation.

values. In the next section we will investigate how such a curve might be computed in a general framework and in section 2.7 we will see how to include the inertia values and the tolerated length in the computation, and what is the definition of the inertia measure that results.

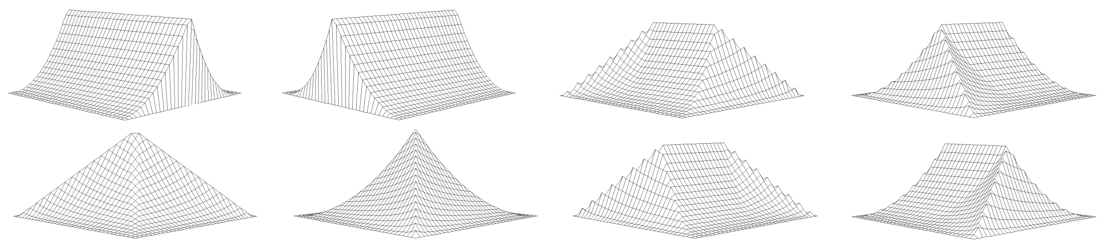


Figure 2.15: Plots of the inertia surfaces for a square for eight orientations

2.6 A Network to Find Frame Curves

In the previous sections, we have presented the Inertia Surfaces and the Tolerated Length. We concluded that skeletons could be defined as long and smooth curves with optimal inertia and tolerated length. In the following sections we will derive a class of dynamic programming algorithms[6] that find curves in an arbitrary graph that maximizes a certain quantity. We will apply these algorithms to finding skeletons in the inertia surfaces.

There has been some work which is relevant. [Mahoney 1987] showed that long and smooth curves in binary images are salient in human perception even if they have multiple gaps and are in the presence of other curves. [Sha'ashua and Ullman 1988] devised a saliency measure and a dynamic programming algorithm that can find such salient curves in a binary image (see also [Ullman 1976]). [Subirana-Vilanova 1990], [Freeman 1992], [Spoerri 1991], [Sha'ashua and Ullman 1991] have presented schemes which use a similar underlying dynamic programming algorithm.

However, all of the above schemes suffer from the same problem: the scheme has an orientation dependency such that certain orientations are favored. This problem was first mentioned in [Subirana-Vilanova 1991], [Freeman 1992]. Here, we will present a scheme which is not orientation dependent. In addition, our scheme perceives grouping of curves in a way which is remarkably similar to that of humans.

In this section we will examine the basic dynamic programing algorithm in a way geared toward demonstrating that the kind of measures that can be computed with the network is remarkably limited. The actual proof of this will be given in section 2.9.

We define a *directed graph with properties* $G = (V, E, P_E, P_J)$ as a graph with a set of *vertices* $V = \{v_i\}$; a set of edges $E \subset \{e_{i,j} = (v_i, v_j) \mid v_i, v_j \in V\}$; a function $P_E : E \rightarrow \Re^m$ that assigns a vector $\vec{p_e}$ of *properties* to each edge; and a function $P_J : J \rightarrow \Re^n$ that assigns a vector $\vec{p_j}$ of *properties* to each *junction* where a junction is a pair of adjacent edges (i.e. any pair of edges that share exactly one vertex) and J is the set of all junctions. We will refer to a curve in the graph as a sequence of

[6] See [Dreyfus 1965], [Gluss 1975] for an introduction to dynamic programming.

connected edges. We assume that we have an *inertia function* S that associates a positive integer $S(C)$ with each curve C in the graph. This integer is the *inertia* or *inertia value* of the curve. The inertia of a curve will be defined in terms of the properties of the elements (vertices, edges, and junctions) of the curve.

We will devise an iterative computation that finds for every point v_i and each of its connecting edges $e_{i,j}$, the curve with the highest inertia. Such a curve will start at the point v_i with the orientation of the edge $e_{i,j}$. This includes defining an inertia function and a computation that will find the frame curves for that function, i.e. those that have the highest values. The applications that will be shown here work with a 2 dimensional grid. The vertices are the points in the grid and the edges the elements that connect the different points in the grid. The junctions will be used to include in the inertia function *properties* of the shape of the curve, such as curvature.

The computation will be performed in a locally connected parallel network with a processor $pe_{i,j}$ for every edge $e_{i,j}$. The processors corresponding to the incoming edges of a given vertex will be connected to those corresponding to the connecting edges at that vertex. We will design the iterative computation so that we know, at iteration n and at each point, what the inertia of the optimum curve of lenght n is. That is the curve with highest inertia among those of size n that start at any given edge. This provides a constraint in the invariant of the algorithm that we are seeking that will guide us to the final algorithm. In order for the computation to have some computing power each processor $pe_{i,j}$ must have at least *one* state variable that we will denote as $s_{i,j}$. Since we want to know the highest inertia among all curves of length n starting with any given edge, we will assume that, at iteration n, $s_{i,j}$ contains that value for the edge corresponding to $s_{i,j}$.

Observe that having only one variable looks like a big restriction, however, we show in Section 2.9 that allowing more state variables does not add any power to the possible functions that can be computed with this network.

Since the inertia of a curve is defined only by the properties of the elements in the curve, it can not be influenced by properties of elements outside the curve. Therefore the computation to be performed can be expressed as:

$$s_{i,j}(n+1) = \mathrm{MAX}\{\mathcal{F}(n+1, \vec{p_e}, \vec{p_j}, s_{j,k}(n)) \mid \quad (j,k) \in E\}$$

$$s_{i,j}(0) = \mathcal{F}(0, \vec{p_e}, \vec{p_j}, 0) \tag{2.2}$$

where $\mathcal{F}$ is the function that will be computed in every iteration and that will lead to the computed total curved value. Observe that given $\mathcal{F}$, the global inertia value of any curve can be found by applying $\mathcal{F}$ recursively on the elements of the curve.

We are now interested in what types of functions S we can use and what type of functions $\mathcal{F}$ are needed to compute them such that the value obtained in the computation is the maximum for the resulting measure S. Using contradiction and induction, we conclude that a function $\mathcal{F}$ will compute the highest inertia for all possible graphs if and only if it is monotonically increasing in its last argument, i.e. iff

$$\forall \vec{p}, x, y \quad x < y \quad \longrightarrow \mathcal{F}(\vec{p}, x) < \mathcal{F}(\vec{p}, y), \tag{2.3}$$

where $\vec{p}$ is used to abbreviate the first three arguments of $\mathcal{F}$.

What type of functions $\mathcal{F}$ satisfy this condition? We expect them to behave freely as $\vec{p}$ varies. And when $s_{j,k}$ varies, we expect $\mathcal{F}$ to change in the same direction with an amount that depends on $\vec{p}$. A simple way to fulfill this condition is with the following function:

$$\mathcal{F}(\vec{p}, x) = f(\vec{p}) + g(x) * h(\vec{p}) \tag{2.4}$$

where f, g and h are positive functions, and g is monotonically increasing.

We now know what type of function $\mathcal{F}$ we should use but we do not know what type of measures we can compute. Let us start by looking at the inertia $s_{1,2}$ that we would compute for a curve of length i. For simplicity we assume that g is the identity function:

- **Iter. 1:** $s_{1,2}(1) = f(\vec{p_{1,2}})$
- **Iter. 2:** $s_{1,2}(2) = s_{1,2}(1) + f(\vec{p_{2,3}}) * h(\vec{p_{1,2}})$
- **Iter. 3:** $s_{1,2}(3) = s_{1,2}(2) + f(\vec{p_{3,4}}) * h(\vec{p_{1,2}}) * h(\vec{p_{2,3}})$
- **Iter. 4:** $s_{1,2}(4) = s_{1,2}(3) + f(\vec{p_{4,5}}) * h(\vec{p_{1,2}}) * h(\vec{p_{2,3}}) * h(\vec{p_{3,4}})$

  ...
- **Iter. i:** $s_{1,2}(i) = s_{1,2}(i-1) + f(\vec{p_{i,i-1}}) * \prod_{k=1}^{k=i-1} h(\vec{p_{k,k+1}}) = \sum_{l=1}^{l=i} f(\vec{p_{l,l-1}}) * \prod_{k=1}^{k=l-1} h(\vec{p_{k,k+1}})$.

At step n, the network will know about the "best" curve of length n starting from any edge. By "best" we mean the one with the highest inertia. Recovering the optimum curve from a given point can be done by tracing the links chosen by the processors (using Equation 2.2).

2.7 Computing Curved Inertia Frames

In this section, we will show how the network defined in the previous section can be used to find frames of reference using the inertia surfaces and the tolerated length as defined in Section 2.5. A directed graph with properties that defines a useful network for processing the image plane has one vertex for every pixel in the image and one edge connecting it to each of its neighbors, thus yielding a locally connected parallel network. This results in a network that has eight orientations per pixel.

The value computed is the sum of the $f(\vec{p_{i,j}})$'s along the curve weighted by the product of the $h(\vec{p_{i,j}})$'s. Using $0 \leq h \leq 1$ we can ensure that the total inertia will be smaller than the sum of the f's. One way of achieving this is by using $h = 1/k$ or $h = \exp(-k)$ and restricting k to be larger than 1. The f's will then be a quantity to be maximized and the k's a quantity to be minimized along the curve. In our skeleton network, f will be the inertia measure and k will depend on the tolerated length and will account for the shape of the curve so that the inertia of a curve is the sum of the inertia values along a curve weighted by a number that depends on the overall smoothness of the curve. In particular, the functions f, g and h (see Equation 2.4) are defined as:

- $f(\vec{p}) = f(\vec{p_e}) = \mathcal{I}(R, r)$,
- $g(x) = x$
- and $h(\vec{p}) = h(\vec{p_j}) = \rho^{\frac{l_{emt}}{\alpha \mathcal{T}(\vec{p_j})}}$.

α, which we call the "circle constant", scales the tolerated length, and it was set to 4 in the current implementation (because $4\ r\pi/2$ is the length of the perimeter of a circle - where r is the radius of the circle). ρ, which we call the "penetration factor", was set to 0.5 (so that inertia values "half a circle" away get factored down by 0.5). And l_{emt} is the length of the corresponding element. Also, $s_{i,j}(0) = 0$ (because the inertia of a skeleton of length 0 should be 0).

With this definition, the inertia assigned to a curve of length L is:

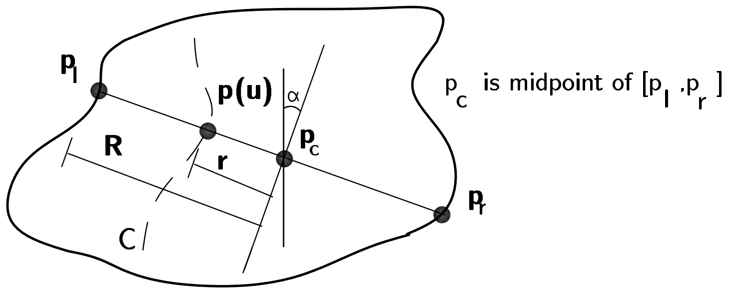
$$S_L = \sum_{l=1}^{l=L} \mathcal{I}(\vec{p_{l,l-1}}) \prod_{k=1}^{k=l-1} \rho^{\frac{l_{emt}}{\alpha \mathcal{T}(\vec{p_k})}} = \sum_{l=1}^{l=i} \mathcal{I}(\vec{p_{l,l-1}}) \rho^{\sum_{k=1}^{k=l-1} \frac{l_{emt}}{\alpha \mathcal{T}(\vec{p_k})}},$$

which is an approximation of the continuous value given in Equation 2.5 below (see also Figure 2.16):

$$S(C) = \int_0^L \mathcal{I}(l) \rho^{\int_0^l \frac{1}{\alpha \mathcal{T}(t)} dt} dl \qquad (2.5)$$

Where S_C is the inertia of a parameterized curve $C(u)$, and $\mathcal{I}(u)$ and $\mathcal{T}(u)$ are the inertia value and the tolerated length respectively at point u of the curve.

The obtained measure favors curves that lie in large and central areas of the shape and that have a low overall internal curvature. The measure is bounded by the area



Symmetry constant $s \in [0,\infty)$

Inertia Value $I(p(u)) = 2R \dfrac{(R-r)^s}{R^s}$

Tolerated length $T(p(u)) = \begin{cases} 0 \text{ if } r_c < r+R \text{ with } r_c = \left(\dfrac{\delta^2 p(u)}{\delta u^2}\right)^{-1} \\ r_c(\pi - \arccos(1 - \frac{r}{r_c})) \text{ otherwise} \end{cases}$

Penetration constant $\rho \in (0,1]$

<u>Curved Inertia of C</u> $S(C) = \int_0^L I(p(u))\, \rho^{\int_0^u \frac{1}{T(p(t))} \delta t}\, \delta u$

Figure 2.16: Definition of curved inertia used by current implementation of Curved Inertia Frames (see text for details).

of the shape: a straight symmetry axis of a convex shape will have an inertia equal to the area of the shape. In the next section we will present some results showing the robustness of the scheme in the presence of noisy shapes.

Observe that if the tolerated length $\mathcal{T}(t)$ at one point $C(t)$ is small then $\int_0^l \frac{1}{\alpha \mathcal{T}(t)} dt$ is large so that $\rho^{\int_0^l \frac{1}{\alpha \mathcal{T}(t)} dt} dl$ becomes small (since $\rho < 1$) and so does the inertia for the curve S_L. Thus, small values of α and ρ penalize curvature favoring smoother curves.

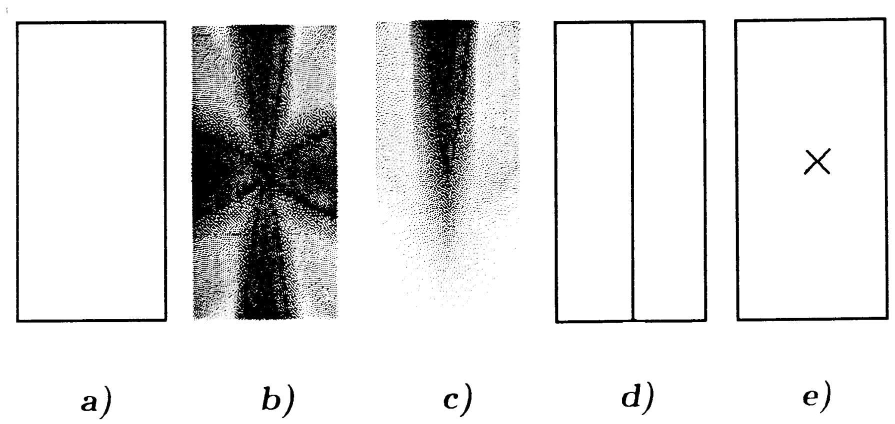


Figure 2.17: a) Rectangle. b) Skeleton sketch for the rectangle. Circles along the contour indicate local maxima in the skeleton sketch. c) Skeleton sketch for the rectangle for one particular orientation, vertical-down in this case. d) Most salient curve. e) Most interesting point for the most salient curve.

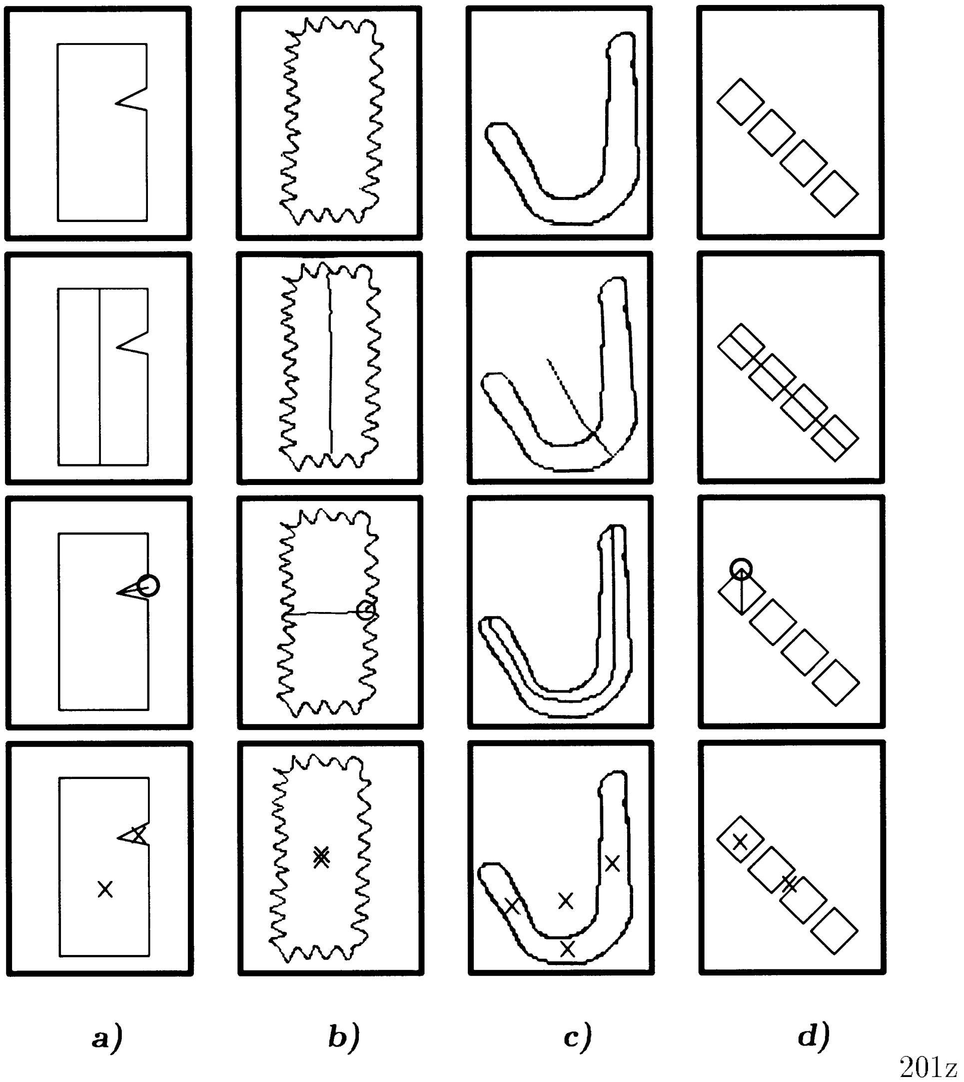


Figure 2.18: *Top:* Four shapes: a notched square, a stamp, a J, and Mach's demonstration. *Second row:* The most salient curve found by the network for each of them. Observe that the scheme is remarkably stable under noisy or bent shapes. *Third row:* The most salient curve starting inside the shown circles. For the J shape the curve shown is the most salient curve that is inside the shape. *Fourth row:* The most interesting point according to the curves shown in the two previous rows. See text for details.

2.8 Results and Applications

In this section we will present some results and applications of the frame computation, and in the following sections we will discuss the limitations of the skeleton network and ways to overcome them.

The network described in the previous section has been implemented on a Connection Machine and tested on a variety of images. As mentioned above, the implementation works in two stages. First, the distance to the nearest point of the shape is computed at different orientations all over the image so that the inertia surfaces and the tolerated length can be computed; this requires a simple distance transform of the image. In the second stage, the network described in section 2.7 computes the inertia of the best curve starting at each point in the image. This is done at eight orientations in our current implementation. The number of iterations needed is bounded by the length of the longest optimal curve but in general a much smaller number of iterations will suffice. In all the examples shown in this Section the images were 128 by 128 pixels and 128 iterations were used. However, in most of the examples, the results do not change after about 40 iterations. In general, the number of iterations needed is bounded by the width of the shape measured in pixels.

2.8.1 *The Skeleton Sketch and the highest inertia skeleton:*

The *skeleton sketch* contains the inertia value for the most salient curve at each point. The skeleton sketch is similar to the saliency map described in [Sha'ashua and Ullman 1988] and [Koch and Ullman 1985] because it provides a saliency measure at every point in the image. It is also related to other image-based representations such as the 2 1/2D sketch [Marr 1982] or the 2-D Scale-Space Image [Saund 1990]. Figure 2.17 shows the skeleton sketch for a rectangle. The best skeleton can be found by tracing the curve starting at the point having the highest skeleton inertia value. Figure 2.18 shows a few shapes and the most salient curve found by the network for each of them. Observe that the algorithm is robust in the presence of non-smooth contours. Given a region in the image we can find the best curve that starts in the region by finding the maxima of the skeleton sketch in the region, see Figure 2.18. In general, any local maximum in the skeleton sketch corresponds to a curve accounting for a symmetry in the image. Local maxima in the shape itself are particularly interesting since they correspond to features such as corners (See Figure 5.12).

2.8.2 *The most "central" point:*

In some vision tasks, besides being interested in finding a salient skeleton, we are interested in finding a particular point related to the curve, shape, or image. This can be due to a variety of reasons: because it defines a point in which to start subsequent processing to the curve, or because it defines a particular place in which to shift our window of attention. Different points can be defined; the point with the highest inertia is one of them, because it can locate relevant features such as corners.

Another interesting point in the image is the "most central" point in a curve which can be defined by our scheme by looking for the inertia along the curve at both directions within the curve. The most central point can be defined as the point where these two values are "large and equal". The point that maximizes $\min(p_l, p_r)$ has been used in the current implementation[7]. See Figure 2.18 for some examples. Observe in Figure 2.18 that a given curve can have several *central points* due to different local maxima.

Similarly, the most central point in the image can be defined as the point that maximizes $\min(p_l, p_r)$ for all orientations.

2.8.3 *Shape description:*

Each locally salient curve in the image corresponds to a *symmetric region* in one portion of the scene. The selection of the set of most interesting frames corresponding to the different parts of the shape yields a part description of the scene. Doing this is not trivial (See [Shashua and Ullman 1990]) because a salient curve is surrounded by

[7]See also [Reisfeld, Wolfson, and Yeshurun 1988] where a scheme to detect interest points was presented. Their scheme is scale dependent contrary to C.I.F. which selects the larger structure as the most interesting one, independently of the scale at which the scene is seen.

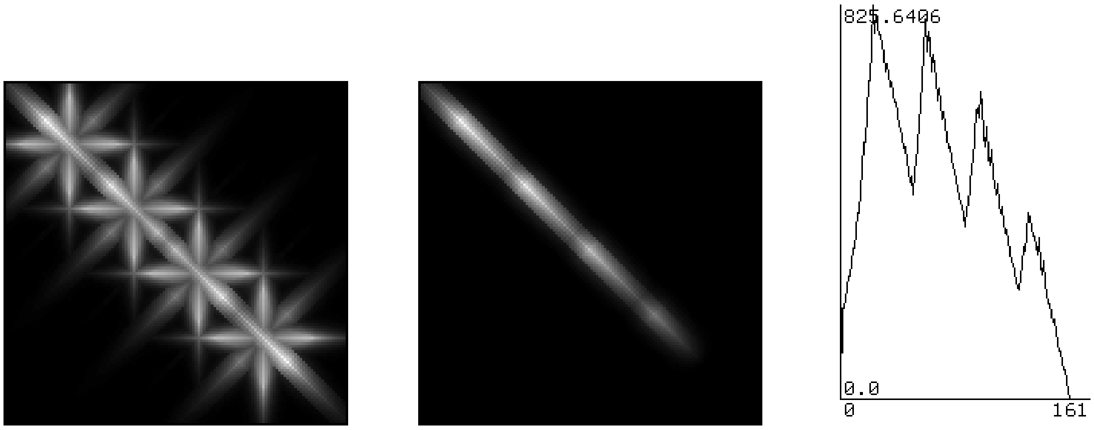


Figure 2.19: *Left:* Skeleton Sketch for Mach's demonstration (Original image in previous Figure top right). *Center:* Skeleton Sketch for *one orientation only.* *Right:* Slice of the "one-orientation" Skeleton Sketch through one of the diagonals of the image. Note that the values decrease across the gaps and increase inside the square (see also [Palmer and Bucher 1981]).

other curves of similar inertia. In general, a curve displaced one pixel to the side from the most salient curve will have an inertia value similar to that of the most salient one and higher than that of other locally most salient curves. In order to inhibit these curves, we color out from a locally maximal curve in perpendicular directions (to the axis) to suppress parallel nearby curves. The amount to color can be determined by the average width of the curve. Once nearby curves have been suppressed we look for the next most salient curve and reiterate this process. Another approach to finding a group of several curves, not just one, is given in [Sha'ashua and Ullman 1990]. Both approaches suffer from the same problem: the groups obtained do not optimize a simple global maximization function.

Figure 2.20 shows the skeleton found for an airplane. The skeleton can then be used to find a part description of the shape in which each component of the frame has different elements associated that describe it: a set of contours from the shape, an inertia measure reflecting the relevance or saliency that the component has within the shape, a central point, and a location within the shape.

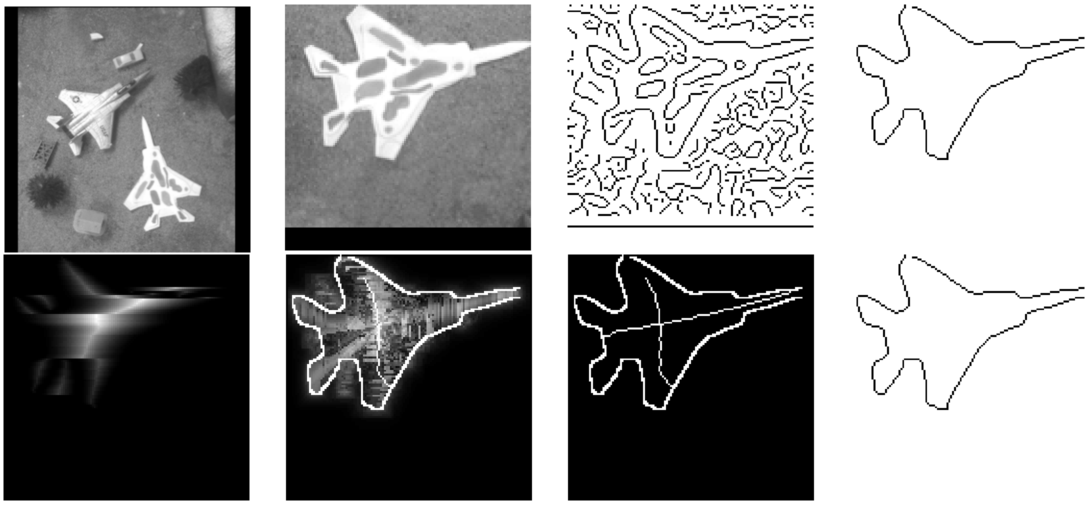


Figure 2.20: *Top:* Image; airplane portion enlarged; its edges; airplane without short edges. *Bottom:* Vertical inertia surface; skeleton sketch; skeleton; most salient point.

2.8.4 *Inside/Outside:*

The network can also be used to determine a continuous measure of inside-outside [Shafrir 1985], [Ullman 1984] (see also [Subirana-Vilanova and Richards 1991], and Appendices A and B. The distance from a point to the frame can be used as a measure of how near the point is to the outside of the shape. This measure can be computed using a scheme similar to the one used to inhibit nearby curves as described in the previous paragraph: coloring out from the frame at perpendicular orientations, and using the stage at which a point is colored as a measure of how far from the frame the point is. The inertia of a curve provides a measure of the area swept by the curve which can be used to scale the coloring process.

2.9 Limitations of Dynamic Programming Approach

In this section we show that the set of possible inertia measures that can be computed with the network defined in Section 2.6 is limited.

Proposition 1 *The use of more than one state variable in the inertia network defined in section 2.6 does not increase the set of possible functions that can be optimized with the network.*

Proof: The notation used in the proof will be the one used in section 2.6. We will do the proof for the case of two state variables; the generalization of the proof to more state variables follows naturally. Each edge will have an inertia state variable $s_{i,j}$, an auxiliary state variable $a_{i,j}$, and two functions to update the state variables:

$$s_{i,j}(n+1) = MAX_k \mathcal{F}(\vec{p}, s_{j,k}(n), a_{j,k}(n))$$
and
$$a_{i,j}(n+1) = \mathcal{G}(\vec{p}, s_{j,k}(n), a_{j,k}(n)).$$

We will show that, for any pair of functions $\mathcal{F}$ and $\mathcal{G}$, either they can be reduced to one function or there is an initialization for which they do not compute the optimal curves.

If $\mathcal{F}$ does not depend on its last argument $a_{j,k}$, then the decision of what is the most salient curve is not effected by the introduction of more state variables (so we can do without them). Observe that we might still use the state variables to compute additional properties of the most salient curve without effecting the actual shape of the computed curve.

If $\mathcal{F}$ does depend on its last argument then there exists some $\vec{p}$, x, y and $w \in \Re$ such that:

$$\mathcal{F}(\vec{p}, y, x) < \mathcal{F}(\vec{p}, y, w).$$

Assuming continuity, this implies that there exists some $\epsilon > 0$ such that:

$$\mathcal{F}(\vec{p},\ y - \epsilon\,, x) < \mathcal{F}(\vec{p}, y, w).$$

Assume now two curves of length n starting from the same edge $e_{i,j}$ such that $s1_{i,j}(n) = y$, $a1_{i,j}(n) = x$, $s2_{i,j}(n) = y - \epsilon$ and $a2_{i,j}(n) = y$. If the algorithm were correct at iteration n it would have computed the values $s1_{i,j}(n) = y$, $a1_{i,j}(n) = x$ for the variables $s_{i,j}$ and $a_{i,j}$. But then at iteration $n+1$ the inertia value computed for an edge $e_{h,i}$ would be $s_{h,i} = \mathcal{F}(\vec{p},\ y - \epsilon\,, x)$ instead of $\mathcal{F}(\vec{p}, y, w)$ that corresponds to a curve with a higher inertia value. □

2.10 Non-Cartesian Networks

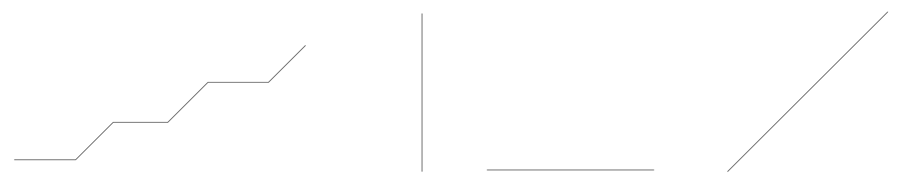


Figure 2.21: For the right-most three curves, curvature can be computed localy (it is 0 everywhere). However, the curve on the left is a discretization on a cartesian network of a line at about 60 degrees. Is also has a zero curvature but it can not be computed localy since local estimates would be biased towards the network orientations.

Straight lines that have an orientation different from one of the eight network orientations generate curvature impulses due to the discretization imposed on them, essentially 45 or 90 degrees. Such impulses are generated in a number of pixels, per unit length, which can be arbitrarily large depending on the resolution of the grid. This results in a reduction of the inertia for such curves, biasing the network towards certain orientations (see Figure 2.21).

Cartesian networks such as these are used in most vision systems. They suffer from three key problems:

- Orientations bias: Certain orientations are different from the rest (typically 0, +/- 45, and 90 degrees). This implies that the result of most schemes depends on how the image array is aligned with the scene. This is also true for hexagonal grids.
- Length bias: Segments at 45 degrees have a longer length than the others.
- Uniform distribution: All areas of the visual array receive an equal amount of processing power.

To prevent the orientation bias, and as mentioned above, we made an implementation of the network that included a smoothing term that enabled the processors to change their "orientation" at each iteration instead of keeping only one of the eight initial orientations. At each iteration, the new orientation is computed by looking at nearby pixels of the curve which lie on a straight line (so that curvature is minimized). This increases the resolution but requires aditional state variables to memorize the orientation of the computed curve.

This allows greater flexibility but at the expense of breaking the optimization relation shown in Equation 2.3 (since additional state variables are used). Note that the value computed for the obtained curve corresponds to its true value. However, the obtained curve is not guaranteed to be the optimum. A similar problem is encountered with the smoothing terms used by [Sha'ashua and Ullman 1988], [Subirana-Vilanova 1990], [Spoerri 1990], [Subirana-Vilanova and Sung 1992], and [Freeman 1992].

One possible solution to this problem is to change the topology of the network (a key insight of this thesis!). Figure 2.22 presents an example in which the network is distributed at different orientations. This solves the orientation problem. If processors are placed with a fixed given size, then the size bias is also solved. However, constant size would result in few intersections (if any). The number of intersections can be increased by relaxing the size requirement.

Another solution is the use of random networks as shown in Figure 2.23. The surprising finding is that the corresponding implementation on a SPARC 10 is about 50 times faster than the CM-2 implementation presented in this thesis. Some results are presented in Figures 2.24, 2.25, and 2.26.

By allowing processors to vary their width 20 %, the expected number of intersections across lines is 1/5 the number of lines. The expected distance between an arbitrary line and the closest network line can also be estimated easily using polar coordinates.

In addition, lines can be concentrated in a certain region (or focus of attention) whose shape may be tuned to the expected scene (e.g. dynamic programming on a combination of circles and lines)[8].

2.11 Review

In this Chapter we have presented C.I.F. (Curved Inertia Frames), a novel scheme to compute curved symmetry axes. The scheme can recover provably global and curve axis and is based on novel non-cartesian networks.

In the next Chapter we will show how C.I.F. can be used to work directly in the image (without edges). The extension is based on a multi-scale vector ridge detector that computes tolerated length and inertia values directly for the image (i.e. without the need for discontinuities and distance transforms).

[8]We are currently working on extending these ideas together with S. Casadei [Subirana-Vilanova and Casadei 1993].

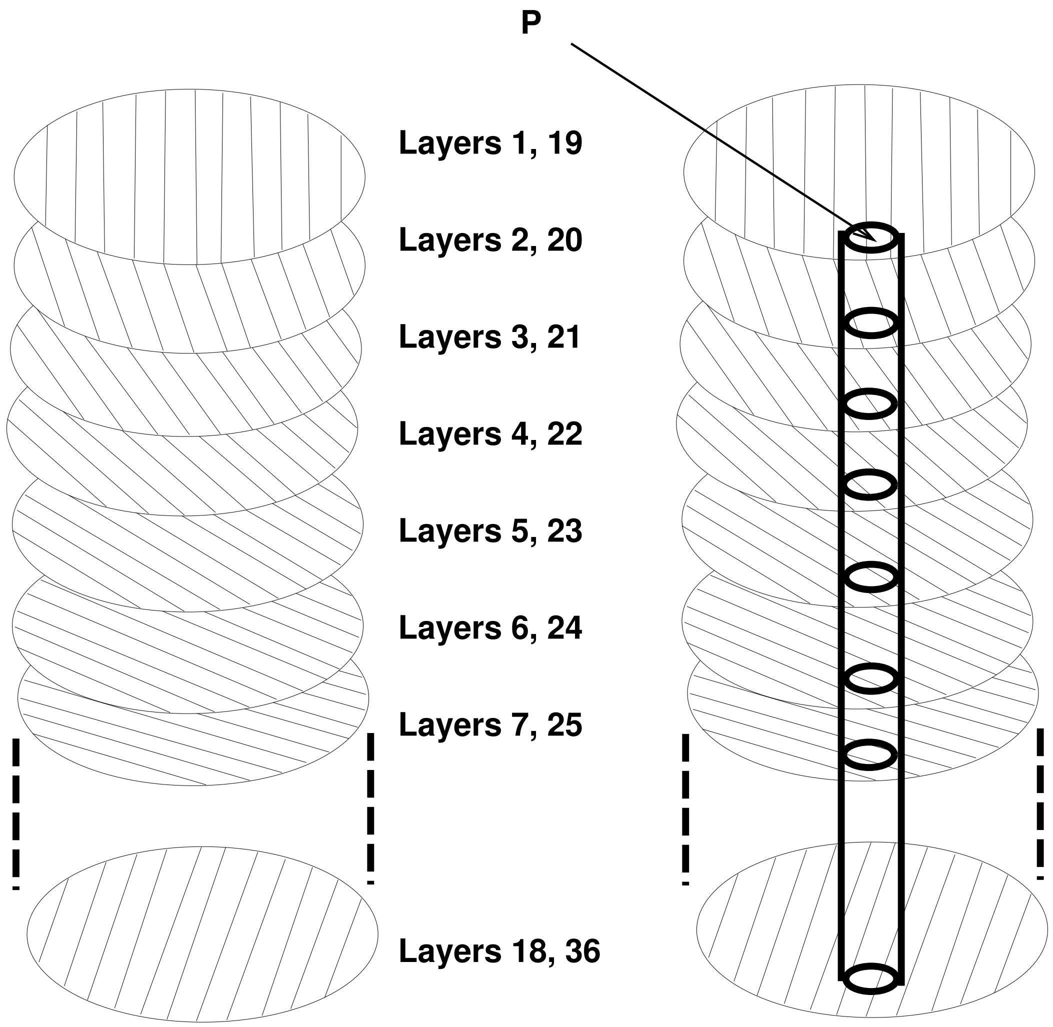
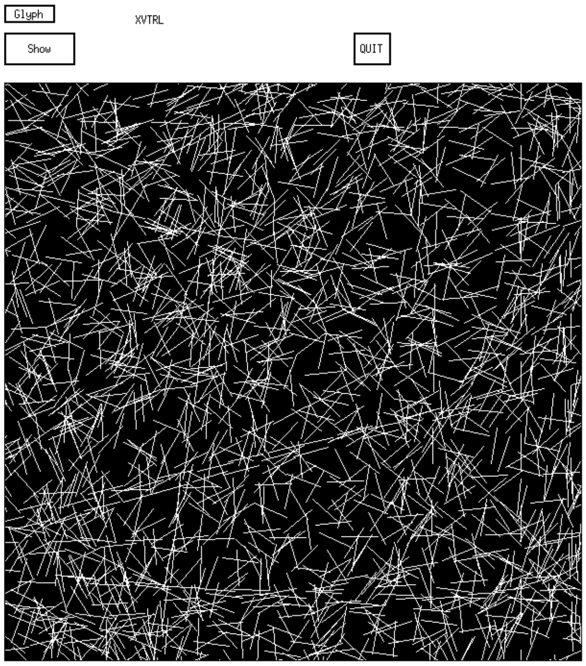


Figure 2.22: *Left:* The network outlined in this Figure avoids smoothing problems because all the elements have the same length and because it has many orientations. *Right:* The connectivity among processors is determined by connecting all processors that fall within a certain cylinder as depicted in this diagram.



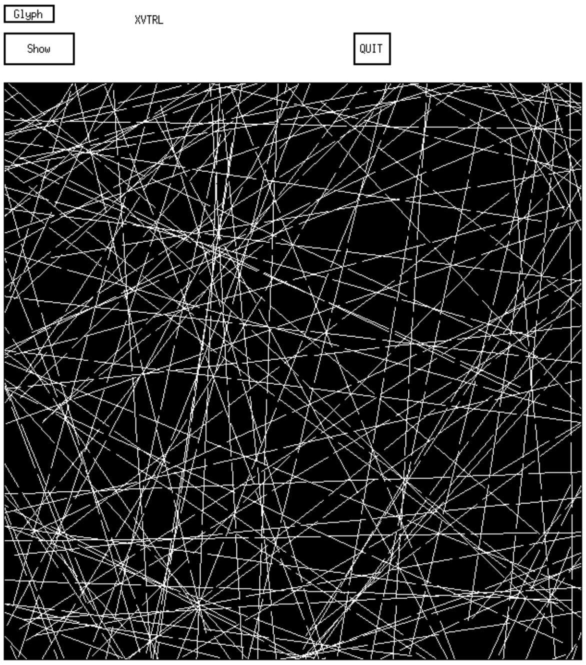


Figure 2.23: Random networks have not been used in vision and provide an elegant and efficient solution to local smoothing problems.

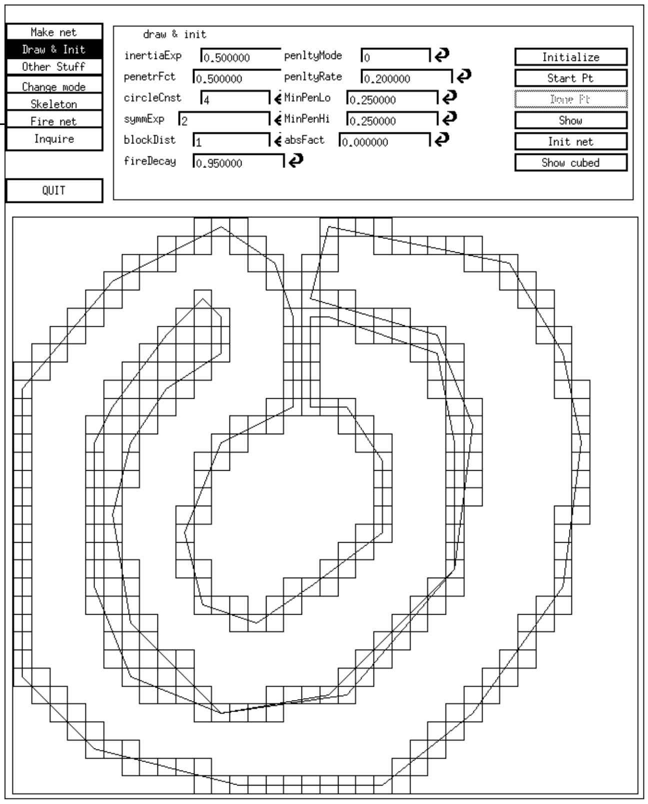


Figure 2.24: The continuous line in this drawing is the one used in the experiment reported in Figure 2.25. The squares show the discretization grid used. As can be seen, the inertia value resolution is low. In fact the pixel distance along the best skeleton's path is at most 5! The rest of the Figure shows part of the interface used in the random network simulator.

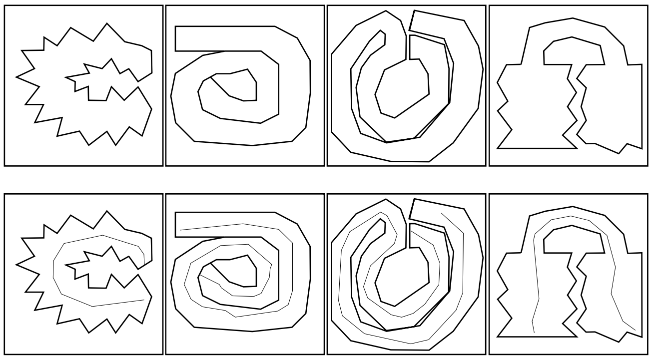


Figure 2.25: *Top:* Drawings. *Right:* Best skeleton obtained with SPARC simulation of 1600 line random network. Initialization was performed with only 32 pixels as shown in Figure 2.24. The processor size is 1/32 of the image's width. In all cases, the best solution was obtained after no more than 20 iterations.

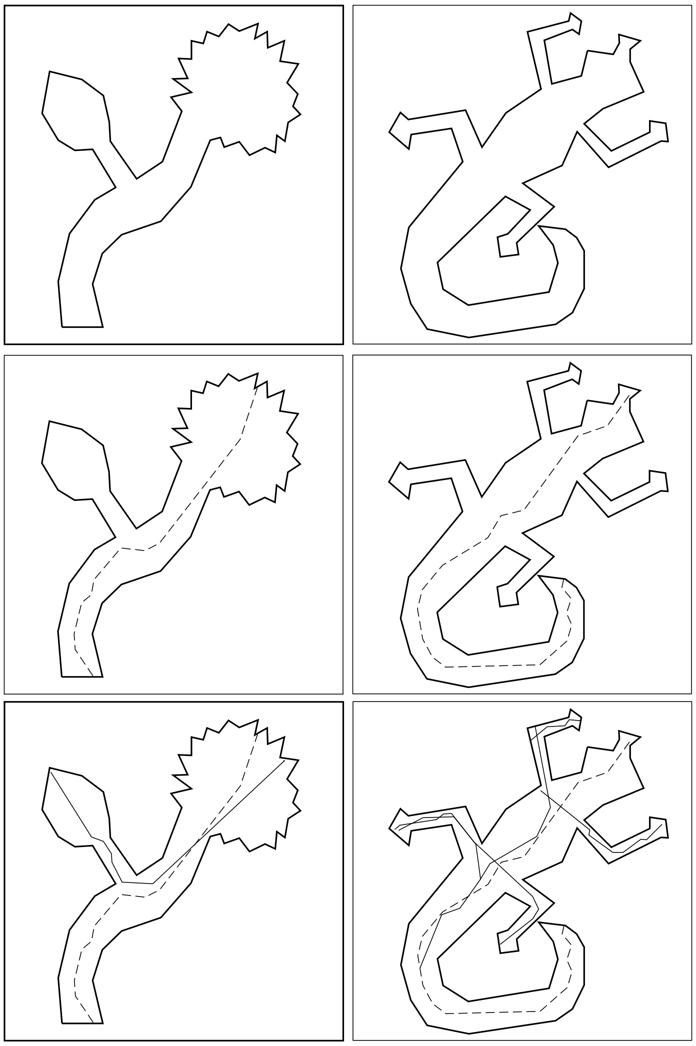


Figure 2.26: *First row:* Drawings of flower and newt. *Second row:* Best skeleton computed by C.I.F. The random network used had 128 pixels in the initialization stage, and 2400 lines in the flower and 3200 lines in the newt. *Third row:* The skeleton sketch has been used to find a complete description of the object. New optimal skeletons have been computed among those in regions without skeletons. This process has been repeated until there were no empty regions left inside the shape.

Non-Rigid Perceptual Organization Without Edges

Chapter 3

3.1 Introduction

In this chapter we continue the study of computations that can recover axes for mid-level vision and unbending elongated and flexible shapes. In Chapter 2 we presented a version of Curved Inertia Frames (C.I.F.) that works on the edges or discontinuities of a shape. In this chapter we will show how the robustness of Curved Inertia Frames can be increased by incorporating information taken directly from the image. This also means that C.I.F. will be able to compute several mid-level properties without the need of explicitly computing discontinuities.

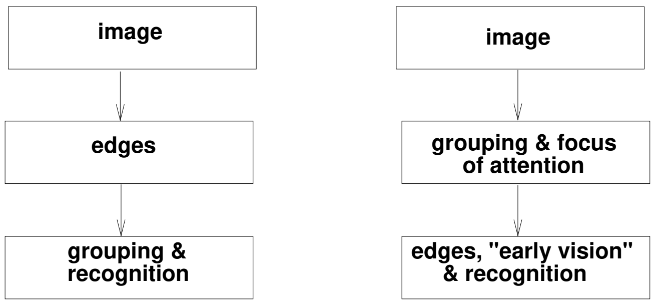


Figure 3.1: Two different views on the role of perceptual organization.

Recent work in computer vision has emphasized the role of edge detection and discontinuities in perceptual organization (and recognition). This line of research stresses that edge detection should be done at an early stage on a brightness representation of the image, and segmentation and other early vision modules computed later on (see Figure 3.1 left). We (like some others) argue against such an approach and present a scheme that segments an image without finding brightness, texture, or color edges (see Figure 3.1 right). In our scheme, C.I.F., discontinuities and a potential focus of attention for subsequent processing are found as a byproduct of the perceptual organization process which is based on a novel ridge detector introduced in this Chapter.

Segmentation without edges is not new. Previous approaches fall into two classes. Algorithms in the first class are based on coloring or region growing [Hanson and Riseman 1978], [Horowitz and Pavlidis 1974], [Haralick and Shapiro 1985], [Clements 1991]. These schemes proceed by laying a few "seeds" in the image and then "grow" these until a complete region is found. The growing is done using a local threshold function, i.e. decisions are made based on local neighborhoods. This results in schemes limited in two ways: first, the growing function does not incorporate global factors, resulting in fragmented regions (see Figure 3.2). Second, there is no way to incorporate *a priori* knowledge of the shapes that we are looking for. Indeed, important Gestalt principles such as symmetry, convexity and proximity (extensively used by current grouping algorithms) have not been incorporated in coloring algorithms. These principles are useful heuristics to aid grouping processes and are often sufficient to disambiguate between alternative interpretations. In this chapter, we present a non-local perceptual organization scheme that uses no edges and that embodies these gestalt principles. It is for this reason that our scheme overcomes some of the problems with region growing schemes – mainly the fragmenting of regions and the merging of overlapping regions with similar region properties.

The second class of segmentation schemes which work without edges is based on computations that find discontinuities while preserving some region properties such as smoothness or other physical approximations [Geman and Geman 1984], [Terzopoulos 86], [Blake and Zisserman 1987], [Hurlbert and Poggio 1988], [Poggio, Gamble and Little 1988], [Trytten and Tüceryan 1990], [Zerubia and Geiger 1991]. These schemes are scale dependent and in some instances depend on reliable edge

Figure 3.2: *(From top-left to bottom right) 1:* Full shirt image. *2:* Canny edges. *3:* Color edges. *4:* An image of a shirt. *5:* Original seeds for a region growing segmentation algorithm. *6:* Final segmentation obtained using a region growing algorithm.

detection to estimate the location of discontinuities. Scale at the discontinuity level has been addressed previously [Witkin 1983], [Koenderink 1984], [Perona and Malik 1990], but these schemes do not explicitly represent regions; often, meaningful regions are not fully enclosed by the obtained discontinuities. As with the previous class, all these algorithms do not embody any of the Gestalt principles and in addition perform poorly when there is a nonzero gradient inside a region. The scheme presented in this Chapter performs perceptual organization (see above) and addresses scale by computing the largest scale at which a structure (not necessarily a discontinuity) can be found in the image.

3.2 In Favor of Regions

What is an edge? Unfortunately there is no agreed definition of it. An edge

Figure 3.3: *Left:* Model of an edge. *Right:* Model of a ridge or box. Are these appropriate?

Figure 3.4: *Left:* Zero-crossings. *Right:* Sign bit. Which one of these is harder to recognize? (Taken from [Marr and Hildreth 1980]).

can be defined in several related ways: as a discontinuity in a certain property[1], as "something" that looks like a step edge (e.g. [Canny 1986] - see Figure 3.3) and by an algorithm (e.g. zero-crossings [Marr and Hildreth 1980], optical filtering [Vallmitjana, Bertomeu, Juvells, Bosch and Campos 1988]). Characterizing edges has proven to be difficult especially near corners, junctions[2], [Cheng and Hsu 1988], [Beymer 1991], [Giraudon and Deriche 1991], [Korn 1988], [Noble 1988], [Gennert 1986], [Singh and Shneier 1990], [Medioni and Yasumoto 1987], [Harris and Stephens 1988] and when the image contains noise, transparent surfaces, edges at multiple scales, or edges different than step edges (e.g. roof edges) [Horn 1977], [Ponce and Brady 1985], [Forsyth and Zisserman 1989], [Perona and Malik 1990].

What is a region? Problems similar to those encountered in the definition of an edge are found when attempting to define regions. Roughly speaking, an image region is a collection of pixels sharing a common property. In this context, an edge is the border of a region. How can we find regions in images? We could proceed in a way similar to that used with edges, so that a region is defined (in one dimension) as

[1] Note that, strictly speaking, there are no discontinuities in a properly sampled image (or that they are present at every pixel)

[2] Junctions are critical for most edge-labeling schemes which do not tolerate robustly missing junctions.

a structure that looks like a box (see Figure 3.3). However, this suffers from problems similar to the ones mentioned for edges.

Thus, regions and edges are two closely related concepts. It is unclear how we should represent the information contained in an image. As regions? As edges? Most people would agree that a central problem in visual perception is finding the objects or structures of interest in an image. These can be defined sometimes by their boundaries, i.e. by identifying the relevant edges in an edge-based representation. However, consider now a situation in which you have a transparent surface as when hair occludes a face, when the windshield in your car is dirty, or when you are looking for an animal inside the forest. An edge-based representation does not deal with this case well, because the region of interest is not well defined by the discontinuities in the scene but by the perceived discontinuities. This reflects an object-based view of the world. Instead, a region-based representation is adequate to represent the data in such an image.

Furthermore, independently of how we choose to represent our data, which structures should we recover first? Edges or regions? Here are four reasons why exploring the computation of regions (without edges) may be a promising approach:

3.2.1 *Human perception*

There is some psychological evidence that humans can recognize images with region information better than they recognize line drawings [Cavanaugh 1991]. However, there is not a clear consensus [Ryan and Schwartz 1956], [Biederman and Ju 1988] (see also Figure 3.4).

3.2.2 *Perceptual organization*

Recent progress in rigid-object recognition has lead to schemes that perform remarkably better than humans for limited libraries of models. The computational complexity of these schemes depends critically on the number of "features" used for matching. Therefore, the choice of features is an important issue. A simple feature that has been used is a point of an edge. This has the problem that, typically, there are many such features and they are not sufficiently distinctive for recognition thereby increasing the complexity of the search process. Complexity can be reduced by grouping these features into lines [Grimson 1990]. Lines in this context are a form of grouping. This idea has been pushed further and several schemes exist that try to group edge segments that come from the same object. The general idea underlying grouping is that "group features" are more distinctive and occur less frequently than individual features [Marroquin 1976], [Witkin and Tenenbaum 1983], [Mahoney 1985], [Lowe 1984, 1987], [Sha'ashua and Ullman 1988], [Jacobs 1989], [Grimson 1990], [Subirana-Vilanova 1990], [Clemens 1991], [Mahoney 1992], [Mahoney 1992b], [Mahmood 1993]. This has the effect of simplifying the complexity of the search space. However, even in this domain where existing perceptual organization has found use, complexity still limits the realistic number of models that can be handled. "Additional" groups obtained with region-based computations should be welcomed.

Representations which maintain some region information such as the sign-bit of the zero-crossings (instead of just the zero-crossings themselves) can be used for perceptual organization. One property that is easy to recover locally in the sign-bit image shown in Figure 3.4 is that of membership in the foreground (or background) of a certain portion of the image since a simple rule can be used: The foreground is black and the background white. (This rule can not be applied in general, however it illustrates how the coloring provided by the sign-bit image can be used to obtain region information.) In the edge image, this information is available but can not be computed locally. The region-based version of C.I.F. presented in this Chapter uses, to a certain extent, a similar principle to the one we have just discussed - namely, that often regions of interest have uniform properties (similar brightness, texture, etc.).

3.2.3 *Non-rigid objects*

Previous research on recognition has focused on rigid objects. In such a domain, one of the most useful constraints is that the image's change in appearance can be attributable mainly to a change in viewing position and luminance geometry[3]. It has been shown that this implies that the correspondence of a few features constrains the viewpoint (so that pose can be easily verified). Therefore, for rigid-objects, edge-based segmentation schemes which look for small groups of features that come from one object are sufficient. Since cameras introduce noise and edge-detectors fail to find some edges, the emphasis has been on making these schemes as robust as possible under spurious data and occlusion.

In contrast, not much research has been addressed to non-rigid objects where, as mentioned in Chapter 1, the change in appearance cannot be attributable solely to a change in viewing direction. Internal changes of the shape must be taken into account. Therefore, grouping a small subset of image features is not sufficient to recover the object's pose. A different form of grouping that can group all (or most of) the object's features is necessary. Even after extensive research on perceptual organization, there are no edge-based schemes that work in this domain (see also the next subsection). This may not be just a limitation on our understanding of the problem but a constraint imposed by the input used by such schemes. The use of more information, not just the edges, may simplify the problem. One of the goals of our research is to develop a scheme that can group features of a flexible object under a variety of settings that are robust under changes in illumination. Occlusion and spurious data should also be considered, but they are not the main driver of our research.

3.2.4 *Stability and scale*

In most images, interesting structures in different regions of the image occur at different scales. This is a problem for edge-based grouping because edge detectors are notably sensitive to the "scale" at which they are applied. This presents two problems for grouping schemes: it is not clear what is the scale at which to apply edge detectors and, in some images, not all edges of an object appear accurately at one single scale. Scale stability is in fact one of the most important sources of noise and spurious data mentioned above.

[3]For polygonal shapes, in most applications luminance can be ignored if it is possible to recover edges sufficiently accurately.

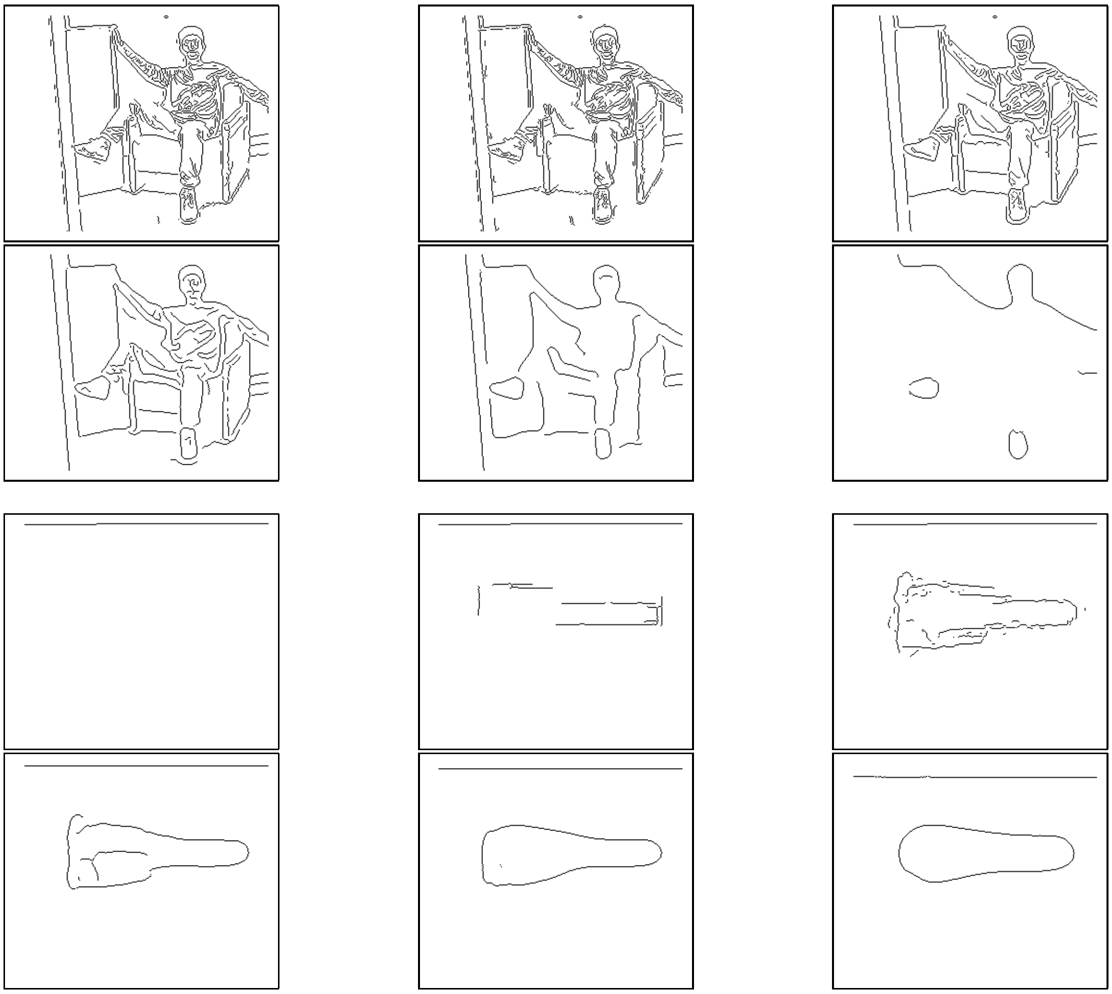


Figure 3.5: Edges computed at six different scales for 227X256 images. The results are notably different. Which scale is best? *Top six:* Image of a person; scales 1, 2, 4, 8, 16, and 32. Note that some of the edges corresponding to the legs are never found. *Bottom six:* Blob image; scales 4, 8, 16, 32, 64, and 128. Note also that the scales in the two images do not correspond.

Consider for example Figure 3.5 where we have presented the edges of a person at different scales. Note that there is no single scale where the silhouette of the person is not broken. For the purposes of recognition, the interesting edges are obviously the ones corresponding to the object of interest. Determining the scale at which these appear is not a trivial task. In fact, some edges do not even appear in any scale (e.g. the knee in Figure 3.5).

This problem has been addressed in the past [Zhong and Mallat 1990], [Lu and Jain 1989], [Clark 1988], [Geiger and Poggio 1987], [Schunck 1987], [Perona and Malik 1987], [Zhuang, Huang and Chen 1986], [Canny 1985], [Witkin 1984], but edge detection has treated scale as an isolated issue, independent of the other edges that may be involved in the object of interest. We believe that the stability and scale of the edges should depend on the region to which they belong and not solely on the discontinuity that gives rise to them. The scheme that we will present looks for the objects directly, not just for the individual edges. This means that in our research we address stability in terms of objects (not edges). In fact, our scheme commits to a scale which varies through the image; usually it varies also within the object. This scale corresponds to that of the object of interest chosen by our scheme.

3.3 Color, Brightness, Or Texture?

Early vision schemes can be divided based on the type of information that they use such as brightness, color, texture, motion, stereo. Similarly, one can design region-based schemes that use any of these sources of information. We decided to extend C.I.F. it to color first, without texture or brightness.

Color based perceptual organization (without the use of other cues) is indeed possible for humans since two adjacent untextured surfaces viewed under iso-luminant conditions can be segmented[4]. In addition, color may be necessary even if there are brightness changes since most scenes contain iso-luminant portions. As we will see later in the chapter, color is interesting because (like color and texture) it can be

[4]However, the human visual system has certain limitations in iso-luminant displays (e.g. [Cavanaugh 1987]).

casted as a vector property. Most schemes that work on brightness images do not extend naturally to vector properties. In contrast, vector ridge-detector techniques should be applicable to color, motion and texture.

3.4 A Color Difference Measure

Under normal conditions, color is a perceived property of a surface that depends mostly upon surface spectral reflectance and very little on the spectral characteristics of the light entering our eyes. It is therefore useful for describing the material composition of a surface (independently of its shape and imaging geometry) [Rubin and Richards 1981]. Lambertian color is indeed uniform over most untextured physical surfaces, and is stable in shadows and under changes in the surface orientation or the imaging geometry. In general, it is more stable than texture or brightness. It has long been known that the perceived color (or intensity) at any given image point depends on the light reflected from the various parts of the image, and not only on the light at that point. This is known as the simultaneous-contrast phenomena and has been known at least since E. Mach reported it at the beginning of the century. [Marr 1982] suggests that such a strategy may be used because one way of achieving some compensation for illuminance changes is by looking at differences rather than absolute values. According to this view, a surface is yellow because it reflects more "yellow" light than a blue surface, and not because of the absolute amount of yellow light reflected (of which the blue surface may reflect an arbitrary amount depending on the incident light).

The exact algorithm by which humans compute perceived color is still unclear. C.I.F. only requires a rough estimate of color which is used to segment the image, see Figure 3.6. We believe perceived color should be computed at a later stage by a process similar to the ones described in [Helson 1938], [Judd 1940], [Land and McCann 1971]. This model is in line with the ones presented in Appendix B and [Jepson and Richards 1991] which suggest that perceptual organization is a very early process which precedes most early visual processing. In our images, color is entered in the computer as a "color vector" with three components: the red, green, and blue channels of the video signal. Our scheme works on color differences $\mathcal{S}_{\otimes}$ between pairs

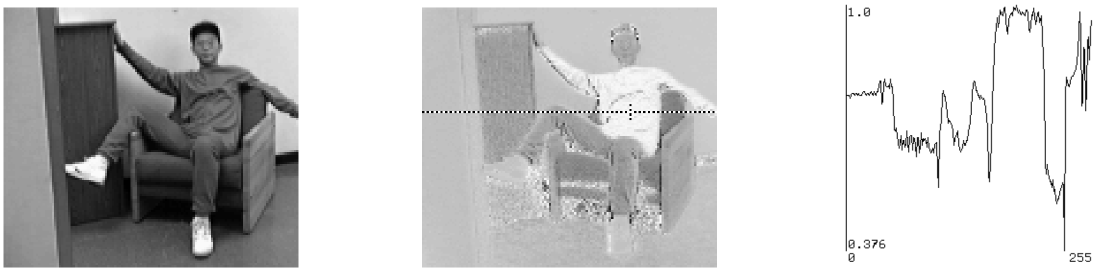


Figure 3.6: The similarity measure described in Equation 1 is illustrated here for an image of a person. *Left:* Image. *Center:* Similarity measure using as reference color the color of the pixel located at the intersection of the two segments shown. *Right:* Plot of the similarity measure along the long segment using the same reference color.

of pixels $\vec{c}$ and $\vec{c_R}$. The difference that we used is defined in equation 3.1 and was taken from [Sung 1991] ($\otimes$ denotes the vector cross product operation) and responds sensitively to color differences between similar colors.

$$\mathcal{S}_{\otimes}(\vec{c}) = 1 - \frac{|\vec{c} \otimes \vec{c_R}|}{|\vec{c}||\vec{c_R}|} \tag{3.1}$$

This similarity measure is a decreasing function with respect to the angular color difference. It assigns a maximum value of 1 to colors that are identical to the reference "ridge color", $\vec{c_R}$, and a minimum value of 0 to colors that are orthogonal to $\vec{c_R}$ in the RGB vector space. The discriminability of this measure can be seen intuitively by looking at the normalized image in Figure 3.6. The exact nature of this measure is not critical to our algorithm. What is important is that when two adjacent objects have different "perceived" color (in the same background) this measure is positive[5] Other measures have been proposed in the literature and most of them could be incorporated in our scheme.

What most color similarity measures have in common is that they are based on vector values and cannot be mapped onto a one-dimensional field [Judd and Wyszecki 75][6]. This makes color perception different from brightness from a computational

[5] Note that the perceived color similarity among arbitrary objects in the scene will obviously not correspond to this measure. Especially if we do not take into account the simultaneous-contrast phenomena.

[6] Note that using the three channels, red, green, and blue independently works for some cases.

point of view since not all the one-dimensional techniques used in brightness images extend naturally to higher dimensions.

3.5 Regions? What Regions?

In the last three sections we have set forth an ambitious goal: Develop a perceptual organization scheme that works on the image itself, without edges and using color, brightness, and texture information.

But what constitutes a good region? What "class" of regions ought to be found? Our work is based on the observation that most objects in nature (or their parts) have a common color or texture, and are long, wide, symmetric, and convex. This hypothesis is hard to verify formally, but it is at least true for a collection of common objects [Snodgrass and Vanderwart 1980] used in psychophysics. And as we will show, it can be used in our scheme yielding seemingly useful results. In addition, humans seem to organize the visual array using these heuristics as demonstrated by the Gestalt Psychologists [Wertheimer 1923], [Koffka 1935], [Köhler 1940]. In fact, these principles were the starting point for much of the work in computer vision on perceptual organization for rigid objects. We use these same principles but in a different way: Without edges and with non-rigid shapes in mind.

In the next section, we describe some common problems in finding regions. To do so, we introduce a one dimensional version of "regions" and discuss the problems involved in this simplified version of the task. A scheme to solve the one dimensional version of the problem is discussed in Sections 3.7 and 3.8. This exercise is useful because both the problems and the solution encountered generalize to the two dimensional version, which is based on an extension of C.I.F..

However, it is possible to construct cases in which it does not, as when an object has two discontinuities, one in the red channel only and the other in one of the other two channels only. In addition, the perceived similarity is not well captured by the information contained in the individual channels alone but on the combined measure.

3.6 Problems in Finding Brightness Ridges

One way of simplifying perceptual organization is to start by looking at a one dimensional version of the problem. This is especially useful if the solution lends itself to a generalized scheme for the two dimensional problem. This would be a similar path to the one followed by most edge detection research. In the case of edge detection, a common one-dimensional version of the problem is a step function (as shown in Figure 3.3). Similarly, perceptual organization without edges can be cast in one dimension, as the problem of finding ridges similar to a hat (as shown in Figure 3.3). A hat is a good model because it has one of the basic properties of a region: it is uniform and has discontinuities in its border. As we will see shortly, the hat model needs to be modified before it can reflect all the properties of regions that interest us.

In other words, the one-dimensional version of the problem that we are trying to solve is to locate ridges in a one-dimensional signal. By ridge we mean something that "looks like" a pair of step edges (see Figure 3.3). A simple-minded approach is to find the edges in the image, and then look for the center of the two edges. This was the approach suggested at the beginning of this Chapter [Subirana-Vilanova 1990]. Another possibility is to design a filter to detect such a structure as in [Canny 1985], [Kass and Witkin 1988], [Noble 1988].

However, the use of such filters as estimators for ridge detection suffers from a number of problems. These problems are not particular to either scheme, but are linked to the nature of ridges in real images[7]. The model of a ridge used in these schemes is similar to the hat model shown in Figure 3.3. This is a limited model since ridges in images are not well suited to it. Perhaps the most obvious reason why such a model is not realistic is the fact that it is tuned to a particular scale, while, in most images, ridges appear at multiple and unpredictable scales. This is not so much of a problem in edge-detection as we have discussed in the previous sections, because the edges of a wide range of images can be assumed to have "a very similar scale." Thus, Canny's ridge detector works only on images where all ridges are of the same scale as is true in the text images shown in [Canny 1983] (see also Figures 3.17

[7]In fact, most of these problems are similar for color and for brightness images.

Figure 3.7: *Left:* Plot with multiple steps. A ridge detector should detect three ridges. *Right:* Plot with narrow valleys. A ridge detector should be able to detect the different lobes independently of the size of the neighboring lobes.

and 3.18).

Therefore, an important feature of a ridge detector is its scale invariance. We now summarize a number of important features that a ridge operator should have (see Figure 3.7):

- *Scale:* See previous paragraph.
- *Non-edgeness:* The filter should give no response for a step edge. This property is violated by [Canny 1985], [Kass and Witkin 1988].
- *Multiple steps:* The filter should also detect regions between small steps. These are frequent in images, for example when an object is occluding the space between two other objects. This complicates matters in color images because the surfaces are defined by vectors not just scalar values.
- *Narrow valleys:* The operator should also work in the presence of multiple ridges even if they are separated by small valleys.
- *Noise:* As with any operator that is to work in real images, tolerance to noise is a critical factor.
- *Localization:* The ridge-detector output should be higher in the middle of the ridge than on the sides.
- *Strength:* The strength of the response should be somehow correlated with the strength of the perception of the ridge by humans.

- *Large scales:* Large scales should receive higher response. This is a property used by C.I.F. and it is important because it embodies the preference for large objects (see also Section 14).

3.7 A Color Ridge Detector

In the previous section we have outlined a number of properties we would like a ridge-detector to have. As we have mentioned, the ridge-detectors of Canny, and Kass and Witkin fail because, among other things, they cannot handle multiple scales. A naive way of solving the scale problem would be to apply such ridge detectors at multiple scales and define, the output of the filter at each point, as the response at the scale which yields a maximum value at that point. This filter would work in a number of situations but has the problem of giving a response for step edges (since the ridge-detector at any single scale responds to edges, so will the combined filter - see Figures 3.17 and 3.18).

One can suppress the response to edges by splitting Canny's ridge operator into two pieces, one for each edge, and then combining the two responses by looking at the minimum of the two. This is the basic idea behind our approach (see Figures 3.8 and 3.9). Figures 3.17 and 3.18 illustrate how our filter behaves according to the different criteria outlined before. The Figures also compare our filter with that of the second derivative of a gaussian, which is a close approximation of the ridge-filter Canny used. There are a number of potential candidates within this framework such as splitting a Canny filter by half, using two edge detectors and many others. We tried a number of possibilities on the Connection Machine using a real and a synthetic image with varying degrees of noise. Table 3.7 describes the filter which gives a response most similar to the inertia values and the tolerated length that one would obtain using similar formulas for the corresponding edges, as described in Chapter 2 [Subirana-Vilanova 1990]. The validity of this approach is further confirmed by analytical results presented in Section 3.8.

This approach uses two filters (see profiles in Figure 3.8, 3.9 and Table 3.1), each of which looks at one side of the ridge. The output of the combined filter is the

VAR.	EXPRESSION	DESCRIPTION
$\mathcal{P}_{max}$ F_s F_c	Free Parameter (3) Free Parameter (8) Free Parameter (1/8)	Gradient penalization coeff. Filter Side Lobe size coeff. Local Neighborhood size coeff.
$g(x)$ g_{max}		Color gradient at location x. Max. color gradient in image.
σ σ_s σ_c	 σ/F_s $F_c\sigma$	Size of Main Filter Lobe. Size of Side Filter Lobe. Reference Color Neighborhood
$\vec{c}(x)$ $\vec{c}_n(x)$ $\vec{c}_r(x)$	$[\mathcal{R}(x)\ \mathcal{G}(x)\ \mathcal{B}(x)]^T$ $\vec{c}(x)/\vert\vec{c}(x)\vert$ $\int_{-\sigma_c}^{\sigma_c} \frac{1}{\sqrt{2\pi}\sigma_c}\ e^{-\frac{r^2}{2\sigma_c^2}}\ \vec{c}_n(x+r)\ dr$	Color vector at location x. Normalized Color at x. Reference Color at x
$\mathcal{F}_L(r)$	$\begin{array}{ll} \frac{r+\sigma}{\sigma^2\sqrt{2\pi}}\ e^{-\frac{(r+\sigma)^2}{2\sigma^2}} & -\sigma < r < \sigma \\ \frac{r+\sigma}{\sigma_s^2\sqrt{2\pi}}\ e^{-\frac{(r+\sigma)^2}{2\sigma_s^2}} & -(\sigma+2\sigma_s) < r < -\sigma \\ 0 & \text{otherwise} \end{array}$	Left Half of Filter
$\mathcal{F}_R(r)$	$\mathcal{F}_L(-r)$	Right Half of Filter
$\mathcal{I}_L(x)$ $\mathcal{I}_R(x)$	$\int_{-(\sigma+\sigma_s)}^{\sigma} S_{\otimes}(\vec{c}_r(x), \vec{c}_n(x+r))\mathcal{F}_L(r)\ dr$ $\int_{-\sigma}^{\sigma+\sigma_s} S_{\otimes}(\vec{c}_r(x), \vec{c}_n(x+r))\mathcal{F}_R(r)\ dr$	Inertia from Left Half Inertia from Right Half
$\mathcal{I}_\sigma(x)$	$\min(\mathcal{I}_L(x), \mathcal{I}_R(x))\ \frac{\sqrt{\sigma}}{(1+\mathcal{P}_{max}\frac{g(x)}{g_{max}})^2}$	Inertia at location x (Scale σ).
$\mathcal{I}(x)$ $\sigma(max)$	$\forall\sigma\ \max(\mathcal{I}_\sigma(x))$ σ such that $\mathcal{I}_\sigma(x)$ is maximized	Overall inertia at location x.
$\mathcal{T}_L(x)$	$\begin{array}{ll} 0 & \text{if } r_c < \sigma(max) \\ r_c(\pi - \arccos(\frac{r_c-\sigma(max)}{r_c})) & \text{otherwise} \end{array}$	Tolerated Length (Depends on radius of curvature r_c)

Table 3.1: Steps for Computing Directional Inertias and Tolerated Length. Note that the scale σ is *not* a free parameter.

Figure 3.8: *Left:* Gaussian second derivative, an approximation of Canny's optimal ridge detector. *Right:* Individual one-dimensional masks used by our operator.

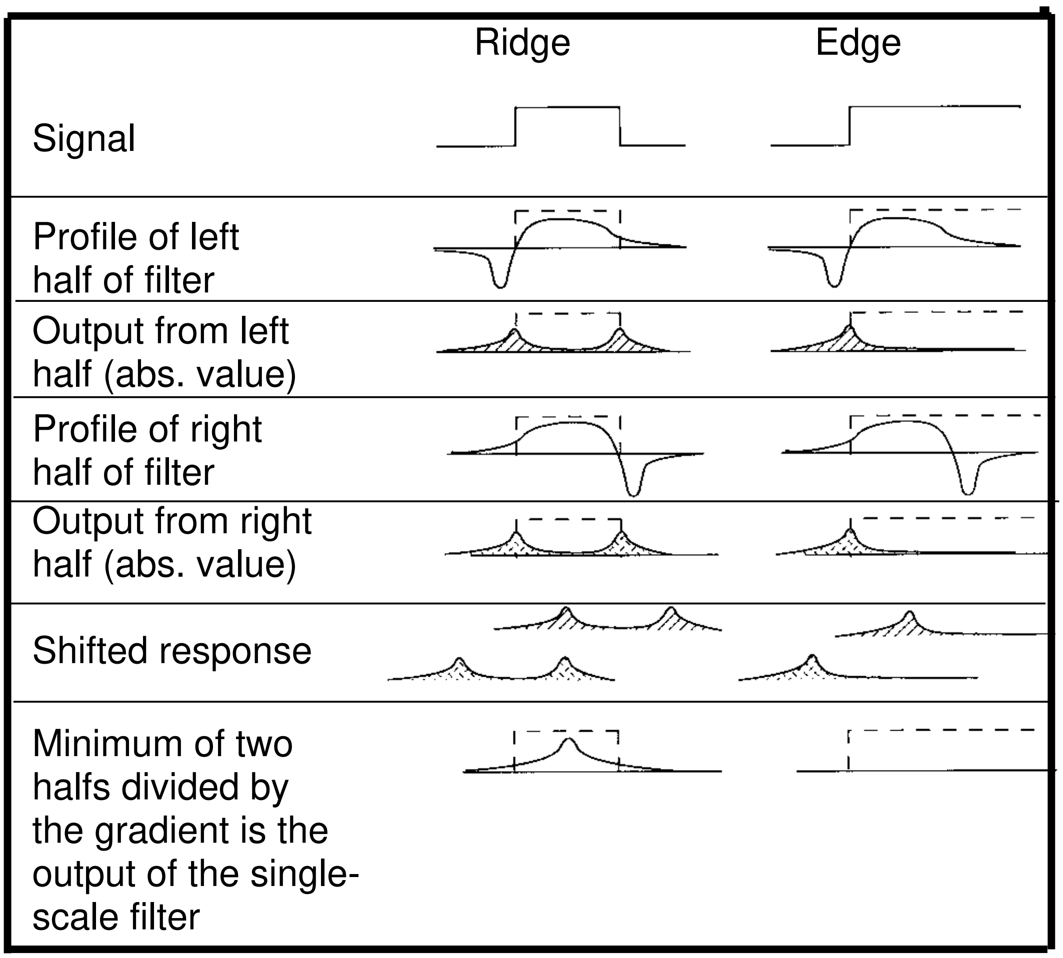


Figure 3.9: Intuitive description of ridge detector output on flat ridge and edge.

minimum of the two responses. Each of the two parts of the filter is asymmetrical, reflecting the fact that we expect the object to be uniform (which explains each filter's large central lobe), and that we do not expect that a region of equal size be adjacent to the object (which explains each filter's small side lobe to accomodate for narrower adjacent regions). In other words, our ridge detector is designed to handle narrow valleys.

The extension to handling steps and color is tricky because there is no clear notion of what is positive and what is negative in vector quantities. We solve this problem by adaptively defining a reference color at each point as the weighted average color over a small neighborhood of the point (about eight times smaller than the scale of the filter, in the current implementation). Thus, this reference color will be different for different points in the image, and scalar deviations from the reference color are computed as defined in Section 3.3.

3.8 Filter Characteristics

This Section examines some interesting characteristics of our filter under noiseless and noisy operating conditions. We begin in Section 3.8.1 by deriving the filter's optimum scale response and its optimum scale map for noiseless ridge profiles, from which we see that both exhibit local output extrema at ridge centers. Next, we examine our filter's scale (Section 3.8.2) and spatial (Section 3.8.3) localization characteristics under varying degrees of noise. Scale localization measures the closeness in value between the optimum mask size at a ridge center and the actual width of the ridge. Spatial localization measures the closeness in position between the filter's peak response location and the actual ridge center. We shall see that both the filter's optimum scale and peak response location remain remarkably stable even at noticeably high noise levels. Our analysis will conclude with a comparison with Canny's ridge detector in Section 3.8.4 and experimental results in Section 3.9.

For simplicity, we shall perform our analysis on scalar ridge profiles instead of color ridge profiles. The extension to color is straightforward if we think of the *reference color* notion and the color similarity measure of equation 3.1 as a transformation

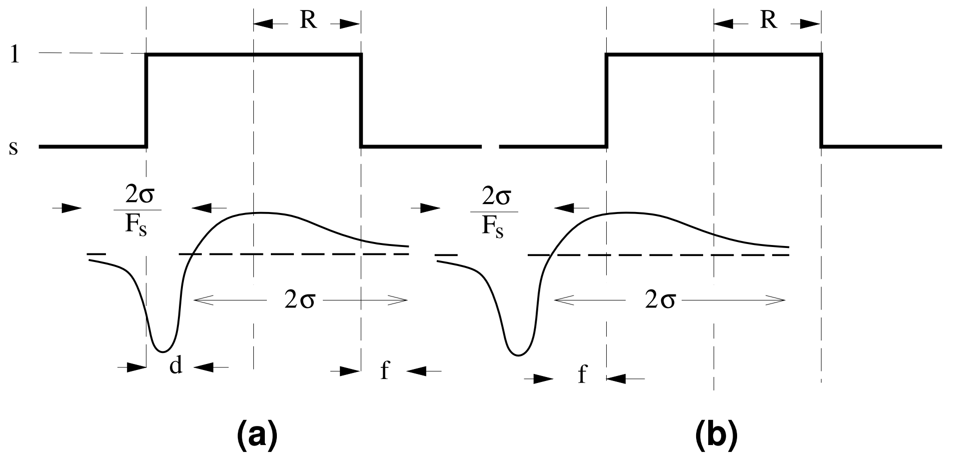


Figure 3.10: Half-mask configurations for computing the optimum scale ridge response of our filter. See text for explanation.

that converts color ridge profiles into scalar ridge profiles.

We shall be using filter notations similar to those given in Table 3.7. In particular, σ denotes the main lobe's width (or scale); F_s denotes the filter's main lobe to side lobe width ratio; and $\mathcal{F}_L(r, \sigma_m, \sigma_s)$ a left-half filter with main lobe size σ_m, side lobe size $\sigma_s = \sigma_m / F_s$, and whose form is a normalized combination of two Gaussian first derivatives. At each point on a ridge profile, the filter output, by definition, is the maximum response for mask pairs of all scales centered at that point.

3.8.1 *Filter response and optimum scale*

Let us first obtain the *single scale* filter response for the two half-mask configurations in Figure 3.10. Figure 3.10(a) shows an off-center left half-mask whose side lobe overlaps the ridge plateau by $0 \leq d \leq 2\sigma/F_s$ and whose main lobe partly falls off the right edge of the ridge plateau by $0 \leq f \leq 2\sigma$. The output in terms of mask dimensions and offset parameters is:

$$O_a(d, f) \;=\; \int_{-(\sigma+\frac{2\sigma}{F_s})}^{-(\sigma+d)} s\mathcal{F}_L(r, \sigma, \frac{\sigma}{F_s})dr + \int_{-(\sigma+d)}^{\sigma-f} \mathcal{F}_L(r, \sigma, \frac{\sigma}{F_s})dr + \int_{\sigma-f}^{\sigma} s\mathcal{F}_L(r, \sigma, \frac{\sigma}{F_s})dr$$

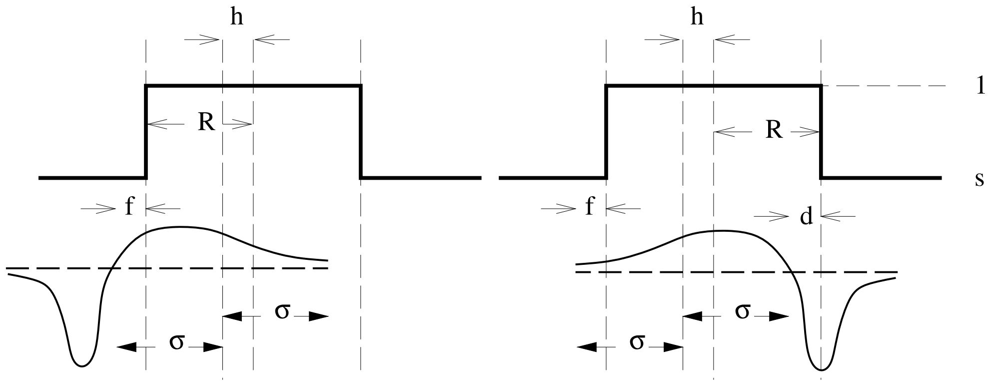


Figure 3.11: Mask pair configuration for computing the all scales optimum ridge response of our filter. See text for explanation.

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi}}\Big[(F_s s - 1)(e^{-2} - 1) \\
&\quad -(1-s)\left(F_s(1 - e^{-\frac{F_s^2 d^2}{2\sigma^2}}) + (e^{-\frac{(2\sigma - f)^2}{2\sigma^2}} - e^{-2})\right)\Big] \qquad (3.2)
\end{aligned}
$$

A value of f greater than d indicates that the filter's main lobe (i.e. its scale) is wider than the ridge and vice-versa. Notice that when $d = f = 0$, we have a perfectly centered mask whose main lobe width equals the ridge width, and whose output value is globally maximum.

Figure 3.10(b) shows another possible left half-mask configuration in which the main lobe partly falls outside the left edge of the ridge plateau by $0 \leq f \leq 2\sigma$. Its output is:

$$
\begin{aligned}
O_b(f) &= \int_{-(\sigma + \frac{2\sigma}{F_s})}^{-(\sigma - f)} s\mathcal{F}_L(r, \sigma, \frac{\sigma}{F_s})dr + \int_{-(\sigma - d)}^{\sigma} \mathcal{F}_L(r, \sigma, \frac{\sigma}{F_s})dr \\
&= \frac{1}{\sqrt{2\pi}}\left[(F_s s - 1)(e^{-2} - 1) - (1-s)(1 - e^{-\frac{f^2}{2\sigma^2}})\right] \qquad (3.3)
\end{aligned}
$$

The equivalent right-half mask configurations are just mirror images of the two left-half mask configurations, and have similar *single scale* ridge response values.

Consider now the *all scales* optimum filter response of a mask pair, offset by h

from the center of a ridge profile (see Figure 3.11). The values of d and f in the figure can be expressed in terms of the ridge radius (R), the filter size (σ) and the offset distance (h) as follows:

$$
\begin{aligned}
d &= R + h - \sigma \\
f &= \sigma + h - R
\end{aligned}
$$

Notice that the right-half mask configuration in Figure 3.11 is exactly the mirror image of the left-half mask configuration in Figure 3.10(a).

Because increasing σ causes f to increase which in turn causes the left-half mask output to decrease, while decreasing σ causes d to increase which in turn causes the right-half mask output to decrease, the *all scales* optimum filter response, $\texttt{Opt}(h, R)$, must therefore be from the scale, σ_o, whose left and right half response values are equal. Using the identities for d and f above with the half-mask response equations 3.2 and 3.3, we get, after some algebriac simplification:

$$\texttt{Opt}(h, R) = \frac{1}{\sqrt{2\pi}} \left[(F_s s - 1)(e^{-2} - 1) - (1 - s)(1 - e^{-\frac{(\sigma_o + h - R)^2}{2\sigma_o^2}}) \right] \tag{3.4}$$

where the *optimum scale*, σ_o, must satisfy the following equality:

$$F_s(1 - e^{-\frac{F_s^2(R+h-\sigma_o)^2}{2\sigma_o^2}}) + (e^{-\frac{(\sigma_o - h + R)^2}{2\sigma_o^2}} - e^{-2}) = (1 - e^{-\frac{(\sigma_o + h - R)^2}{2\sigma_o^2}}). \tag{3.5}$$

The following bounds for σ_o can be obtained:

$$\frac{R + h}{1 + \frac{\sqrt{2}}{F_s} \ln(\frac{F_s}{F_s - 1 - e^{-2}})} < \sigma_o < (R + h). \tag{3.6}$$

For our particular implementation, we have $F_s = 8$ which gives us: $0.9737(R + h) < \sigma_o < (R + h)$. Since $h \geq 0$, Equation 3.6 indicates that the optimum filter scale, σ_o, is a local *minimum* at ridge centers where $h = 0$.

To show that the *all scales* optimum filter response is indeed a local *maximum* at ridge centers, let us assume, using the inequality bounds in Equation 3.6, that $\sigma_o = k(R+h)$ for some fixed k in the range:

$$\frac{1}{1+\frac{\sqrt{2}}{F_s}\ln(\frac{F_s}{F_s-1-e^{-2}})} < K < 1.$$

Equation 3.4 becomes:

$$\texttt{Opt}(h,R) = \frac{1}{\sqrt{2\pi}}\left[(F_s s-1)(e^{-2}-1)-(1-s)(1-e^{-\frac{((1+k)h-(1-k)R)^2}{2k^2(R+h)^2}})\right]. \qquad (3.7)$$

Differentiating the above equation with respect to h, we see that $\texttt{Opt}(h,R)$ indeed decreases with increasing h for values of h near 0.

3.8.2 *Scale localization*

We shall approach the scale localization analysis as follows (see Figure 3.12(a)): Consider a radius R ridge profile whose *signal to noise ratio* is $(1-s)/n_o$, where $(1-s)$ is the height of the ridge signal and n_o^2 is the noise variance. Let $d = |R-\sigma_o|$ be the size difference between the ridge radius and the optimum filter scale at the ridge center. We want to obtain an estimate for the magnitude of d/R, which measures the *relative* error in scale due to noise.

Figures 3.12(b) (c) and (d) show three possible left-half mask configurations aligned with the ridge center. In the absence of noise (i.e. if $n_o = 0$), their respective output values (O_s) are:

$$\begin{aligned}(\sigma = R) \qquad : O_s &= \int_{-(\sigma+\frac{2\sigma}{F_s})}^{-\sigma} s\mathcal{F}_L(r,\sigma,\frac{\sigma}{F_s})dr + \int_{-\sigma}^{\sigma} \mathcal{F}_L(r,\sigma,\frac{\sigma}{F_s})dr \\ &= \frac{1}{\sqrt{2\pi}}(1-e^{-2})(1-sF_s)\end{aligned}$$

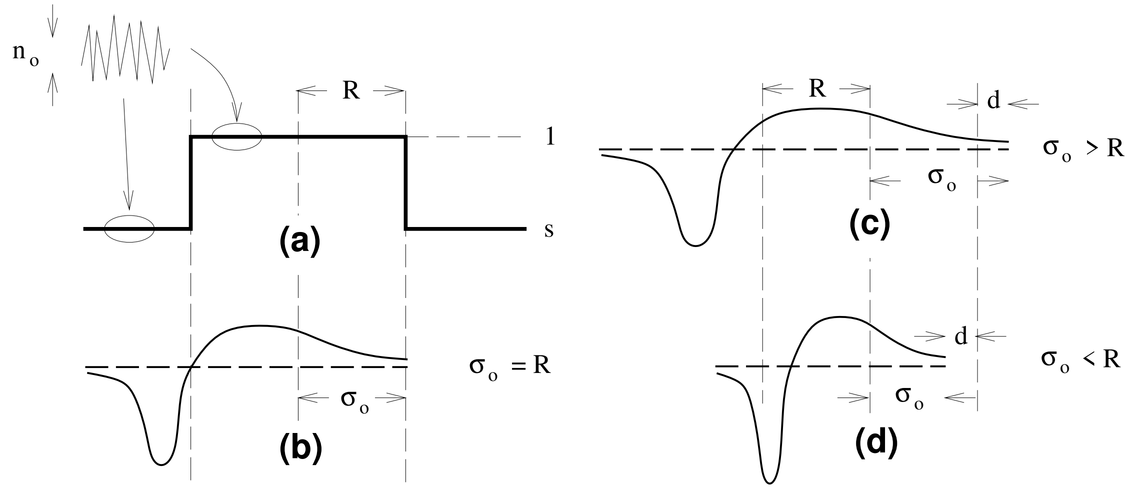


Figure 3.12: Mask configurations for scale localization analysis. (a) A radius R ridge profile with *noise to signal* ratio $n_o/(1-s)$. (b) A mask whose scale equals the ridge dimension. (c) A mask whose scale is larger than the ridge dimension. (d) A mask whose scale is smaller than the ridge dimension.

$$
\begin{aligned}
(\sigma = R + d) : O_s(d) &= \int_{-(\sigma+\frac{2\sigma}{F_s})}^{-(\sigma-d)} s\mathcal{F}_L(r,\sigma,\frac{\sigma}{F_s})dr + \int_{-(\sigma-d)}^{\sigma-d} \mathcal{F}_L(r,\sigma,\frac{\sigma}{F_s})dr \\
&\quad + \int_{\sigma-d}^{\sigma} s\mathcal{F}_L(r,\sigma,\frac{\sigma}{F_s})dr \\
&= \frac{1}{\sqrt{2\pi}} \Big[(1-e^{-2})(1-sF_s) \\
&\quad +(1-s)(e^{-2} + e^{-\frac{d^2}{2(R+d)^2}} - e^{-2}e^{\frac{2d}{R+d}}e^{-\frac{d^2}{2(R+d)^2}} - 1)\Big] \\
(\sigma = R - d) : O_s(d) &= \int_{-(\sigma+\frac{2\sigma}{F_s})}^{-(\sigma+d)} s\mathcal{F}_L(r,\sigma,\frac{\sigma}{F_s})dr + \int_{-(\sigma+d)}^{\sigma} \mathcal{F}_L(r,\sigma,\frac{\sigma}{F_s})dr \\
&= \frac{1}{\sqrt{2\pi}} \left[(1-e^{-2})(1-sF_s) + (1-s)F_s(e^{-\frac{F_s^2 d^2}{2(R-d)^2}} - 1)\right]
\end{aligned}
\tag{3.8}
$$

Let us now compute O_n, the noise component of the filter output. Since the noise signal is white and zero mean, we have $\mathtt{E}[O_n] = 0$, where $\mathtt{E}[x]$ stands for the expected value of x. For noise of variance n_o^2, the variance of O_n is:

$$\begin{aligned}\mathtt{Var}[O_n] &= \int_{-(\sigma+\frac{2\sigma}{F_s})}^{\sigma} n_o^2 \mathcal{F}_L^2(r,\sigma,\frac{\sigma}{F_s})dr \approx \int_{-\infty}^{\infty} n_o^2 \mathcal{F}_L^2(r,\sigma,\frac{\sigma}{F_s})dr \\ &= \frac{1+F_s}{8\sigma\sqrt{\pi}} \approx \frac{1+F_s}{8R\sqrt{\pi}},\end{aligned} \tag{3.9}$$

or equivalently, the standard deviation of O_n is:

$$\mathtt{SD}[O_n] = \sqrt{\frac{1+F_s}{8R\sqrt{\pi}}}. \tag{3.10}$$

A loose upper bound for d/R can be obtained by finding d, such that the noiseless response for a size $\sigma = R+d$ (or size $\sigma = R-d$) mask is within one noise output standard deviation of the optimum scale response (ie. the response for a mask of size $\sigma_o = R$). We examine first, the case when $\sigma = R+d$. Subtracting O_s for $\sigma = R$ from $O_s(d)$ for $\sigma = R+d$ (both from the series of equations 3.8) and equating the difference with $\mathtt{SD}[O_n]$, we get:

$$(1-s)(1-e^{-2}+e^{-\frac{d^2}{2(R+d)^2}}-e^{-2}e^{\frac{2d}{R+d}}e^{-\frac{d^2}{2(R+d)^2}}) = \sqrt{\frac{1+F_s}{8R\sqrt{\pi}}},$$

which, after some algebra and simplifying approximations, becomes:

$$\begin{aligned} d/R &\approx \frac{\sqrt{2K}}{1-\sqrt{2K}} \qquad (0 \leq \frac{n_o}{1-s} < (1-e^{-2})(1-e^{-\frac{1}{2}})\sqrt{\frac{8R\sqrt{\pi}}{1+F_s}}) \\ &\quad \text{where :} \qquad K = -\ln\left(1-\frac{n_o}{1-s}\frac{1}{1-e^{-2}}\sqrt{\frac{1+F_s}{8R\sqrt{\pi}}}\right). \end{aligned} \tag{3.11}$$

Figure 3.13(a) graphs d/R as a function of the *noise to signal ratio* $n_o/(1-s)$. We remind the reader that our derivation is in fact a probabilistic upper bound for d/R. For d/R to exceed the bound, the $\sigma = R+d$ filter must actually produce a combined signal and noise response, greater than that of all the other filters with sizes from $\sigma = R$ to $\sigma = R+d$.

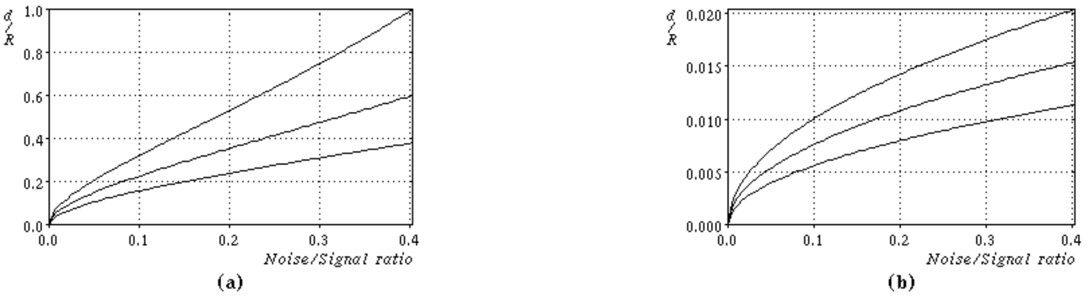


Figure 3.13: Relative scale error (d/R) as a function of *noise to signal ratio* ($n_o/(1-s)$) for *(a) Equation 3.11 where* $\sigma_o > R$, and *(b) Equation 3.12 where* $\sigma_o < R$. For both graphs, $F_s = 8$, top curve is for $R = 10$, middle curve is for $R = 30$ and bottom curve is for $R = 100$.

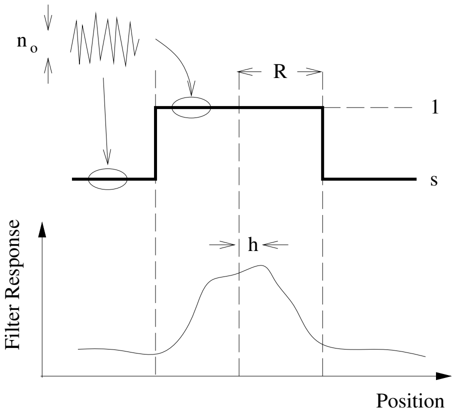
A similar analysis for the $\sigma = R - d$ case yields (see Figure 3.13(b) for plot):

$$
\begin{aligned}
d/R \;&\approx\; \frac{\sqrt{2K}}{F_s + \sqrt{2K}} \\
&\text{where}: \quad K = -\ln\left(1 - \frac{n_o}{1-s}\sqrt{\frac{1+F_s}{8F_s^2 R\sqrt{\pi}}}\right). \qquad (3.12)
\end{aligned}
$$

3.8.3 *Spatial localization*

Consider the radius R ridge in Figure 3.14 whose signal to noise ratio is $(1-s)/n_o$. As before, $(1-s)$ is the height of the ridge signal and n_o^2 is the noise variance. Let h be the distance between the actual ridge center and the peak location of the filter's *all scales* ridge response. Our goal is to establish some magnitude bound for h/R that can be brought about by the given noise level.

To make our analysis feasible, let us assume, using Equation 3.6, that the optimum filter scale at distance h from the ridge center is $\sigma_o = R + h$. Notice that for our typical values of F_s, the uncertainty bounds for σ_o are relatively small. The optimum scale filter output without noise is therefore:



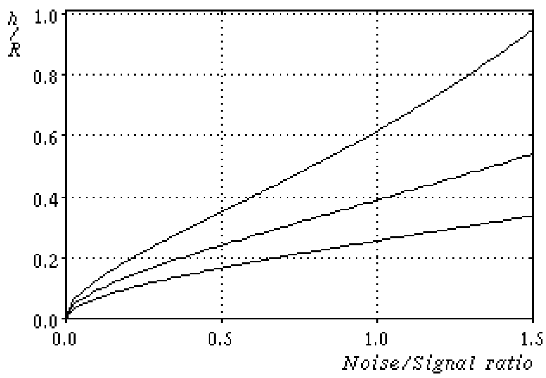


Figure 3.14: **Left:** Mask configurations for scale localization analysis. An all scales filter response for a radius R ridge profile with *noise to signal* ratio $n_o/(1-s)$. h is the distance between the actual ridge center and the filter response peak location. **Right:** Relative spatial error (h/R) as a function of *noise to signal ratio* ($n_o/(1-s)$), where $F_s = 8$, top curve is for $R = 10$, middle curve is for $R = 30$ and bottom curve is for $R = 100$. See Equation 3.15.

$$\mathtt{Opt}(h, R) \approx \frac{1}{\sqrt{2\pi}}\left[(F_s s - 1)(e^{-2} - 1) - (1 - s)(1 - e^{-\frac{4h^2}{2(R+h)^2}})\right], \tag{3.13}$$

and the difference in value between the above and the noiseless optimum scale output at ridge center is:

$$\mathtt{Opt}(0, R) - \mathtt{Opt}(h, R) \approx (1 - s)(1 - e^{-\frac{4h^2}{2(R+h)^2}}). \tag{3.14}$$

As in the scale localization case, we obtain an estimate for h/R by finding h such that the difference in Equation 3.14 equals one noise output standard deviation of the optimum scale filter at ridge center (see Equation 3.10). We get:

$$(1 - s)(1 - e^{-\frac{4h^2}{2(R+h)^2}}) = n_o\sqrt{\frac{1 + F_s}{8R\sqrt{\pi}}},$$

which eventually yields (see Figure 3.14 for plot):

$$h/R = \frac{\sqrt{K}}{\sqrt{2}-\sqrt{K}} \qquad (0 \le \frac{n_o}{1-s} < (1-e^{-2})\sqrt{\frac{8R\sqrt{\pi}}{1+F_s}})$$
$$\text{where}: \quad K = -\ln\left(1 - \frac{n_o}{1-s}\sqrt{\frac{1+F_s}{8R\sqrt{\pi}}}\right). \tag{3.15}$$

3.8.4 *Scale and spatial localization characteristics of the Canny ridge operator*

We compared our filter's scale and spatial localization characteristics with those of a Canny ridge operator. This is a relevant comparison because the Canny ridge operator was designed to be optimal for simple ridge profiles (see [Canny 1985] for details on the optimality criterion). The normalized form of Canny's ridge detector can be approximated by the shape of a scaled Gaussian second derivative:

$$\mathcal{C}(r,\sigma) = \frac{1}{\sqrt{2\pi}\sigma^3}(\sigma^2 - r^2)e^{-\frac{r^2}{2\sigma^2}}. \tag{3.16}$$

We begin with scale localization. For a noiseless ridge profile with radius R and height $(1-s)$, the optimum scale ($\sigma = R$) Canny filter response at the ridge center is:

$$O_s(\sigma = R) = \sqrt{\frac{2}{\pi}}(1-s)e^{-\frac{1}{2}}. \tag{3.17}$$

Similarly, the ridge center filter response for a mismatched Canny mask ($\sigma = R + d$) is:

$$O_s(\sigma = R+d) = \sqrt{\frac{2}{\pi}}\frac{R}{R+d}(1-s)e^{-\frac{R^2}{2(R+d)^2}},$$

where the scale difference, d, can be either positive or negative in value.

We want an estimate of d/R in terms of the noise to signal ratio. Consider now the effect of white Gaussian noise (zero mean and variance n_o^2) on the optimum scale Canny filter response. The noise output standard deviation is:

$$\begin{aligned}\mathtt{SD[O_n]} &= \sqrt{\int_{-\infty}^{\infty} n_o^2 \mathcal{C}^2(r, \sigma = R) dr} \\ &= n_o \sqrt{\frac{3}{8R\sqrt{\pi}}}. \end{aligned} \tag{3.18}$$

Performing the same scale localization steps as we did for our filter, we get:

$$n_o \sqrt{\frac{3}{8R\sqrt{\pi}}} = \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}} (1-s) - \sqrt{\frac{2}{\pi}} \frac{R}{R+d} e^{-\frac{R^2}{(R+d)^2}} (1-s),$$

which reduces to the following equation that implicitly relates d/R to $\frac{n_o}{1-s}$:

$$\frac{n_o}{1-s} = \sqrt{\frac{16R}{3\sqrt{\pi}}} \left[e^{-\frac{1}{2}} - \frac{R}{R+d} e^{-\frac{R^2}{2(R+d)^2}} \right]. \tag{3.19}$$

For spatial localization, we want an estimate of h/R in terms of $\frac{n_o}{1-s}$, where h is the distance between the actual ridge center and the *all scales* Canny operator peak output location. At distance h from the ridge center, the optimum Canny mask scale (σ_o) is bounded by:

$$\sqrt{\mathrm{R}^2 + \mathrm{h}^2 - 2\mathrm{Rh} \frac{1 - \mathrm{e}^{-\frac{4\mathrm{Rh}}{2(\mathrm{R}-\mathrm{h})^2}}}{1 + \mathrm{e}^{-\frac{4\mathrm{Rh}}{2(\mathrm{R}-\mathrm{h})^2}}}} \leq \sigma_\mathrm{o} \leq \sqrt{\mathrm{R}^2 + \mathrm{h}^2 - 2\mathrm{Rh} \frac{1 - \mathrm{e}^{-\frac{4\mathrm{Rh}}{2(\mathrm{R}+\mathrm{h})^2}}}{1 + \mathrm{e}^{-\frac{4\mathrm{Rh}}{2(\mathrm{R}+\mathrm{h})^2}}}},$$

and the noiseless optimum scale filter response is:

$$O_s(h) = \frac{2}{\sqrt{2\pi}\sigma_o} (1-s) e^{-\frac{R^2+h^2}{2\sigma_o^2}} \left[R \cosh(\frac{Rh}{\sigma_o^2}) - h \sinh(\frac{Rh}{\sigma_o^2}) \right].$$

Setting $O_s(0) - O_s(h) = \mathtt{SD}[O_n]$, we arrive at the following implicit equation relating h/R and $n_o/(1-s)$:

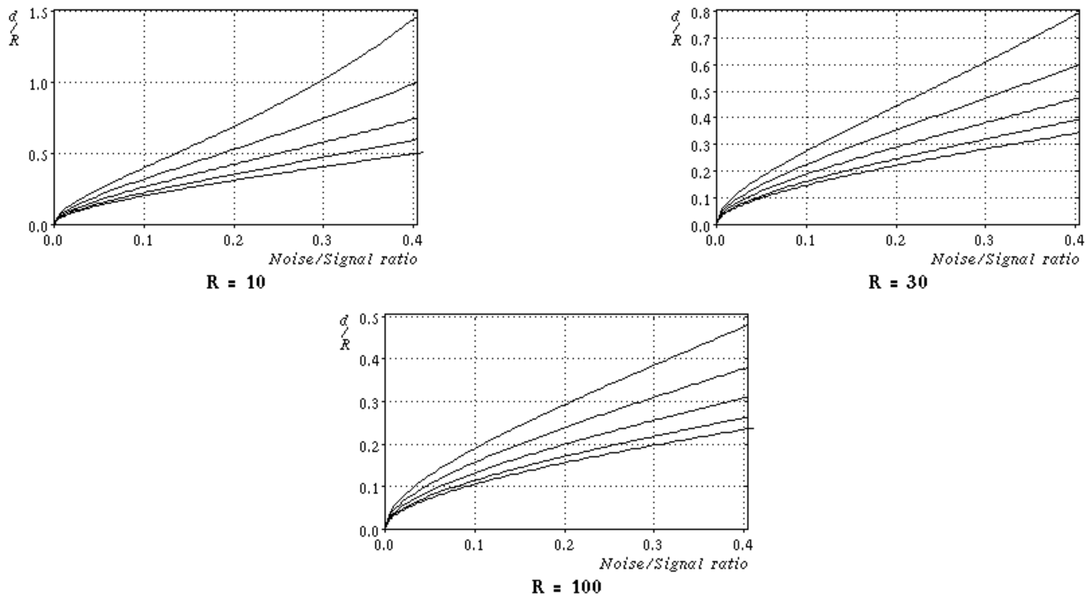


Figure 3.15: Comparison of relative scale error (d/R) as a function of *noise to signal ratio* ($n_o/(1-s)$) between our filter ($\sigma > R$ case) and the Canny ridge filter. See Equations 3.11 and 3.19. *Top Left:* $R = 10$. *Top Right:* $R = 30$. *Bottom:* $R = 100$. For each graph, curves from top to bottom are those of: $F_s = 16$, $F_s = 8$, $F_s = 4$, $F_s = 2$, and Canny.

$$\frac{n_o}{1-s} \approx \sqrt{\frac{4R}{3\sqrt{\pi}}}\left[e^{-\frac{1}{2}} - \frac{1}{\sigma_o}e^{-\frac{R^2+h^2}{2\sigma_o^2}}\left(R\cosh(\frac{Rh}{\sigma_o^2}) - h\sinh(\frac{Rh}{\sigma_o^2})\right)\right], \qquad (3.20)$$

where $\sigma_o \approx \sqrt{R^2 + h^2 - 2Rh(1 - e^{-\frac{4Rh}{2R^2}})/(1 + e^{-\frac{4Rh}{2R^2}})}$ (valid for small h/R values).

We see from Figures 3.15 and 3.16 that at typical F_s ratios, our filter's scale and spatial localization characteristics are comparable to those of the Canny ridge operator.

3.9 Results

We have tested our scheme (filter + network) extensively; Figures 3.17 and 3.18

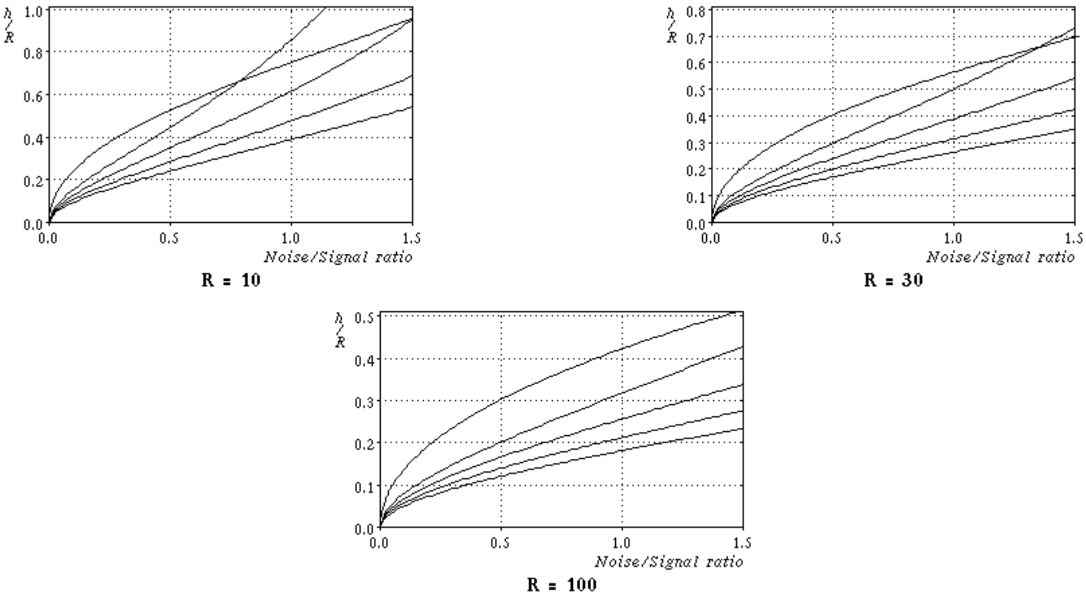


Figure 3.16: Comparison of relative spatial error (h/R) as a function of *noise to signal ratio* ($n_o/(1-s)$) between our filter and the Canny ridge filter. See Equations 3.15 and 3.20. *Top Left:* $R = 10$. *Top Right:* $R = 30$. *Bottom:* $R = 100$. For each graph, the Canny curve is the top curve between $n_o/(1-s) = 0$ and $n_o/(1-s) = 0.5$. The other curves from top to bottom are for: $F_s = 16$, $F_s = 8$, $F_s = 4$ and $F_s = 2$.

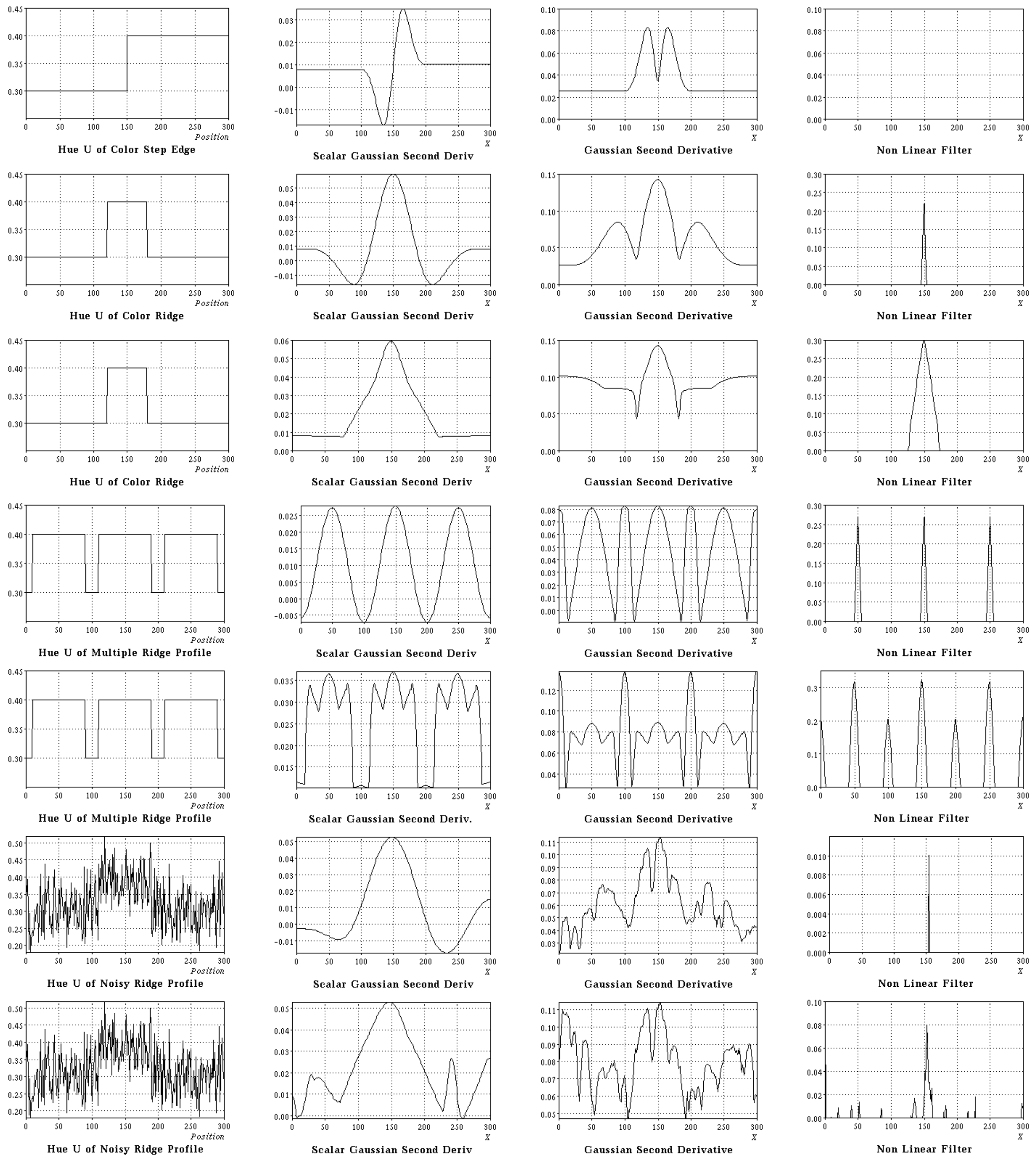


Figure 3.17: *First column:* Different input signals. *Second column:* Output given by second derivative of the gaussian. *Third column:* Output given by second derivative of the gaussian using reference color. *Fourth column:* Output given by our ridge detector. The *First, Second, Fourth and Sixth rows* are results of a single scale filter application where σ is tuned to the size of the largest ridge. The *Third, Fifth and Seventh rows* are results of a multiple scale filter application. Note that no scale parameter is involved in any multiple-scale case.

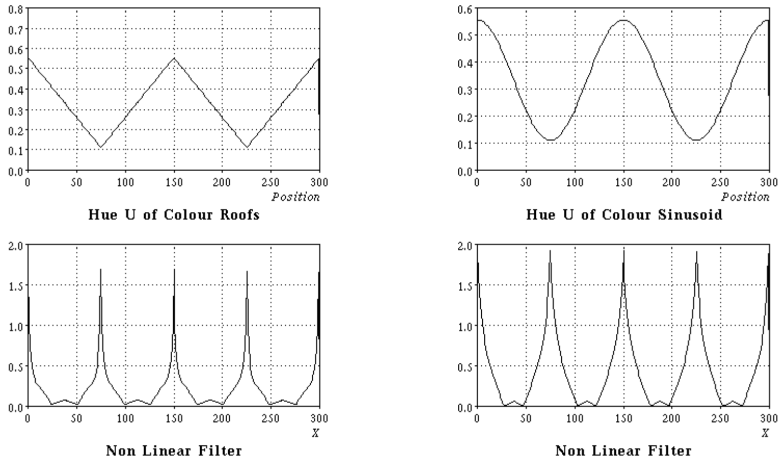


Figure 3.18: Comparing multiple scale filter responses to color profiles. *Top:* Hue U channel of roof and sinusoid color profiles. *Bottom:* Multi-scale output given by *color* convolution of our non-linear mask with the color profiles. Even though our filter was designed to detect flat regions, it can also detect other types of regions.

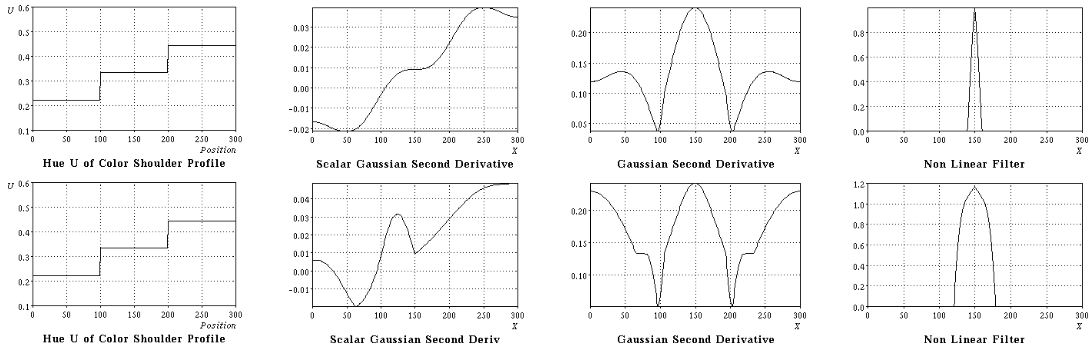


Figure 3.19: *First column:* Multiple step input signal. *Second column:* Output given by second derivative of the gaussian. *Third column:* Output given by second derivative of the gaussian using reference color. *Fourth column:* Output given by our ridge detector. The *first row* shows results of a single scale filter application where σ is tuned to the size of the largest ridge. The *second row* shows results of a multiple scale filter application. Note that no scale parameter is involved in multiple-scale case.

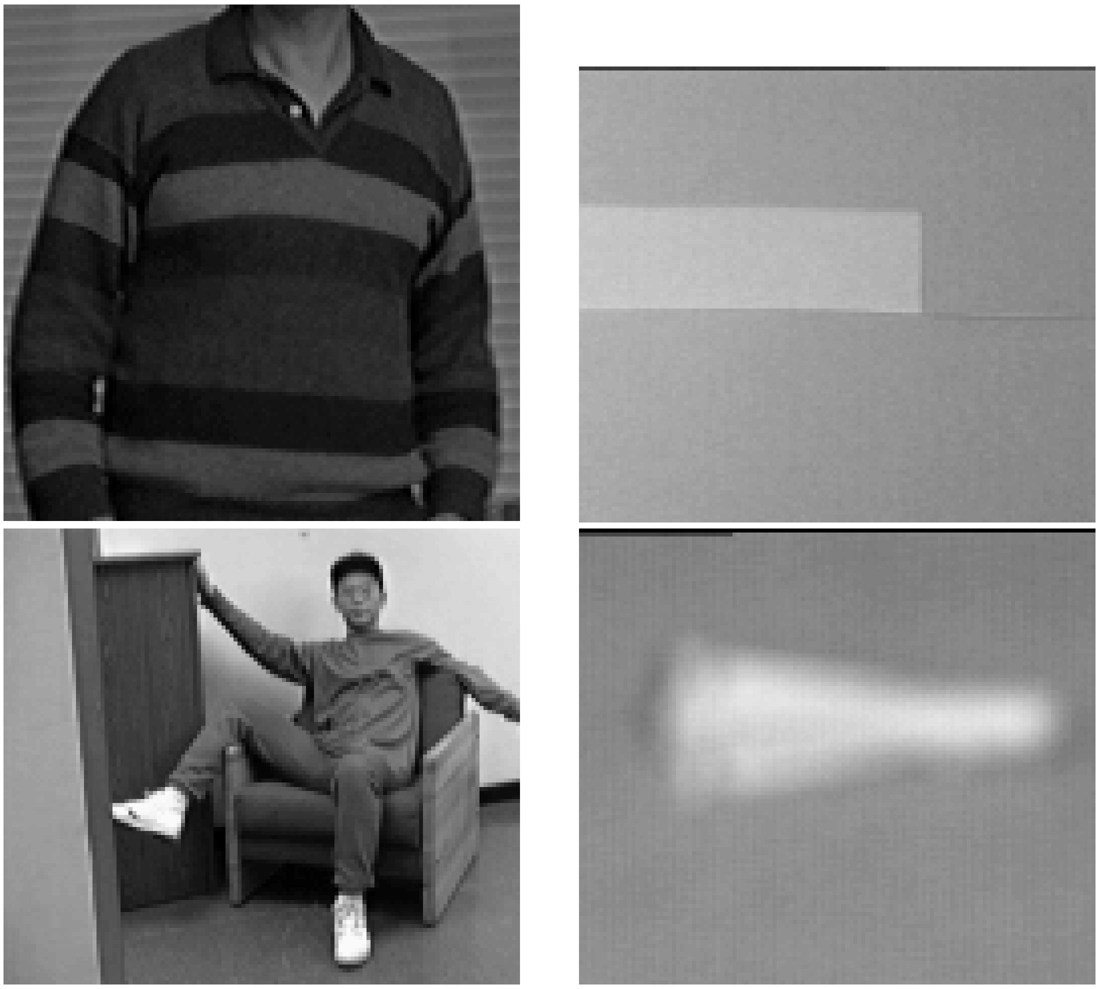


Figure 3.20: Four images. *Left to Right:* Sweater image, Ribbons image, Person image and Blob image. See inertia surfaces for these images in Figures 3.21 and 3.22 and the Canny edges at different scales for the Person and Blob image in Figure 3.5. Note that our scheme recovers the Person and blob at the right scale, without the need of specifying the scale.

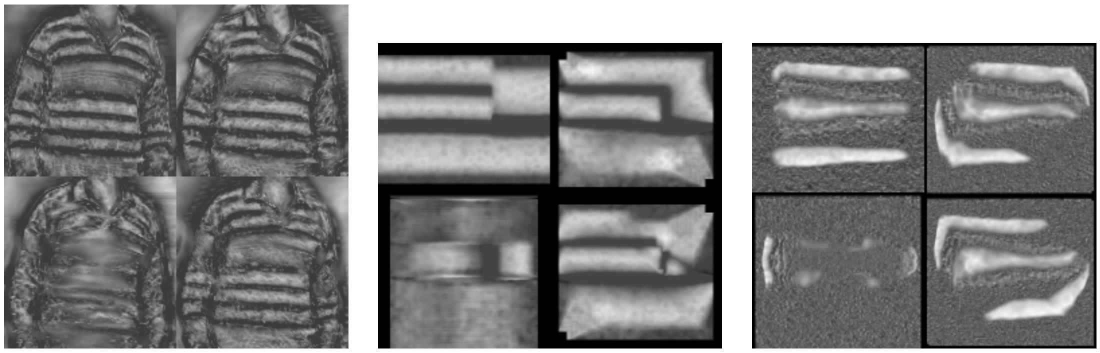


Figure 3.21: Inertia surfaces for three images at four orientations (clockwise 12, 1:30, 3 and 4:30). Note that *exactly* the same lisp code (without changing the parameters) was used for all the images. *From Left to Right:* Shirt image, Ribbon image, Blob image.

show that our filter produces sharper and more stable ridge responses than the second derivative of a gaussian filter, even when working with the notion of reference colors for color ridge profiles. First, our filter localizes all the ridges for a single ridge, for multiple or step ridges and for noisy ridges. The second derivative of the gaussian instead fails under the presence of multiple or step ridges. Second, the scale chosen by our operator matches the underlying data closely while the scale chosen by the second derivative of the gaussian does not match the underlying data (see Figures in Section 3.8). This is important because the scale is necessary to compute the tolerated length which is used in the second stage of our scheme to find the Curved Inertia Frames of the image. And third, our filter does not respond to edges while the second derivative of the gaussian does.

In the previous paragraph, we discussed the one-dimensional version of our filter. The same filter can be used as a directional ridge operator for two-dimensional images. Figure 3.21 shows the directional output (a.k.a. inertia surfaces) of our filter on four images. The two-dimensional version of the filter can be used with different degrees of elongation. In our experiments, we used one pixel width to study the worst possible scenario. An elongated filter would smooth existing noise; however, large scales are not good because they smooth the response near discontinuities and in curved areas of the shape (this can be overcome by using curved filters [Malik and Gigus 1991]).

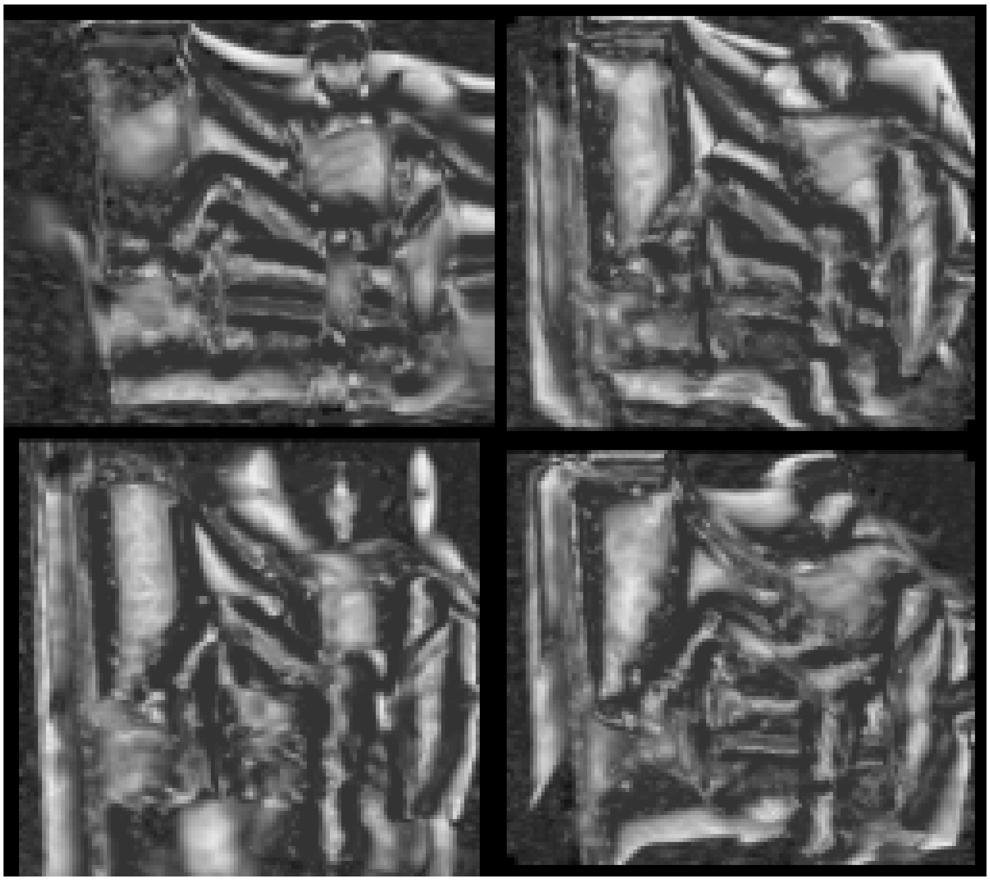


Figure 3.22: Inertia surfaces for the person image at four orientations. Note that *exactly* the same lisp code (without changing the parameters) was used for these images and the others shown in this Chapter.

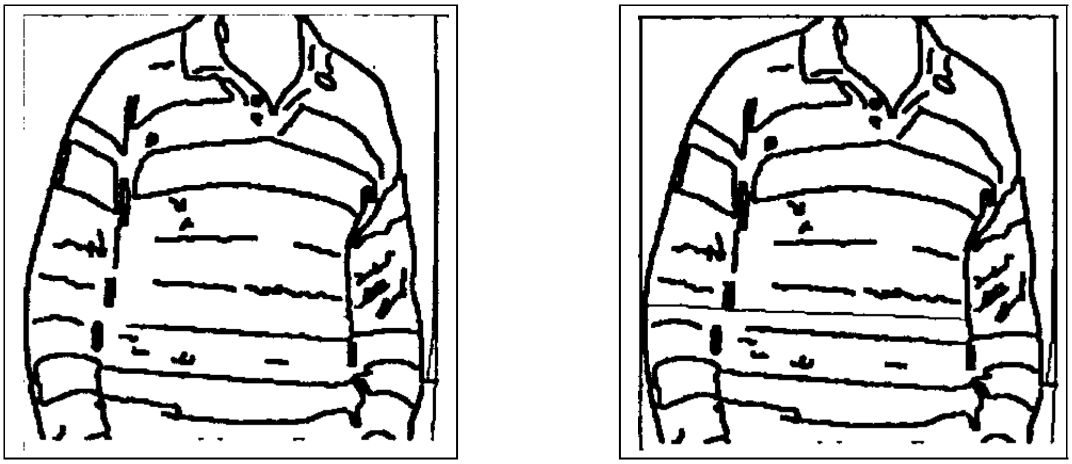


Figure 3.23: Most salient Curved Inertia Frame obtained in the shirt image. Note that our scheme recovers the structures at the right scale, without the need to change any parameters. *Left:* Edge map of shirt image without most salient curved inertia frame. *Right:* With most salient curved inertia frame superimposed.

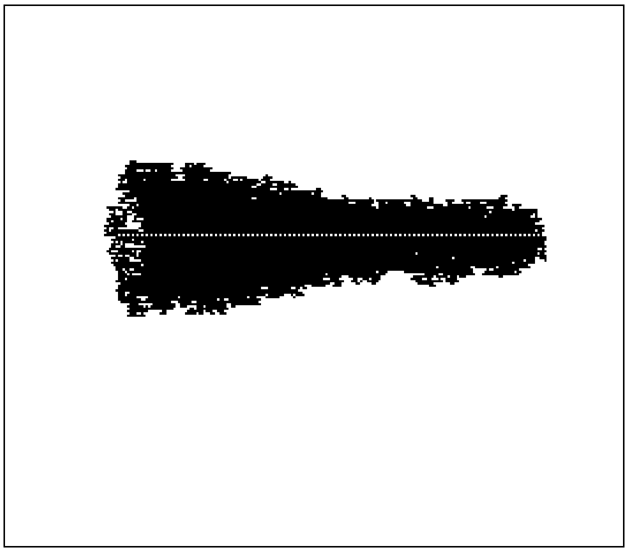


Figure 3.24: Blob with skeleton obtained using our scheme in the blob image. Note that our scheme recovers the structures at the right scale, without the need to change any parameters.

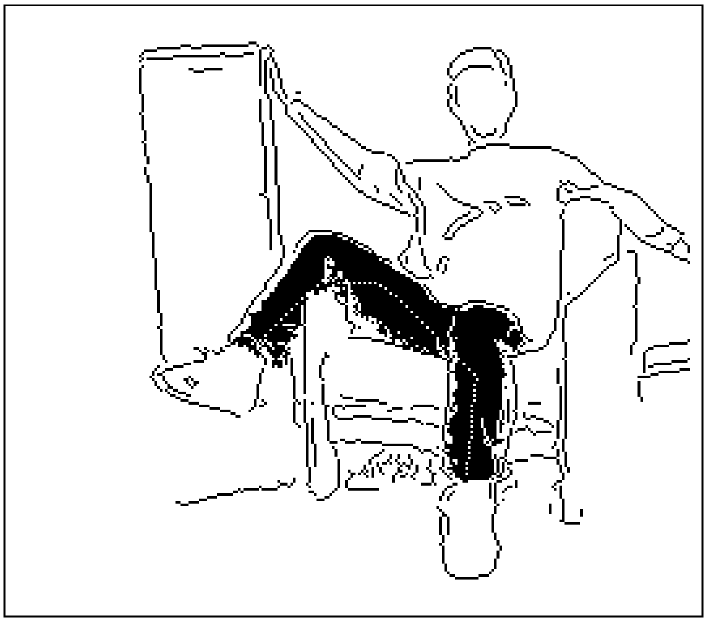


Figure 3.25: Pant region obtained in person image. The white curve is the Curved Inertia Frames from which the region was recovered.

The inertia surfaces and the tolerated length are the output of the first stage of our scheme. In the second stage, we use these to compute the Curved Inertia Frames (see Chapter 2) as shown in Figures 3.23, 3.24, 3.25, 3.26, and 3.27. These skeleton representations are used to grow the corresponding regions by a simple region growing process which starts at the skeleton and proceeds outward (this can be thought of as a visual routine [Ullman 1984] operating on the output of the dynamic programming stage or skeleton sketch). This process is stable because it can use global information provided by the frame, such as the average color or the expected size of the enclosing region. See Figures 3.23, 3.24, 3.25, 3.26, and 3.27 for some examples of the regions that are obtained. Observe that the shapes of the regions are accurate, even at corners and junctions. Note that each region can be seen as an individual test since the computations performed within it are independent of those performed outside it.

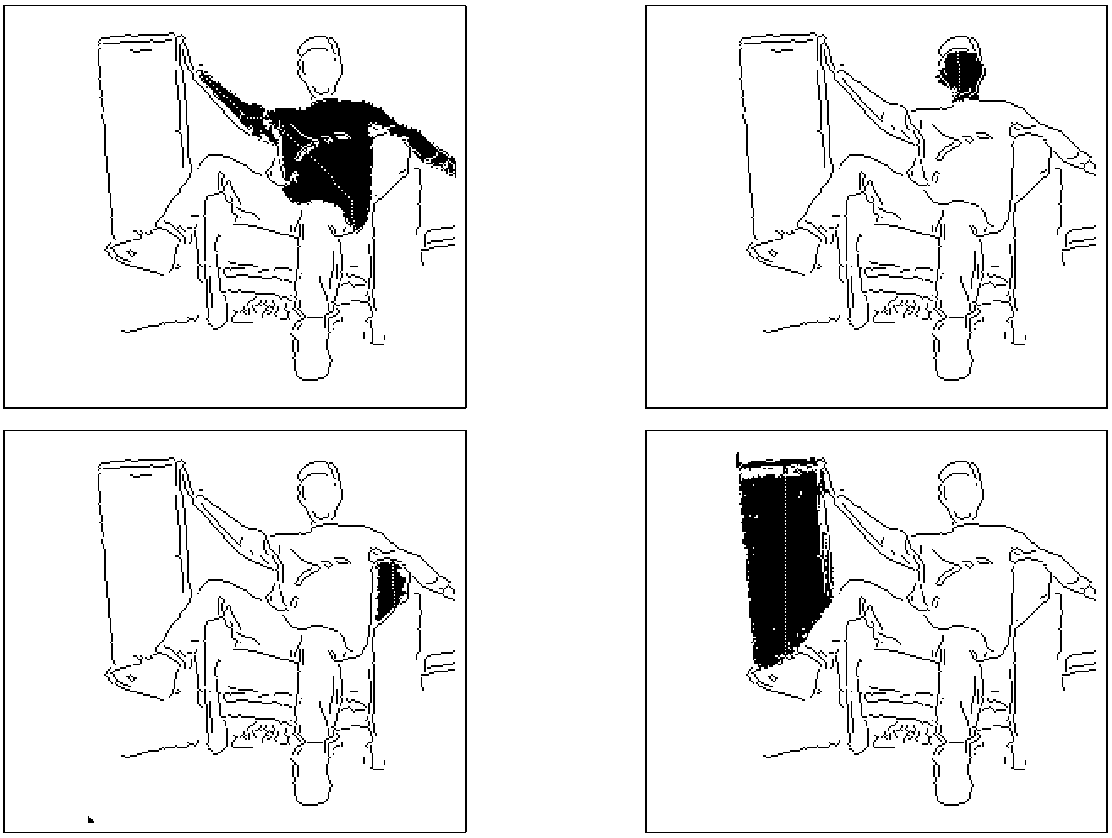


Figure 3.26: Four regions obtained for the person image. The white curves are the Curved Inertia Frame from which the regions where recovered.

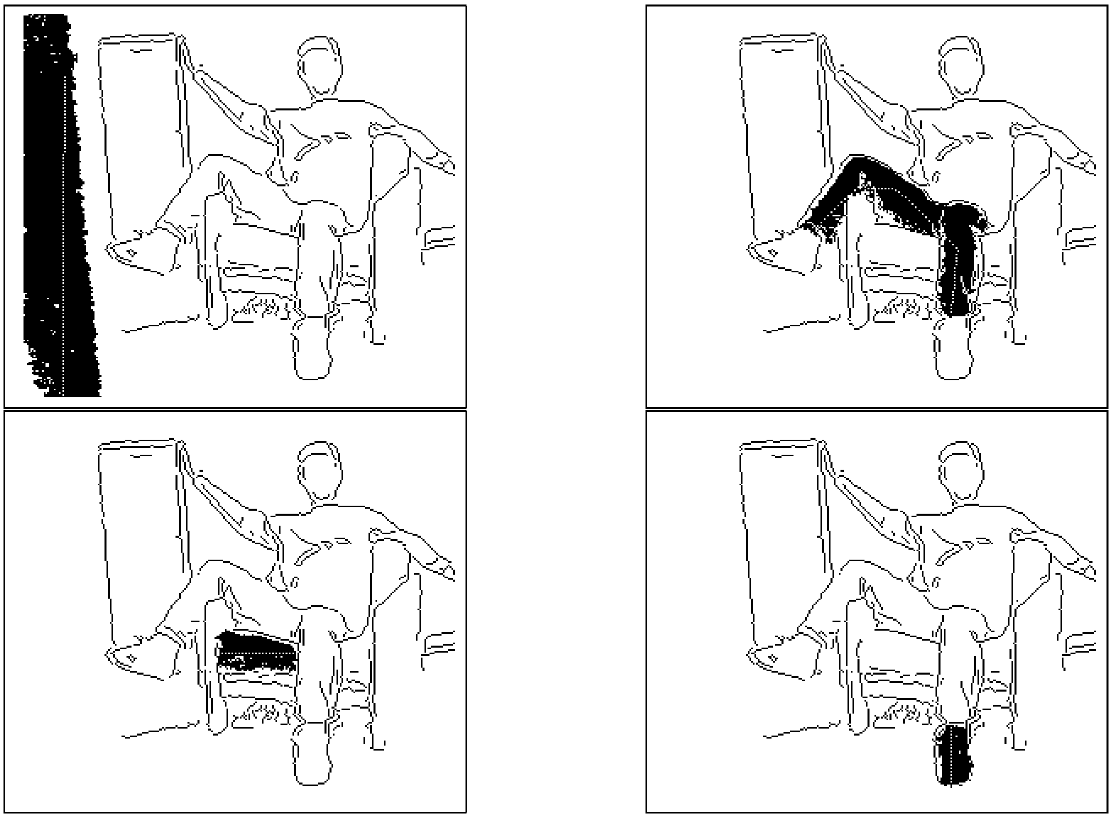


Figure 3.27: Four other regions obtained for the person image. The white curves are the Curved Inertia Frames from which the regions were recovered.

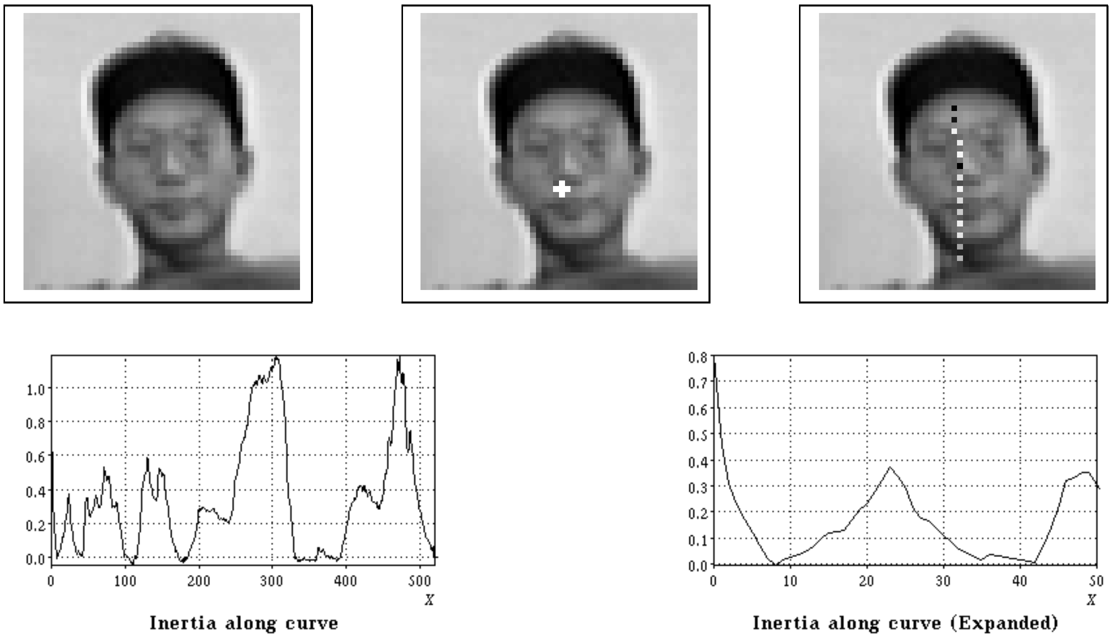


Figure 3.28: This Figure illustrates how the scheme can be used to guide attention. *Top left:* Close up image of face. *Top center:* Maximum inertia point. *Top right:* Skeletal curve through face. *Second row:* Inertia map along entire skeletal curve and along face.

3.10 Discussion: Image Brightness Is Necessary

We have implemented our scheme for color segmentation on the Connection Machine. The scheme can be extended naturally to brightness and texture (borrowing from the now popular filter-based approaches applied to the image, see [Knuttson and Granlund 1983], [Turner 1986], [Fogel and Sagi 1989], [Malik and Perona 1989], [Bovik, Clark and Geisler 1990], [Thau 1990]). The more cues a system uses, the more robust it will be. In fact, image brightness is crucial in some situations because luminance boundaries do not always come together with color boundaries (e.g. cast shadows).

But, should these different schemes be applied independently? Consider a situation in which a surface is defined by an iso-luminant color edge on one side and by a brightness edge (which is not a color edge) on the other. This is in fact the case for one of the shirt image's ribbons. Our scheme would not recover this surface because the two sides of our filter would fail (on one side for the brightness module and on the other for the iso-luminant one). We believe that a combined filter should be used to obtain the inertia values and the tolerated length in this case. The second stage would then be applied only to one set of values. Instead of having a filter with two sides, our new combined filter should have four sides. Two responses on each side, one for color $R_{c,i}$ and one for brightness $R_{b,i}$, the combined response would then be:

$$min(max(R_{b,left}, R_{c,left}), max(R_{b,right}, R_{c,right})).$$

Contour Texture and Frame Curves

Chapter 4

4.1 Introduction

Can we use Curved Inertia Frames and frame alignment to recognize contour textures such as hair, waves, clouds, complex tools, and house ornaments? If so, how should we change the scheme developed in Chapters 2 and 3? Our work is built on the premise that two non-rigid objects may require different recognition strategies. In fact, we believe that a different frame alignment technique is needed to recognize contour textures.

This Chapter addresses contour textures and presents a filter-based scheme to recognize Contour Texture. The scheme can be seen as an instance of frame alignment, as described in Section 1.5 (see Figure 1.9), and a skeleton computation similar to the one presented in Chapter 3 can still be used to recover the main axis of the contour.

4.2 Contour Texture and Non-Rigid Objects

Oak leaves are readily distinguished from other types of leaves (see Figure 4.1). The ability to distinguish leaves can not be attributed solely to an exact shape property since the leaf contours change significantly from one leaf to another. Instead, another property, more statistical in nature, must be used. We call this property contour texture. Contour texture has not been studied extensively in the past and is the subject of this Chapter. Most of the work has been on recognizing repetitive one dimensional patterns that are fully connected [Granlund 1972], [Zahn and Roskies 1972], [Nahin 1974], [Richard and Hemami 1974], [Eccles and Mc Queen and Rosen 1977], [Giardina and Kuhl 1977], [Person and Foo 1977], [Wallace and Wintz 1980], [Crimmins 1982], [Kuhl and Giardina 1982], [Etesami and Uicker 1985], [Persoon and Fu 1986], [Strat 1990], [Van Otterloo 1991], [Dudek 1992], [Maeder 1992], [Uras and Verry 1992]. We are interested in a more general description that does not rely on connectivity. Lack of connectivity is very common for two reasons. First, edge detectors often break contours and it is hard to recover a connected contour. Second, some shapes such as trees or clouds do not have a well-defined contour but a collection of them.

Contour texture is interesting because there are many non-rigid or complex objects with distinctive contour textures (such as clouds, trees, hair, and mountains) and because images without contour texture appear less vivid and are harder (or even impossible) to recognize (see cartoons in Figure 4.2). In fact, many non-rigid objects have boundaries which can be described as contour textures. Rigid-object recognition schemes do not handle contour textures because they rely on "exact" shape properties. This does not occur in contour textures (or in other non-rigid objects[1]) and alternative approaches for handling this case must be developed. In addition, contour texture may help perceptual organization and indexing schemes (see Figure 5.8).

Not all objects can be described just by their contour textures. Leaves are a good example of this [Smith 1972]. In fact, botanists have divided leaves using two

[1]Contour texture is common in classification problems; however, it is also common in recognition problems where the shapes are complex such as in the skyline of a town.

attributes one of which is based on contour texture (they use the term "leaf margin"). This attribute generates several classes such as dentate, denticulate, incised, or serrulate. Botanists also use another attribute (called "leaf shape") which is complementary to contour texture. Leaf shape categories include oval, ovate, cordate, or falcate (see [Smith 1972] for a complete list). The distinction between contour texture and shape is particularly important for deciding what type of representation to use, a question which we will address in Section 4.5.

Since contour textures appear together with other non-rigid transformations, we are particularly interested in finding a useful and computable shape representation for contour textures. If it is to be useful in a general purpose recognition system, such a representation should support simultaneously recognition techniques for contour texture and other non-rigid transformations. The findings presented in this Chapter argue in favor of a two-level representation for contour textures such that one level, which we call the frame curve, embodies the "overall shape" of the contour and the other, the contour texture, embodies more detailed information about the boundary's shape. These two levels correspond closely to the two attributes used to describe leaves by botanists.

The notion of contour texture prompts several questions: Can we give a precise definition of contour texture? What is the relation between two-dimensional texture and contour texture? Is there a computationally-efficient scheme for computing contour texture? There are several factors that determine the contour texture of a curve: for example, the number and shape of its protrusions. What other factors influence the contour texture of a shape? In particular, does shape influence contour texture?

In this Chapter we suggest a filter-based model for contour texture recognition and segmentation (Figure 5.8). The scheme may also be used for contour completion (Figure 5.6), depth perception (Figure 5.5), and indexing (Figure 5.8). The rest of the Chapter addresses these questions and is organized as follows: In Section 4.3 we discuss the definition of contour texture and its relation to 2D texture. In Sections 4.4 and 4.5 we discuss the relation that contour texture has to scale and inside/outside relations respectively. In Section 4.6 we present an implemented filter-based scheme for contour texture. In the next Chapter, we suggest an application for contour

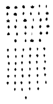


Figure 4.1: Which of these leaves are oak leaves? Some objects are defined by the contour texture of their boundaries. Can you locate the oak leave at the bottom among the other leaves? It is much easier to classify oak leaves from non-oak leaves than to locate individual oak leaves.

texture which serves to illustrate how contour texture may be used for learning.

4.3 Contour Texture and Frame Curves

2D texture has received considerable attention, both in the computational and psychological literature. However, there is no unique definition for it. Roughly speaking, 2D texture is a statistical measure of a two-dimensional region based on local properties. Such properties typically include orientation and number of terminations of the constituent elements (a.k.a. textons).

In this Chapter we argue that contour texture, a related but different concept, is a relevant non-rigid transformation and plays an important role in human visual perception; contour texture can be defined as a statistical measure of a curve based on local properties (See Figure 4.5). We call such a curve the "frame curve". Figure 4.3 shows some contours with different contour textures, all of which have "invisible"

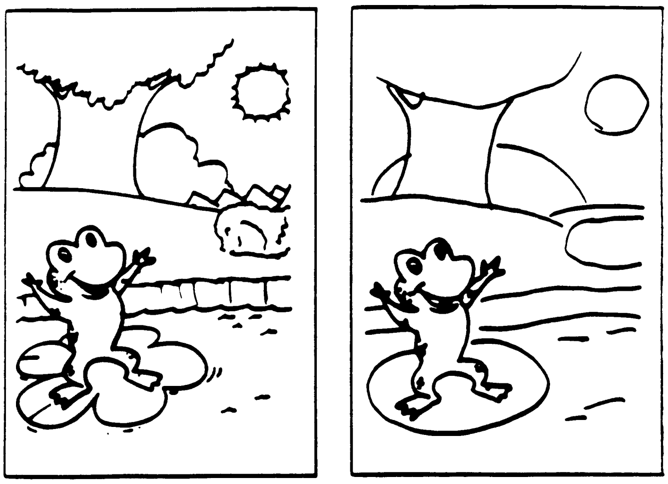


Figure 4.2: *Left:* Cartoon with contour texture. *Right:* Another version of the same cartoon with frame curves; i.e. these two images are identical with the exception that one of them has been drawn by removing the contour texture of its curves. The image without contour texture appears less vivid and has ambiguity. In other words, significant information is lost when the contour texture of a contour is replaced by its frame curve.

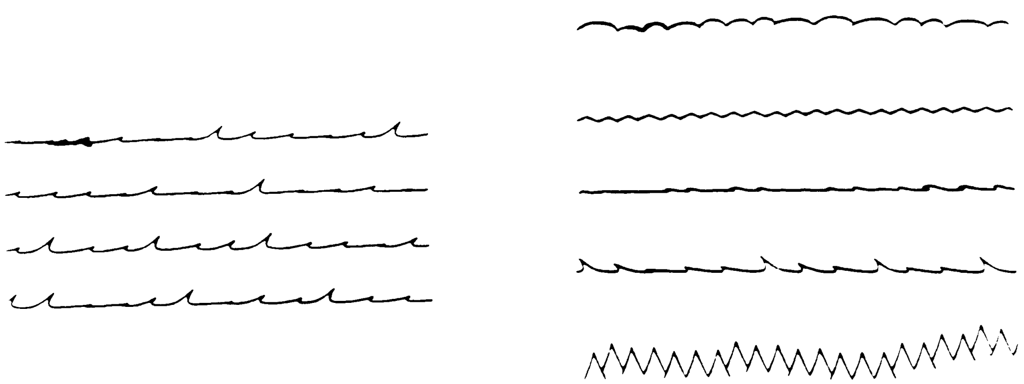


Figure 4.3: *Left:* Different curves with similar contour texture. The frame curve in all these cases is a horizontal line. *Right:* Some curves with different contour textures.

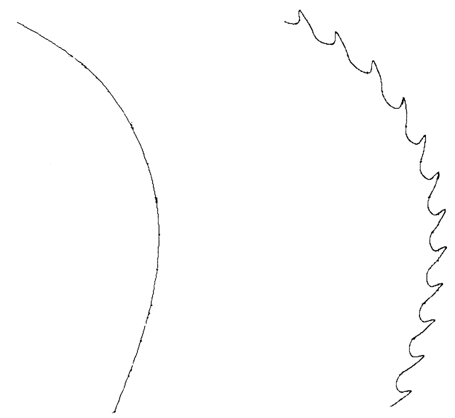


Figure 4.4: *Left:* A curved frame curve. *Right:* The same frame curve with a contour texture (drawn by our implemented contour texture generator).

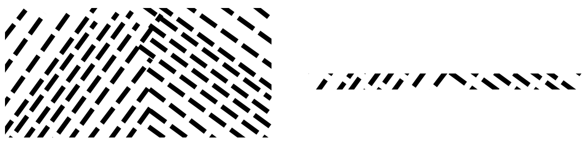


Figure 4.5: *Left:* An image with a vertical two-dimensional texture discontinuity. The discontinuity is defined by the average orientation of the segments near a point in the image. Such orientation is different for the two regions surrounding the central vertical line. *Right:* The tilted segments in this image define a horizontal line. A contour texture discontinuity in such a line is perceived in the middle of it. The discontinuity is defined also by the average orientation of the segments surrounding a point. One of the differences between contour texture and two-dimensional texture is that the statistics are computed over a curve in one case and on a two-dimensional region on the other. Other differences and similarities are discussed in the text.

horizontal lines as frame curves. The contours were drawn by an implemented contour texture generator which takes as input a sample drawing of a COntour TExture ELement or "Cotel" (akin to texton and protrusion[2]) and produces as output a concatenation of one or more of these cotels subject to certain random transformations. The program draws the cotels using as a base a frame curve drawn by the user.

The notion of a frame curve as presented here is closely related to the one presented in [Subirana-Vilanova and Richards 1991] (see also Appendix B). A frame curve is defined there as *a virtual curve in the image which lies in "the center" of the figure's boundary.* In the context of this Chapter, the whole contour texture is the figure. Note that the figure is defined there as the collection of image structures supporting visual analysis of a scene.[3]

A frame curve can also be used for other applications. For example, a frame curve can be used as a topological obstruction to extend size functions [Uras and Verri 1992] to non-circular shapes. As another example, frame curves can be used to compute a part-description of a shape as shown in Appendix B.

4.4 Inside/Outside and Convexity

There are several factors that determine contour texture. In this Section we argue that the side of the contour perceived as inside influences contour texture perception. Consider the examples in Figure 4.8. The left and right stars in the third row have similar outlines since one is a reversed[4] version of the other. They are partially smoothed versions of the center star but each of them looks different from the others; in fact, the left one is more similar to the center star ($N > 20$) despite the fact that both have the same number of smoothed corners. We first made this

[2]This does not mean that we support the texton model of 2D texture perception. We include it as a way to clarify the similarities between contour texture and 2D texture.

[3]This definition is in direct conflict with the classical definition of figure-ground since it is based on attention as oppose to depth. It is for this reason that the Appendix, an updated version of [Subirana-Vilanova and Richards 1991], renames "figure" the attentional frame (to avoid confusions with the classical definition of figure).

[4]By "reversed" we mean that a mirror image of one of the two contours around the frame curve yields the other.

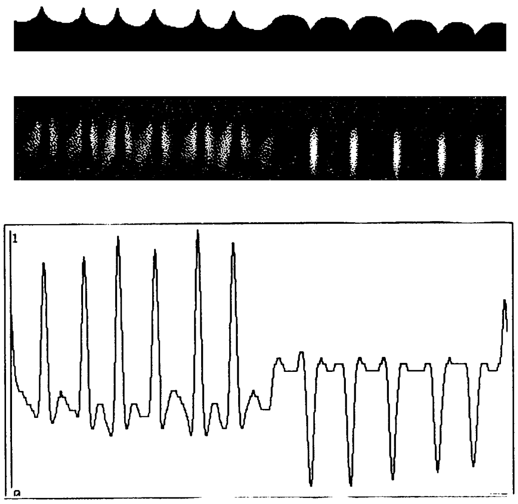


Figure 4.6: *Top:* contour. *Second row:* Normalized output of the selected filter. *Third row:* Cross section of the filter output along the *frame curve*, horizontal in this image, before post-processing. The post-processing removes the peaks by spreading the maxima of the filter output. This contour is interesting because one side is a reversed version of the other yet a clear boundary may be perceived (as if one was "pointy" and the other "smooth"). See text for details.

observation in [Subirana-Vilanova and Richards 1991] and proposed that it is due to a bias which makes the outside of the shapes more "salient"[5].

In the context of contour texture, the findings of [Subirana-Vilanova and Richards 1991] imply that the contour texture of a shape depends on which side is perceived as inside[6]. Their findings suggest that inside/outside relations are computed before contour texture descriptions can be established, and agree with a model which starts by computing an attentional reference frame. Such a frame includes the frame curve and an inside/outside assignment. The model that we present in Section 4.6 conforms to these suggestions.

The inside and outside of a contour are not always easy to define. What is the outside of a tree? Is the space between its leaves inside or outside it? In fact, inside/outside and contour texture are closely related to non-rigid boundaries. In Appendix B we look at this connection more in detail from the point of view of human perception.

4.5 The Role of Scale and Complexity in Shape and Contour Texture

As mentioned in Section 4.1, the notion of contour texture is meant to be used in the differentiation of shapes belonging to different perceptual categories (e.g. an oak vs. an elm leaf) and not to distinguish shapes belonging to similar perceptual categories (e.g. two oak leaves). This raises the following questions: Are two types of representations (shape and contour texture) necessary? When are two objects in the same category? When is a contour texture description appropriate? We address these questions later in the Chapter by presenting an implemented contour texture scheme designed to determine contour similarity based on contour texture descriptors.

---

[5]This bias can be reversed depending on the task, see [Subirana-Vilanova and Richards 1991].

[6]See [Subirana-Vilanova and Richards 1991] or Appendix B for a more detailed discussion on the definition of inside/outside relations, and on the influence of convexity.

In this Section we argue that the difference between shape and contour texture is relevant to computer vision (regardless of implementation details) and, in particular, that it is important to find schemes which automatically determine whether a shape "is" a contour texture or not. For contour textures, it is also important to embody both representations (shape *and* contour texture) for every image contour. At the beginning of this Chapter, we have already presented one of our strongest arguments: Some shapes can not be distinguished by exact shape properties (while others can).

We will now present three other psychological observations that support this difference. First, studies with pigeons have shown that they can discriminate elements with different contour textures but have problems when the objects have similar contour textures [Herrnstein and Loveland 1964], [Cerella 1982]. This suggests that different schemes may be needed for the recognition of shape and contour texture.

Second, consider the object in the top of Figure 4.7. Below the object, there are two transformations of it: the left one is a pictorial enlargement, and the right one is an enlargement in which the protrusions have been replaced by a repetition of the contour (preserving the contour texture). The shape on the left appears more similar to the one on the right[7]. We contend that this is true in general if the shapes have a small number of protrusions (i.e. their "complexity" is low). In these cases, contour texture does not seem to have an important role in their recognition[8]. However, when the shapes are more complex (see bottom-most three shapes in Figure 4.7), the similarity is not based on an exact pictorial matching. Instead, the enlarged shape with the same contour texture is seen as more similar. For "complex" shapes, the visual system tends to abstract the contour texture from the shape and the "enlargement" of such a property is done at a symbolic level. In addition to supporting the distinction between contour texture and shape (first question above), this observation suggests that complexity and scale play a role in determining what type of description (shape or contour texture) should be used in each case: simple shapes are fully represented;[9] and complex shapes are represented with abstract contour texture descriptors. [Goldmeier 1972], [Kertesz 1981], [Hamerly and Springer 1981],

[7]No detailed experiment was performed and the intuitions of the reader will be relied upon.

[8]The contour texture may play an intermediate role in the recognition of the shape by helping indexing.

[9]In the sense that the location of all boundary points are memorized

[Palmer 1982] and [Kimchi and Palmer 1982] each performed similar experiments on a two-dimensional texture version of the problem. [Goldmeier 1972] also presents some one-dimensional contour-texture-like examples (using the notions presented here) which support the role of complexity described here.

The third study which agrees with our distinction between contour texture and shape is that of [Rock, Halper, and Clyton 1972]. They showed subjects a complex figure and later showed them two figures which had the same overall shape and contour texture (using our terms), but only one of which was exactly the same. The subjects were to find which was the previously seen shape. They found that subjects performed only a slightly better than random. This suggests, again, that they were just remembering the overall shape and an abstract description of the contour texture of the boundary's shape. When subjects were presented with non-complex versions of the same shapes the distinctions were based on the exact shapes themselves, which agrees with the model given here.

4.6 A Filter-Based Scheme

The definitions of contour texture and two-dimensional texture, given in Section 4.3, point out some of the relationships between them: both notions are based on statistics of local properties. However, they differ in the extent of such statistics - a curve for contour texture and a surface for two-dimensional texture. In fact, most existing schemes for two-dimensional textures can be applied, after some modifications, to contour texture. Some of the problems that have to be solved in doing so are the computation of frame curves and inside/outside relations.

Thus, it is worthwhile to review work on 2D texture. Theories of two-dimensional texture are abundant, but we will mention just a few. Preattentive texture discrimination has been attributed to differences in nth-order statistics of stimulus features such as orientation, size, and brightness [Julesz and Bergen 1983], [Julesz 1986], [Beck 1982], and [Voorhees and Poggio 1988]. Other theories have been proposed, especially ones that deal with repetitive textures (textures in which textons are similar and on a regular pattern [Hamey 1988]), such as Fourier Transform based models

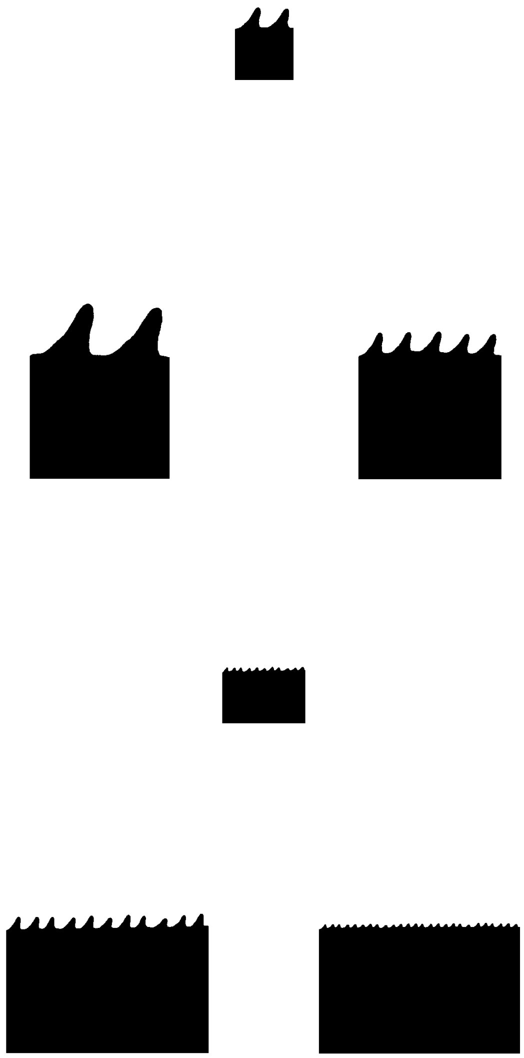


Figure 4.7: This Figure provides evidence that for simple objects like the top one the matching across scales is done pictorially (see second row). For more complex shapes, on the other hand, such as the one on the third and fourth rows, the matching is performed by maintaining the contour texture description. See text for details.

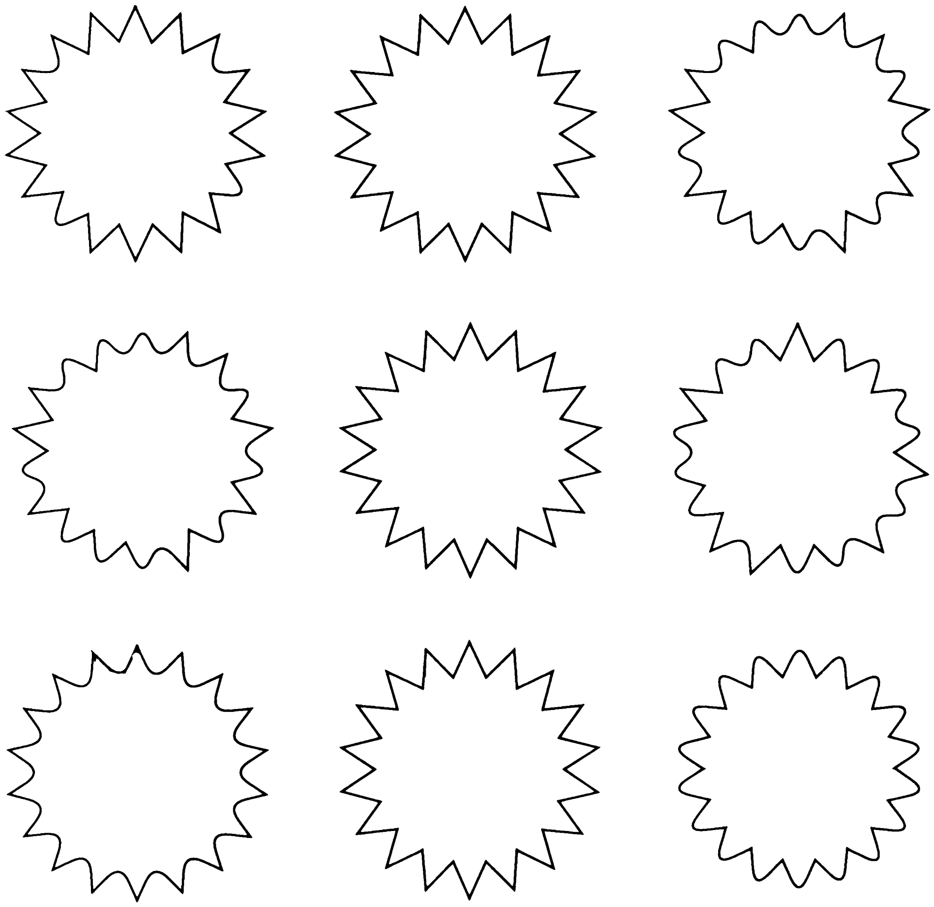


Figure 4.8: *Top row:* Use the middle pattern as reference. Most see the left pattern as more similar to the reference. This could be because it has a smaller number of modified corners (with respect to the center) than the right one, and therefore, a pictorial match is better. *Second row:* In this case, the left and right stars look equally similar to the center one. This seems natural if we consider that both have a similar number of corners smoothed. *Third row:* Most see the left pattern as more similar despite the fact that both, left and right, have the same number of smoothed corners with respect to the center star. Therefore, in order to explain these observations, one can not base an argument on just the number of smoothed corners. The positions of the smoothed corners need be taken into account, i.e. preferences are not based on just pictorial matches. Rather, here the convexities on the *outside* of the patterns seem to drive our similarity judgement. (These Figures were taken from [Subirana-Vilanova and Richards 1991].)

[Bajcsy 1973] and histogramming of displacement vectors [Tomita, Shirai, and Tsuji 1982]. All of these theories tend to work well on a restricted set of textures but have been proved unable to predict human texture perception with sufficient accuracy in all of its spectrum. In addition, it is unclear how these schemes could compute frame curves or inside/outside relations, specially in the presence of fragmented and noisy contours.

Another popular approach has been to base texture discrimination on the outputs of a set of linear filters applied to the image (see [Turner 1986], [Montes, Cristóbal, and Bescós 1988], [Fogel and Sagi 1989], [Malik and Perona 1989], and [Bovik, Clark, and Geisler 1990]). These approaches differ among themselves on the set of selected filters and/or on the required post-processing operations. A purely linear scheme can not be used (see for example [Malik and Perona 1989]), justifying the need for non-linear post-processing operations. [Malik and Perona 1989] compare the discriminability in humans to the maximum gradient of the post-processed output of the filters they use, and find a remarkable match among them. The approach is appealing also because of its simplicity and scope, and because it is conceivable that it may be implemented by cortical cells. In addition, there exists a lot of work on filter based representations for vision [Simoncelli and Adelson 1989], [Simoncelli, Freeman, Adelson, and Heeger 1991].

Some work exists on curve discrimination which could be applied to contour texture discrimination. However, previous approaches are designed to process fully connected curves [Granlund 1972], [Zahn and Roskies 1972], [Nahin 1974], [Richard and Hemami 1974], [Eccles and Mc Queen and Rosen 1977], [Giardina and Kuhl 1977], [Person and Foo 1977], [Wallace and Wintz 1980], [Crimmins 1982], [Kuhl and Giardina 1982], [Etesami and Uicker 1985], [Persoon and Fu 1986], [Strat 1990], [Van Otterloo 1991], [Dudek 1992], [Maeder 1992], [Uras and Verry 1992]. Our model, instead, works directly on images and does not require that the contour be fully connected. The ability to process the contour directly on the image enables the scheme to extend naturally to fragmented curves and to curves without a single boundary (e.g. a contour composed of two adjacent curves).

Our scheme segments and recognizes the curves based on their contour texture and consists of the following steps:

1. Find the frame curves of the contour to be processed.
2. Decide which is the inner side of the contour and color (label) it.
3. Filter the image I with a set of oriented and unoriented filters F_i at different scales, which yields $I * F_i^+$ and $I * F_i^-$, the negative and positive responses to the filters[10]. We have used the same filters used by [Malik and Perona 1989].
4. Perform nonlinear operations on the outputs obtained, such as spreading the maxima and performing lateral inhibition (in the current implementation).
5. Normalize the orientation of the directional filters to the orientation of the frame curve's tangent.

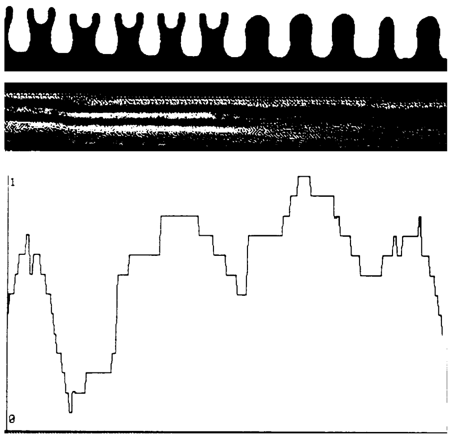


Figure 4.9: *Top:* Colored contour. *Second row:* Normalized output of one filter (horizontal odd-symmetric). *Third row:* Cross section of the processed filter output along the frame curve, horizontal in this image. We have chosen to display a filter which does not yield a discontinuity if we spread the maxima. However, if we spread the number of maxima, a discontinuity appears. The discontinuity in this case can also be found by spreading the maxima of another filter.

Contour texture discontinuities can be defined as places of maximum gradient (along the direction of the frame curve) in the obtained responses, and recognition can be done by matching such responses. Steps 3, 4, and 5 have been implemented on the Connection Machine and tried successfully on a variety of segmentation examples (see Figures 4.9, 4.10, and 4.6).

[10]We used filters similar to those used in [Malik and Perona 1989] in the context of 2D texture. All the examples shown here where run at 4 x 1.5 deg. Note that it is unclear though if only even symmetric filters are needed for Contour Texture as proposed there for 2D texture.

4.6.1 *Computing Frame Curves*

Finding the frame curve is not straight forward. The natural solution involves smoothing [Subirana-Vilanova 1991] but has problems with the often-occurring complex or not fully connected curves.

However, frame curves tend to lie on the ridges of one of the filter's response. This suggests that frame curves can be computed by a ridge detector that can locate long, noisy, and smooth ridges of variable width in the filter's output. One such approach, Curved Inertia Frames, was presented in Chapter 3. Note that computing ridges is different from finding discontinuities in the filter's response, which is what would be used to compute two-dimensional texture discontinuities in the schemes mentioned above.

Therefore, our model uses the filters of step 3 twice, once before step 1, where we compute the frame curve, and once after step 2, to compute contour texture descriptors.

4.6.2 *Coloring*

Step 2, coloring, is needed to account for the dependence of contour texture on the side perceived as inside, as discussed above (see Figure 4.6). Coloring may also be useful in increasing the response of the filters when the contrast is low. However, coloring runs into problems if the contour is not fully connected or if the inner side of the contour is hard to determine. Possible alternatives include using the frame curve as a basis to spread and stop the coloring, and enlarging the size of the contours to increase the response of the filters used in the third step.

4.7 Discussion

In this Chapter, we have proposed an image descriptor, contour texture, that can be used for real-time contour classification in several applications such as long-range tracking, segmentation, and recognition of some non-rigid objects. Contour textures can not distinguish any pair of objects but require almost no computation time and are necessary to distinguish some non-rigid transformations. In Chapter 5 we suggest several applications of contour texture, among them its use as a learnable feature for robot navigation, indexing, and recognition.

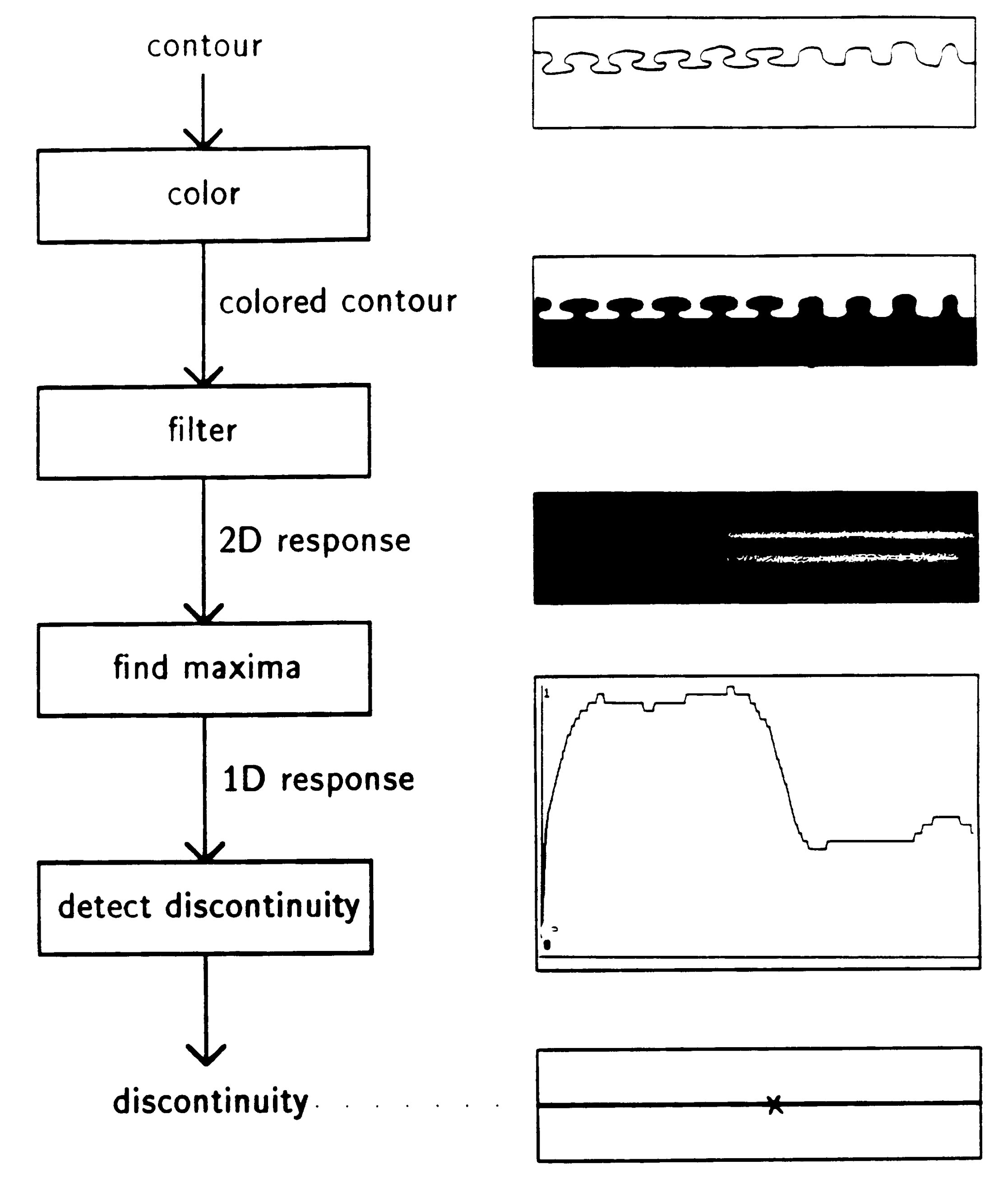


Figure 4.10: Stages of filter-scheme proposed. See text for details.

Conclusion and Future Research

Chapter 5

5.1 Discussion

The starting point of this research was that visual object recognition research to date has focused on rigid objects. The study of the recognition of rigid objects leaves open issues related to the recognition of non-rigid objects. We have presented different types of non-rigid objects, mid-level computations that work in the presence of non-rigid objects, and clarified the role of non-rigid boundaries.

In this thesis we have introduced frame alignment, a computational approach to recognition applicable to both rigid and non-rigid objects. The scheme has been illustrated on two types of non-rigid objects, elongated flexible objects and contour textures, and has three stages. In the first stage, the framing stage, an anchor curve is computed. We have given two different names to such a curve: "frame curve" in contour texture, and "skeleton" in elongated flexible objects. The framing stage works bottom-up, directly on the image, and performs perceptual organization by selecting candidate frame curves for further processing. In the second stage, the unbending stage, such a curve is used to "unbend" the shape. This stage is not necessary in rigid objects. In the third stage, the matching stage, the unbent description is used for matching. In this thesis we have concentrated on Curved Inertia Frames which can be used in the framing stage. In particular, we have illustrated how C.I.F. can find skeletons for elongated objects.

Frame alignment leads to a shape representation with two-levels: a part description capturing the large scale of the shape, and a complementary boundary description capturing the small scale (see Chapter 4 and Section 4.5). The former is computed directly by C.I.F. and the latter by contour texture filters. Such a description makes explicit the difference between contour texture and shape and may be used to simultaneously support several non-rigid transformations of the same object.

The methodology used in this thesis is summarized in Figure 5.1. The methodology is based on studying non-rigid objects by working on four aspects. First, identify a useful application in which to test work on a non-rigid transformation. Second, search for physical models reflecting the underling nature of the objects in consideration. Third, research mid-level computations that can recover global structures useful in the recognition of the objects. Fourth, incorporate signal processing techniques into the mid-level computations (rather than working on the output of early vision mechanisms).

In the rest of this Chapter we review more in detail the the novel findings of this dissertation (Section 5.2) and give suggestions for future research (Section 5.3).

5.2 What's New

In this Section we will review the two areas addressed in this thesis: Non-rigid object recognition (Section 5.2.1) and mid-level vision (Section 5.2.2).

5.2.1 *Recognition of non-rigid objects and frame alignment*

Elongated and Flexible Objects

We have suggested that frame alignment be used to recognize elongated flexible objects by "unbending" them using C.I.F.. We have demonstrated the "unbending" transformation on the simple shapes shown in Figure 2.4. This is useful because flexible objects can be matched as if they where rigid once they have been transformed to a canonical frame position. The canonical orientation needs not be straight. If the objects generally deviate from a circular arc then the canonical representation could store the object with a circular principal axis.

In Appendix B we present evidence against the use of the unbending stage in human perception. The evidence is based on an example in which a bending transformation changes the similarity preferences in a set of three objects. It is important to note that this evidence does not imply that frame curves are not used by human perception. It simply suggests that, in human perception, their role may not be that of frame structures for unbending.

Contour Texture Recognition

Contour texture had received almost no attention in the past, yet we suggest that it plays an important role in visual perception, and in particular, in the shape recognition of some non-rigid or complex objects and possibly in grouping, attention, indexing, and shape-from-contour. We also propose that complex contours, (i.e. non-smooth or disconnected) be represented by abstract contour texture descriptors while simple ones be represented by the detailed location of the contour's points.

A filter-based approach to contour texture is simple and yields useful results in a large number of cases. In addition, we have shown that scale and inside/outside relations play an important role in the perception of contour texture by humans.

5.2.2 *Mid-level vision*

C.I.F.

Curve Inertia Frames is the first computation that recovers candidate discrete curves which are truly global. By global we mean that it provides two warrants: first, there is a mechanism to define a function that associates a given value to any curve in a discrete space; second, there is an algorithm that is guaranteed to find the curve with the highest possible value in such discrete space. Previous approaches provide theoretical justification in the continuous but can only approximate it in the discrete. In contrast, Curved Inertia Frames can be shown to be global in the parallel network where it is implemented. The proof is similar in flavor to well-known proofs in theory of algorithms.

C.I.F. (Curved Inertia Frames) was presented in Chapter 2 and is a novel scheme to compute curved symmetry axes. Previous schemes either use global information, but compute only straight axes, or compute curved axes and use only local information. C.I.F. can extract curved symmetry axes and use global information. This gives the scheme some clear advantages over previous schemes: 1) It can be applied directly in the image, 2) it automatically selects the scale of the object at each image location, 3) it can compute curved axes, 4) it provides connected axes, 5) it is remarkably stable to changes in the shape, 6) it provides a measure associated with the relevance of the axes inside the shape, which can be used for shape description and for grouping based on symmetry and convexity, 7) it can tolerate noisy and spurious data, 8) it provides central points of the shape, 9 it is truly global!.

Curved Inertia Frames is based on two novel measures: the inertia surfaces and the tolerated length. Similar measures may be incorporated in other algorithms such as snakes [Kass, Witkin, and Terzopoulos 88], [Leymarie 1990], extremal regions [Koenderink and van Doorn 1981], [Koenderink 1984], [Pizer, Koenderink, Lifshits, Helmink and Kaasjager 1986], [Pizer 1988], [Gauch 1989], [Gauch and Pizer 1993], dynamic coverings [Zucker, Dobbins, and Iverson 1989], deformable templates [Lipson, Yuille, O'Keefe, Cavanaugh, Taaffe and Rosenthal 1989], [Yuille, Cohen and Hallinan 1989] and physical models [Metaxas 1992] to compute skeletons.

Ridge Detection

In Chapter 3, we have argued that early visual processing should seek representations that make regions explicit, not just edges; furthermore, we have argued that region representations should be computed directly on the image (i.e. not directly from discontinuities). These suggestions can be taken further to imply that an attentional "coordinate" frame (which corresponds to one of the perceptual groups obtained) is imposed in the image prior to constructing a description for recognition (see also

Appendix B).

According to this model, visual processing starts by computing a set of features that enable the computation of a frame of reference. C.I.F. also starts by computing a set of features all over the image (corresponding to the inertia values and the tolerated length). This can be thought of as "smart" convolutions of the image with suitable filters plus some simple non-linear processing.

This has been the motivation for designing a new non-linear filter for ridge-detection. Our ridge detector has a number of advantages over previous ones: it selects the appropriate scale at each point in the image, does not respond to edges, can be used with brightness as well as color data, tolerates noise, and can find narrow valleys and multiple ridges.

The resulting scheme can segment an image without making explicit use of discontinuities and is computationally efficient on the Connection Machine (takes time proportional to the size of the image). The performance of the scheme can in principle be attributed to a number of intervening factors; in any case, one of the critical aspects of the scheme is the ridge-detector. Running the scheme on the edges or using simple gabor filters would not yield comparable results. The effective use of color makes the scheme robust.

Human Perception

In Appendix B we provide evidence against frame alignment in human perception by showing examples in which the recognition of elongated objects does not proceed by unbending the shapes as suggested by our model.

In summary, the implication of the evidence presented in Appendix B, is that frame curves in visual perception are set prior to constructing a description for recognition, have a fuzzy boundary, their outside/top/near/incoming regions are more salient (or not, depending on the task), and that visual processing proceeds by the subsequent processing of convex structures (or holes).

5.3 Future Research

In the following two subsections we will suggest future work related to recognition of non-rigid objects and mid-level vision, respectively.

5.3.1 *Non-rigid object recognition and frame alignment*

Frame alignment can be applied to recognize rigid objects (see Figure 5.11), elongated flexible objects, and contour textures. However, there are other types of non-rigid objects such as handwritten objects, warped objects, and symbolic objects for which the notion of frame curve may not have much use Faces may require several frame curves, rather than one, in order to perform alignment correctly. Figure 5.11 outlines how frame alignment may be used to recognize script.

This thesis has used a mehtodology outlined in Figure 5.1. The Figure also outlines how the methodology used in this thesis may be applied to crumpled objects.

We now present more detailed suggestions for elongated flexible objects and contour textures.

Elongated and Flexible Objects

We have presented evidence that elongated and flexible objects are not recognized in human perception using the unbending transformation. However, this does not mean that frame curves are not used, nor that unbending is not used in a different way. For example, our findings are also consistent with a model in which inside/outside assignments are performed prior to an unbending stage. More experiments could be performed to clarify this issue further. Can we recognize common rigid and elongated objects when they are bent? It is clear that this is possible as is demonstrated by Figure 5.2. Does this recognition depend on the recognition of the local rigid sub-parts of the bent plane? Or does it reflect an unbending transformation?

The unbending stage could benefit from recent progress in image warping using

NON-RIGID TRANSFOR-MATION	APPLICATION	MODELING	ALGORITHM	SIGNAL PROCESSING
Elongated and flexible	Medical imag. People pose estimation as interface Handwriten character rec..	Curved axis of inertia	Dynamic pro-gramming Random networks	Ridge detector Color/brightness
Contour texture	Face database indexing Robotics: Lear-ning world loca-tions - navigation	Neuroscience Fractals Learning	Frame curve Smoothing net-work	Non-linear filter base Texture + early level vision
Crumpling	Recycling: pa-per, plastic, metal Garbage sort Medical imag.	Material sci. Elasticity Graphics	2 1/2 area network Peak localizer	Shape from shading 2 1/2D cues

Figure 5.1: This Figure illustrates the methodology used in this Thesis for investigating the recognition of non-rigid objects (see text for details). The first two rows have already been discussed in previous Chapters. The last row presents a suggestion for future work on crumpling transformations.

skeletons or other geometric transformations [Goldberg 1988].

Contour Texture

We have demonstrated the use of Contour Texture filters in segmentation. Other possible uses include:

- A robot looking at a scene receives over 50Mb of information every second. However, robots often need a much smaller amount of information such as "an object location" or "a direction in which to go". Perceptual schemes that automatically provide this information have proved difficult to construct and most existing approaches do not run on real-time.

  Several real-time systems exist but these are mostly in the area of tracking. In

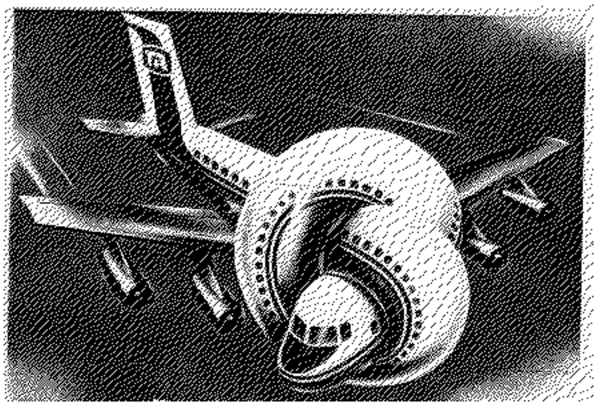


Figure 5.2: Do we recognize this plane by unbending it?

addition, the use of recognition abilities by robots has been generally restricted to rigid objects, small libraries, or off-line applications.

- Contour texture may also be used as an indexing cue in tasks such as face recognition (see Figures 5.4 and 5.3).

- Use contour texture filters to estimate 3D pose[1] (see Figure 5.5). The use of sterable filters may help in estimating a pose and depth using a combination of views.

- Use contour texture as a cue to perceptual organization and attention in combination with other cues such as symmetry and convexity (see Figures 5.8 and 5.7)

5.3.2 *C.I.F. and mid-level vision*

C.I.F. can support many mid-level computations. Its success can be attributed to the fact that it can compute probably global curves directly in the image. Provably global schemes that could recover more complex structures, such as regions or sets of curves, directly in the image would undoubtly surpass the performance of C.I.F..

[1]Note that a lot of work on recovering depth from texture exists [Stevens 1980], [Witkin 1980].

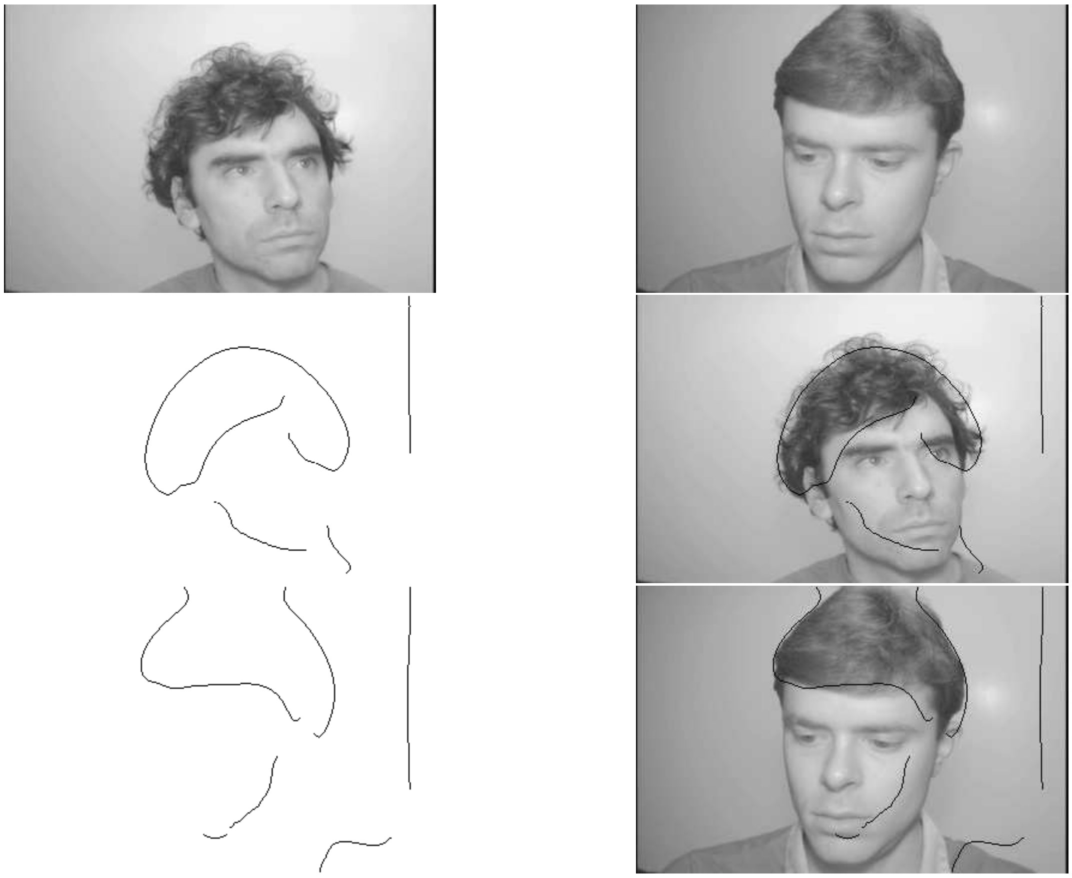


Figure 5.3: *Top:* input images. *Bottom two rows:* frame curves computed with Canny edge detector (left) and frame curves overlaid on input image (right). Face indexing can be done by looking at contour texture descriptors along the frame curves.

Frame curves

It is still unclear to what extent Curved Inertia Frames can be used to recover the frame curve of a contour texture. However, a ridge detector such as the one presented in Chapter 3 may be useful in finding frame curves in the output of contour texture filters.

As we have suggested in Chapter 4, frame curves can be computed on the output of an edge detector by smoothing the contour. Another possibility may be to use Curved Inertia Frames to "smooth" the curve. This may be done by using a non-cartesian network with processors allocated around the original contour.

3D Skeletons

Curved Inertia Frames as presented here computes skeletons in 2 dimensional images. The network can be extended to finding 3 dimensional skeletons [Brady 1983], [Nackman and Pizer 1985] from 3 dimensional data since the local estimates for orientation and curvature can be found in a similar way and the network extends to 3 dimensions - this, of course, at the cost of increasing the number of processors. The problem of finding 3D skeletons from 2D images is more complex; however, in most cases the projection of the 3D skeleton can be found by working on the 2D projection of the shape, especially for elongated objects (at least for the shapes shown in [Snodgrass and Vanderwart 1980]).

Edge detection and early vision

Recently, filter-based approaches to early vision have been presented. These include texture [Knuttson and Granlund 1983], [Turner 1986], [Fogel and Sagi 1989], [Malik and Perona 1989], [Bovik, Clark and Geisler 1990], stereo [Kass 1983], [Jones and Malik 1990], brightness edge detection [Canny 1986], [Morrone, Owens and Burr 1987, 1990], [Freeman and Adelson 1990], and motion [Heeger 1988]. (See also [Abramatic and Faugeras 1982], [Marrone and Owens 1987]). In most of these schemes, discontinuities are defined as maxima on the filter output. Such maxima can be seen as ridges on the filter output.

Thus, Curved Inertia Frames may be used to compute discontinuities in different early vision modules such as edge detection, stereo, motion, and texture using suitable filters to estimate inertia values and tolerated length. Curved Inertia Frames could also be used to look for regions, not discontinuities, by extending the vector color ridge-detector to work on other types of early vision information.

Other Applications Of Reference Frames: Attention, Feature and Corner Detection, Part Segmentation, Spatial Reasoning, and Shape Description

The use of reference frames, and therefore that of Curved Inertia Frames, need not be restricted to recognition nor to the specific types of objects covered here. Figure 5.11 illustrates how frame curves may be applied to rigid objects and handwritten character recognition (see [Edelman 1988] for an example of how alignment may be used in OCR). Skeletons may be used in several of the approaches developped for rigid objects [Ullman 1986], [Grimson and Lozano-Pérez 1988], [Huttenlocher 1988], [Cass 1992], [Breuel 1992], [Wells 1993]. Non-recognition examples where C.I.F. may be useful include: finding an exit path in the maze of Figure 5.9, finding the corner in Figure 5.10, finding features for handwritten recognition in Figure 5.12, finding skewed symmetries [Friedberg 1986], finding the blob in Figure 5.13, determining figure-ground relations in Figure 2.10 and finding the most interesting object in Figure 2.20. In these applications, the main advantage of our scheme over previously presented grouping schemes [Marroquin 1976], [Witkin and Tenenbaum 1983], [Mahoney 1985], [Haralick and Shapiro 1985], [Lowe 1984, 1987], [Sha'ashua and Ullman 1988], [Jacobs 1989], [Grimson 1990], [Subirana-Vilanova 1990], [Clemens 1991] is that it can find complete global, curved, symmetric, and large structures directly on the image without requiring features like straight segments or corners. In this context, perceptual organization is related to part segmentation [Hollerbach 1975], [Marr 1977], [Duda and Hart 1973], [Binford 1981], [Hoffman and Richards 1984], [Vaina and Zlateva 1990], [Badler and Bajcsy 1978], [Binford 1971], [Brooks, Russel, and Binford 1979], [Brooks 1981], [Biederman 1985], [Marr and Nishihara 1978], [Marr 1982], [Guzman 1969], [Pentland 1988] and [Waltz 1975]. In part segmentation one is interested in finding an arrangement of structures in the image, not just on finding them. This complicates the problem because one skeleton alone is not sufficient (See section 2.8.3).

Parameter estimation and shape description

Curved Inertia Frames has several parameters to tinker with. Most notably the penetration constant (controlling the curvature) and the symmetry constant (controlling the deviation from symmetry that is allowed). These parameters are not hard to set when looking for a long and elongated axis. However, there are some instances in which different parameters may be needed, as illustrated in Figure 5.14. The role of context in shape description and perceptual organization deserves more work.

The Skeleton Sketch

The Skeleton Sketch suggests a way in which interest points can be computed bottom-up. These points may be useful as anchor structures for aligning model to object. The Skeleton Sketch also provides a continuous measure useful in determining the distance from the center of the object, suggesting a number of experiments. For example, one could test whether the time to learn/recognize an object depends on the fixation point. The relation may be similar to the way in which a dependence has been found in human perception between object orientation and recognition time/accuracy (see references in the Appendices). This could be done on a set of similar objects of the type shown in Figure A.3.

Random networks

More work is necessary to clarify the type of non-cartesian network that is best suited to Curved Inertia Frames. This requires more theoretical progress, borrowing from the fields of random algorithms [Rahavan 1990] and geometric probability [Solomon 1978]. Random networks could find use in other existing algorithms such as the dynamic programming aproach to shape from shading presented in [Langer and Zucker 1992].

Other mid-level vision tasks

In this thesis we have concentrated our efforts into networks that perform perceptual organization by computing frames that go through high, long, symmetric, and smooth regions of the shape. Other mid-level tasks may be performed with similar computations. For example, that of finding a description around a focus point or that of computing transparent surfaces. These networks may complement Curved Inertia

Frames, for example by incorporating the notion of small structures as described next.

Small structures

The scheme presented in this thesis has a bias for large structures. This is generally a good rule, except in some cases, see Figures A.2, A.1 and A.4. The example of Figure A.1 provides evidence that the preference for small objects can not be due only to pop-out effects. This distinction had not been made before and deserves further treatment.

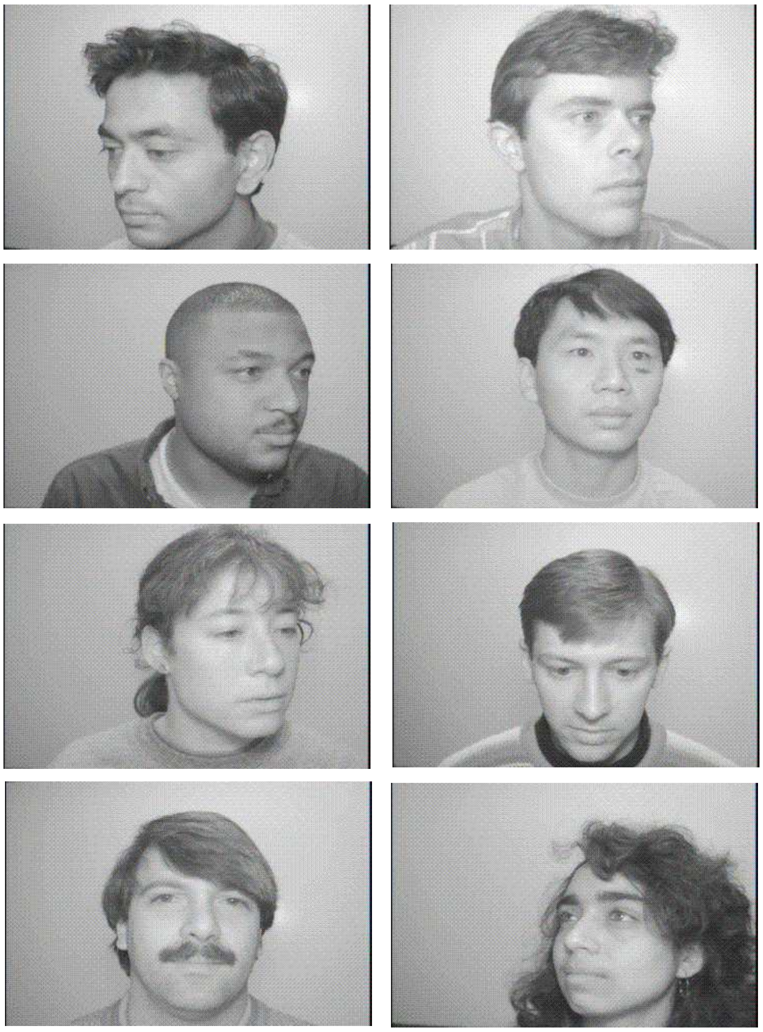


Figure 5.4: Contour texture could be used as a learnable feature for for recognition. The outputs of the filters in the frame curve could be learned using a connectionist approach. This could be used to index into a face database.

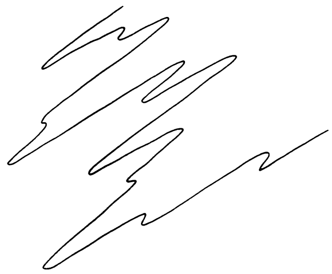


Figure 5.5: Contour texture can provide three-dimensional information [Stevens 1980]. This could be recovered by finding the filter outputs that respond to a given contour.

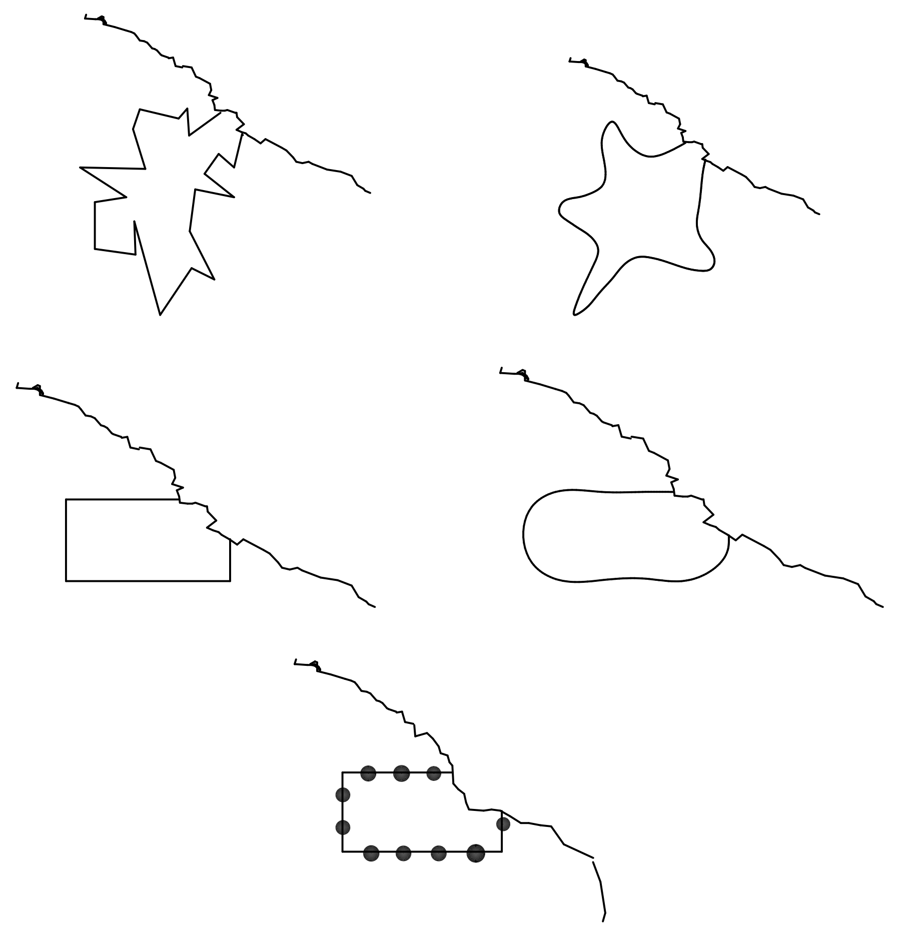


Figure 5.6: Evidence that contour texture is used when completing occluded contours.

Figure 5.7: Evidence that discontinuities in contour texture are easy to detect.

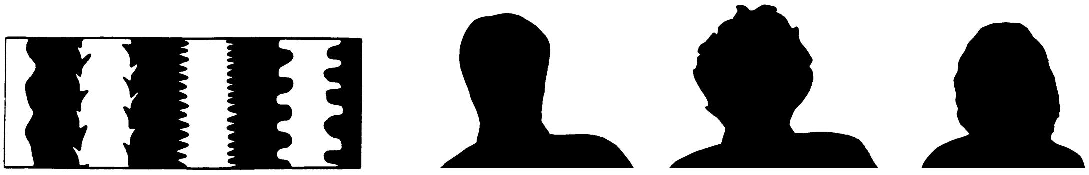


Figure 5.8: This Figure illustrates some of the applications of contour texture. *Left:* Contours can be grouped based on their contour texture. *Right:* The contour texture can be used as a powerful indexing measure into large databases of objects.

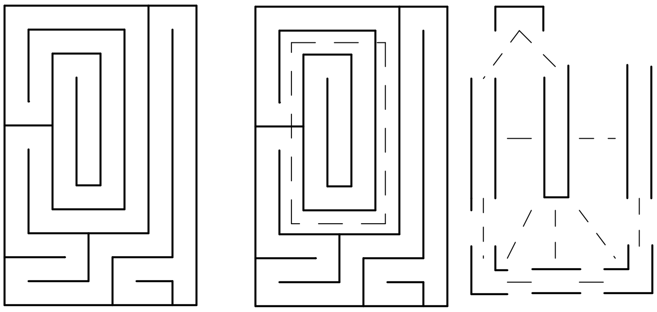


Figure 5.9: *Left:* Maze. *Center:* A portion of the maze is highlighted with a discontinuous line. *Right:* A candidate description of the highlighted region that may be useful to find an exit from the maze. In this Chapter, we are interested in finding frames or skeletons that can yield high-level descriptors like the one shown.

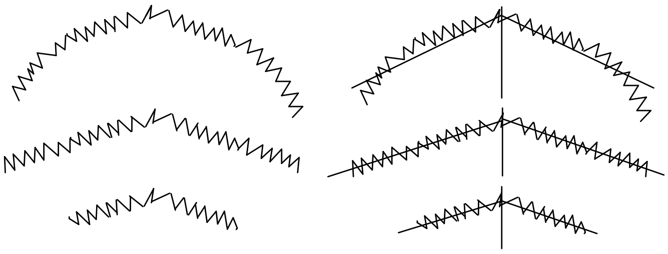


Figure 5.10: Finding corners is hard because they depend on scale. Here we present compelling evidence adapted from [Lowe 1988]. A suggestion of this thisis is that the scheme presented in this Chapter can locate corners of this type, independently of scale, because it looks for the largest possible scale. The scheme would look for position along the contour where there are local maximum in the skeleton sketch.

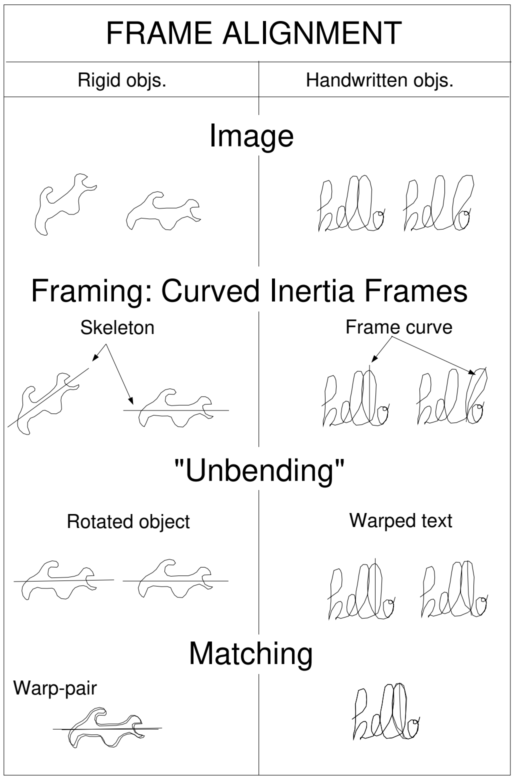


Figure 5.11: *Left:* Frame alignment applied to rigid objects. The "bending" stage is reduced to a rotation. *Right:* Frame alignment applied to handwritten objects. One frame is not sufficient to "unbend" the whole object. Separate curves for each letter need to be computed.

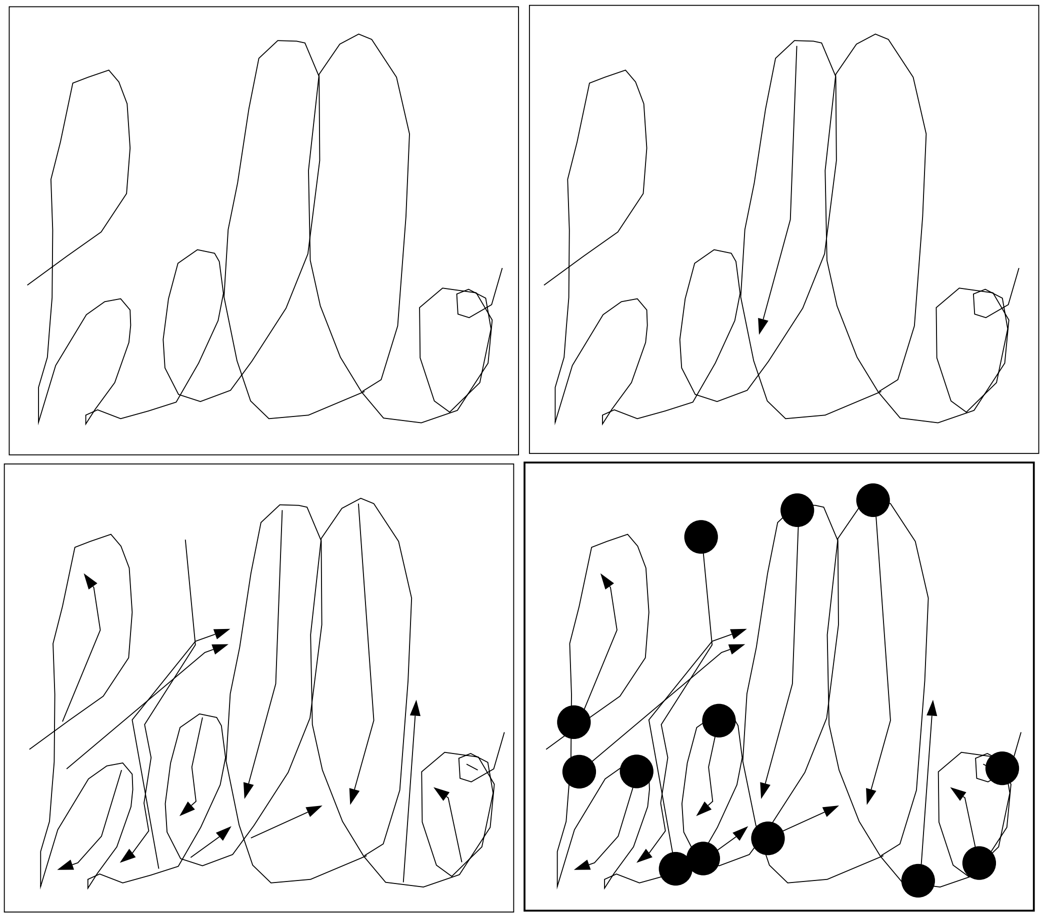


Figure 5.12: *Top left:* Handritten word. *Top right:* Highest inertia skeleton. *Bottom left:* Set of Curved Inertia Frames found by random network. The arrow indicates the "end" point of the skeleton. *Bottom right:* Features along the contour have been found by locating local maxima in the Skeleton Sketch.

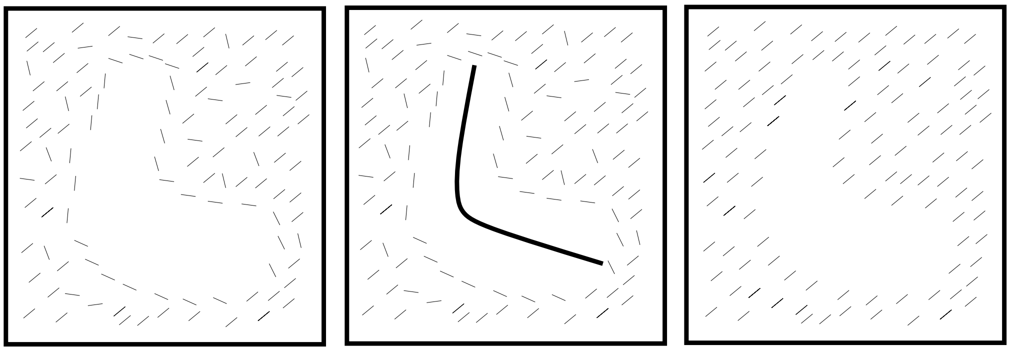


Figure 5.13: Finding the bent blob in the *left* image would be easy if we had the bent frame shown in the *center*. *Right:* Another blob defined by orientation elements of a single orientation. The scheme presented in this thesis needs some modifications before it can attempt to segment the blob on the right (see text).

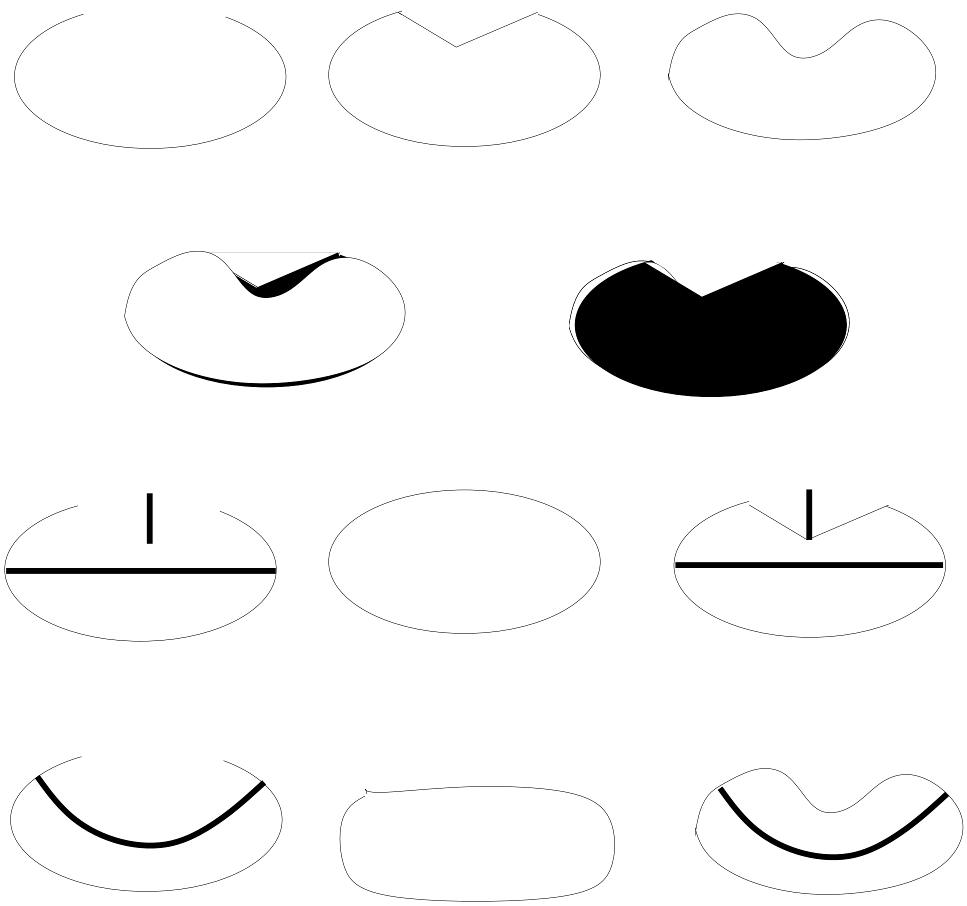


Figure 5.14: This figure reinforces the notion that small changes may result in remarkably different perceptions. *Top row:* One unfinished shape and two ways of completing the shape. *Second row:* the two completed shapes are almost identical as it is shown by overimposing the two shapes. *Third and fourth rows:* Two possible descriptions for the two shapes. This figure provides evidence that local cues such as curvature are important to determine what is the appropriate skeleton description for a shape. Curved Inertia Frames can be adapted to different task requirements by adjusting the penetration and symmetry constants. With the parameters used in the experiments, the output would be that of the bent axis. However, C.I.F. finds a straight axis if the penetration constant is reduced to 0.2 (instead of 0.5). Example adapted from [Richards 1989].

Curve Inertia Frames and Human Perception

Appendix A

A.1 Introduction

For a given shape, the skeleton found by Curved Inertia Frames, as described in Chapters 2 and 3, corresponds roughly to the central regions of the shape. In this Appendix we show how C.I.F. can handle several peculiarities of human perception: bias in frame of reference and occlusion (Sections A.2 and A.3), size perception (Section A.4), perception of discontinuities (Section A.5), and frames of reference in recognition (Section A.6). Our analysis will lead us to present several open problems and new perceptual phenomena. Appendix B will present evidence in favor of the use of frame curves in human perception.

A.2 Frames of Reference

Important frames of reference in the perception of shape and spatial relations by humans include: that of the perceived object, that of the perceiver, and that of the environment. So far in this thesis, we have concentrated on the first. A considerable amount of effort has been devoted to study the effects of the orientation of such a frame (relevant results include, to name but a few [Attneave 1967], [Shepard and Metzler 1971], [Rock 1973], [Cooper 1976], [Wiser 1980, 1981], [Schwartz 1981], [Shepard and Cooper 1982], [Jolicoeur and Landau 1984], [Jolicoeur 1985], [Palmer 1985], [Palmer and Hurwitz 1985], [Corballis and Cullen 1986], [Maki 1986], [Jolicoeur, Snow and Murray 1987], [Parsons and Shimojo 1987], [Robertson, Palmer and Gomez 1987], [Rock and DiVita 1987], [Bethel-Fox and Shepard 1988] [Shepard and Metzler 1988], [Corballis 1988], [Palmer, Simone, and Kube 1988], [Georgopoulos, Lurito, Petrides, Schwartz, and Massey 1989], [Tarr and Pinker 1989]). C.I.F. suggests a computational model of how such an orientation may be computed, the orientation is that of the most salient skeleton assuming it is restricted to be straight (α and ρ close to 0).

The influence of the environment on the frame has been extensively studied, too [Mach 1914], [Attneave 1968], [Palmer 1980], [Palmer and Bucher 1981], [Humphreys 1983], [Palmer 1989]. In some cases the perception of the shape can be biased by the frame of the environment. In particular, humans have a bias for the vertical in shape description (see [Rock 1973]) so that some shapes are perceived differently depending on the orientation at which they are viewed. For example, a rotated square is perceived as a diamond (see Figure A.5). This bias can be taken into account in C.I.F. by adding some constant value to the inertia surface that corresponds to the vertical orientation so that vertical curves receive a higher inertia value. Adding the bias towards the vertical is also useful because it can handle non-elongated objects that are not symmetric, so that the preferred frame is a vertical axis going through the center of the shape[1].

In other cases, the preferred frame is defined by the combination of several otherwise non salient frames. This is the case in Mach's demonstration, first described by E. Mach at the beginning of this century (see Figure 2.18). C.I.F. incorporates this behavior because the best curve can be allowed to extend beyond one object increasing the inertia of one axis by the presence of objects nearby, especially when the objects have high inertia aligned axes. This example also illustrates the tolerance of the scheme to fragmented shapes.

The shape of the frame has received little attention. In Chapter 2, we proposed that in frame alignment a curved frame might be useful (see also Figure 2.4 and [Palmer 1989]). In particular, we have proposed to recognize elongated curved objects by unbending them using their main curved axis as a frame to match the unbent versions. In Appendix B it is shown that such a strategy is not always used in human perception.)

In figure-ground segregation, reference frame computation, and perceptual organization it is well known that humans prefer symmetric regions over those that are not (see Figures 2.10, A.6, and the references above[2]). Symmetric regions can be discerned in our scheme by looking for the points in the image with higher skeleton inertia values. However, [Kanisza and Gerbino 1976] have shown that in some cases convexity may override symmetry (see Figure 2.10).

[1] As discussed in section 2.5, another alternative is to define a specific computation to handle the portions of the shapes that are circular [Fleck 1986], [Brady and Scott 1988].

[2] The role of symmetry has been studied also for random dot displays [Barlow and Reeves 1979], [Barlow 1982] and occlusion [Rock 1984].

Convexity information can be introduced in the inertia surfaces by looking at the distances to the shape, and at the convexity at these points. This information can be used so that frames inside a convex region receive a higher inertia value. Observe that the relevant scale of the convexity at each point can be determined by the distances to the shape (R and r as defined in Chapter 2).

The location of the frame of reference [Richards and Kaufman 1969], [Kaufman and Richards 1969], [Carpenter and Just 1978], [Cavanagh 1978, 1985], [Palmer 1983], [Nazir and O'Reagan 1990] is related to attention and eye movements [Yarbus 1967], and influences figure-ground relations (e.g. Figure D9 in [Shepard 1990]).

We have shown how certain salient structures and individual points can be selected in the image using the Skeleton Sketch; subsequent processing stages can be applied selectively to the selected structures, endowing the system with a capacity similar to the use of selective attention in human vision. The points provided by the Skeleton Sketch are in locations central to some structures of the image and could guide processing in a way similar to the direction of gaze in humans (e.g. [Yarbus 1967]).

[Palmer 1983] studied the influence of symmetry on figural goodness. He computed a "mean goodness rating" associated to each point inside a figure. For a square (see Figure 4 in [Palmer 1983]), he found a distribution similar to that of the skeleton sketch shown in Figure 2.17. The role of this measure is unclear but our scheme suggests that it can be computed bottom-up and hence play a role prior to the recognition of the shape.

Perhaps this measure is involved in providing translation invariance so that objects are first transformed into a canonical position. This suggestion is similar to others that attempt to explain rotation invariance (see references in Appendix B) and it could be tested in a similar way. For example, one can compute the time to learn/recognize an object (from a class sharing a similar property such as the one shown in Figure A.3) in terms of a given displacement in fixation point[3] (or orientation in the references above).

[3]Note that most computational schemes differe from this model and assume that translation invariance is confounded with recognition [Fukushima 1980], [LeCun, Boser, Benker, Henderson, Howard, Hubbard and Jackel 1989], [Földiák 1991].

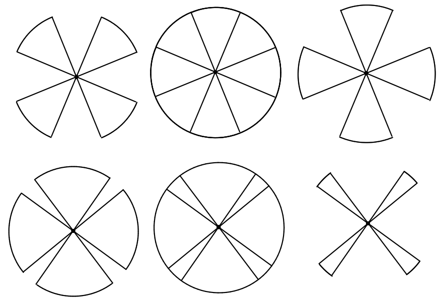


Figure A.1: *Top Center:* Figure is often seen as shown on the *right*, (and ground as *on the left*) due to vertical bias. *Bottom Center:* Preference for the vertical, and preference for large objects is over-ridden here by the preference for small structures (after [Rock 1985]). The network presented in this thesis would find the *left* object as figure due to its preference for large structures. Further research is necessary to clarify when small structures are more salient.

A.3 What Occludes What?

C.I.F. solves the problem of finding different overlapping regions by looking at the large structures one by one. In C.I.F., the larger structures are the first ones to be recovered. This breaks small structures covered by larger structures into different parts. As a result, C.I.F. embodies the constraint that larger structures tend to be perceived as occluding surfaces [Petter 1956]. (See also Figure A.7).

A blob that is salient in one image might not be so when other elements are introduced. An example due to J. Lettvin that supports this claim is shown in figure A.4. We contend that, using our terms, the largest scale is automatically selected by the human visual system. In other words, when there are "overlapping" interpretations the human visual system picks "the largest".

A.4 Small Is Beautiful Too

As mentioned in Chapter 2, the emphasis of C.I.F. is towards finding large structures. In some cases, this may be misleading as evidenced by Figures A.2 and A.1. In these

Figure A.2: Drawing from Miró. As in the previous Figure, small structures define the object depicted in this image. This image would confuse the network presented in [Sha'ashua and Ullman 1988].

examples the interesting structure is not composed of individual elements that pop-out in the background. Instead, what seems to capture our attention can be described as "what is not large". That is, looking for the large structures and finding what is left would recover the interesting structure as if we where getting rid of the background. It is unclear, though, if this observation would hold in general and further research is necessary.

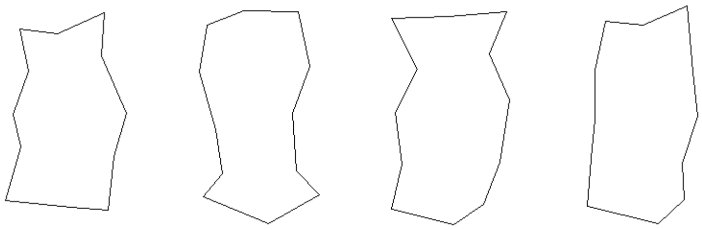


Figure A.3: Does the time to recognize/learn these objects depend on the fixation point? (See text for details.)

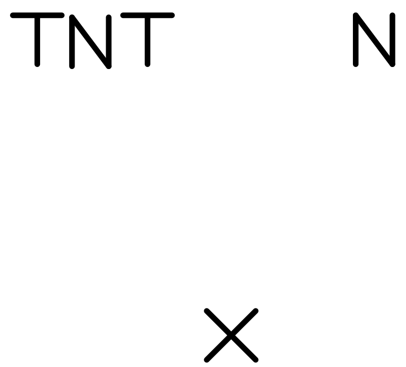


Figure A.4: This Figure provides evidence that a salient blob in one image might not be so when other elements are introduced. By fixating at the **X** try to identify the letter N in the left and in the right of the image. The one on the left is not identifiable. We contend that this is due to the fact that the human visual system selects the larger scale in the left case yielding an horizontal blob. This example is due to J. Lettvin (reference taken from [Ullman 1984]).

A.5 Are Edges Necessary?

A central point in Chapter 3 is that the computation of discontinuities should not precede perceptual organization. Further evidence for the importance of perceptual organization is provided by an astonishing result obtained by [Cumming, Hurlbert, Johnson, and Parker 1991]: when a textured cycle of a sine wave in depth (the upper half convex, the lower half concave) is seen rotating, both halves may appear convex[4], despite the fact that this challenges rigidity[5] (in fact, a narrow band between the two ribbons is seen as moving non-rigidly!). This, at first, seems to violate the rigidity assumption. However, these results provide evidence that before finding the structure from motion, the human visual system may segment the image into

[4] The surface can be described by the equation $Z = sin(y)$ where Z is the depth from the fixation plane. The rotation is along the Y-axis by $+/-$ 10 degrees at 1 Hz.

[5] This observation is relevant because it supports the notion that perceptual organization is computed in the image before structure from motion is recovered.

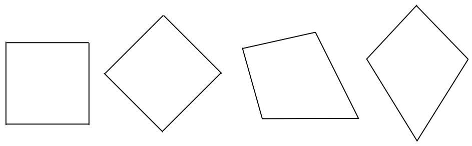


Figure A.5: A square has four symmetry axes all of which could potentially be used to describe it. Depending which one of them is chosen, this shape appears as a square or as a diamond. This suggests that when there is ambiguity the vertical can play an important role. The two trapezoids, on the right further illustrate that even when a shape has several symmetry axis the vertical might be preferred even if it does not correspond to a perfect symmetry axis. Observe that the vertical might be overridden by an exterior frame which can be defined by the combination of several otherwise not salient frames from different shapes such as Mach demonstration, see Figure 2.18.

different components. Within each of these, rigidity can prevail.

Evidence against any form of grouping prior to stereo is provided by the fact that we can understand random dot stereo diagrams (R.D.S.) even though there is no evidence at all for perceptual groups in one single image. However, it is unclear from current psychological data if these displays take longer time. If they do, one possible explanation (which is consistent with our suggestions) may be that they impair perceptual organization on the individual images and therefore on stereo computations. We believe that the effect of such demonstrations has been to focus the attention on stereo without grouping. But perhaps grouping is central to stereo and R.D.S. are just an example of the stability of our stereo system (and its stereo grouping component!).

A second central point of Chapter 3 is that edge detection may not precede perceptual organization. However, there are a number of situations in which edges are clearly necessary as when you have a line drawing image[6] or for the Kanizsa figures. Nevertheless some sort of region processing must be involved since surfaces are also perceived. We (like others) believe that region-based representations should be sought even in this case. In fact, as we noted in section 2, line drawings are harder

[6]Although note that each line has 2 edges (not just one), generally it is assumed that when we look at such drawings we ignore one of the edges. An alternative possibility is that our visual system assembles a region-based description from the edges without merging them.

to recognize (just like R.D.S. seem to be - but see [Biederman 1988]). The role of discontinuities versus that of regions is still unclear.

A.6 Against Frame Alignment; Or Not?; Or What?

Elongated objects, like an I, when bent become like a C or an O developing a hole in the process (see Figure 2.6). In other words, there is a transformation that relates elongated objects to objects with a hole.

The notion that C's (elongated flexible objects) and O's (holes) can both be seen as bent I's suggests that there may be a similar algorithm to recognize both types of objects which uses this fact. As suggested in Chapter 2, elongated flexible objects can be recognized in some cases using frame alignment by transforming the image to a canonical version of itself, in which the object has been unbent (see Figure 2.4). With this scheme, the skeleton of the shape is used as an anchor structure for the alignment process. Can this scheme be extended to handle objects with holes? Does human perception use such a scheme?

One property that distinguishes the two (I and O) is that the inside of a hole can be perceived as a stable entity (present in different instances of the shape) while in the bent I the cavity is unstable due to changes depending on the degree of bending. On the other hand, the hole of a bent I occurs at a predictable distance from the outside of the shape while in more symbolic "hole-like" descriptions the location is more unpredictable.

Notions of inside/outside are also key in part-like descriptions since we are bound to determine what is the inside of a part before we can compute its extent. Note that the definition of part given by [Hoffman and Richards 1984] depends on the side perceived as inside. This brings us to the issue of non-rigid boundaries: How is inside/outside determined? What extent does a boundary have? If holes are independent, how does processing proceed? In Appendix B we present evidence in favor of a scheme in which visual processing proceeds by the successive processing of convex chunks (or "hole chunks"!). This lead us to analyze the process by which images are described. We will suggest that a part description can be computed using the outside rule (defined in Section B.3) which defines them as outside chunks of convex chunks. We will also describe the implications of our findings for existing recognition models.

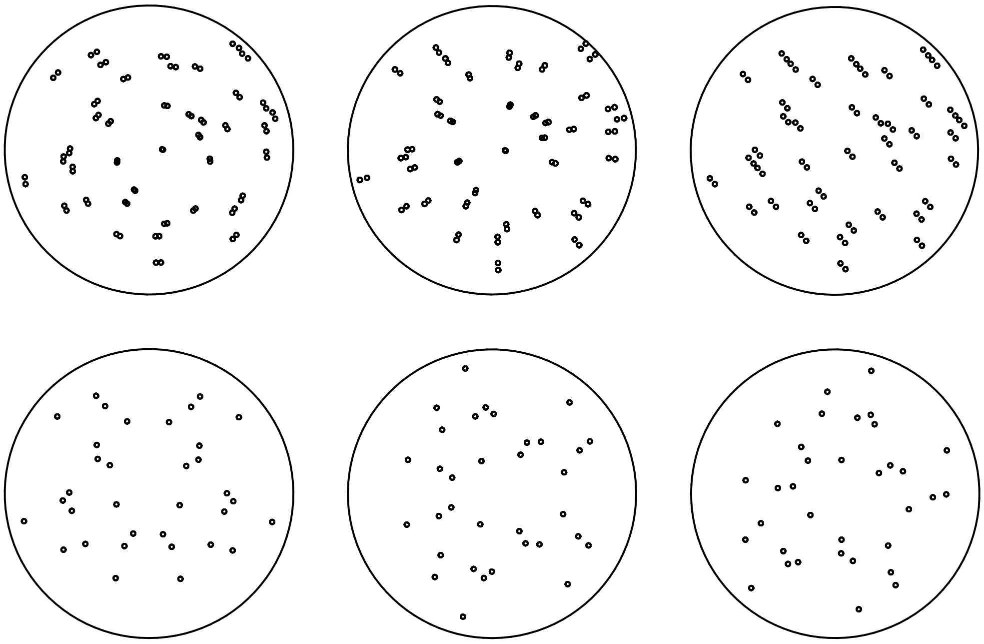


Figure A.6: The patterns in this Figure have been made by superimposing an image made of random dots with a transformation of itself. Transformations shown include rotation (top left), expansion (top middle), translation (top right), symmetry (bottom left). The two other images (bottom middle and right) are a rotation of the bottom left image. Other random dot transformations include stereo and motion. Surprisingly enough, we can detect most of these transformations. However, many studies suggest that our detection of symmetry in these cases is limited - being restricted to the vertical and to low frequency information. We contend that humans do not have symmetry detectors in the same way that it has been proposed that we have orientation detectors. Instead, we propose that symmetry is detected at a later stage and based in salient structures.

Figure A.7: Large shapes occlude small ones. From [Kanizsa 1979].

Frame Curves and Non-Rigid Boundaries

Appendix B

Does the human visual system compute frame curves or skeletons? If so, are they computed prior to recognition? Does the human visual system use frame curves to unbend objects before recognizing them? In this Appendix[1] we will present some suggestions that attempt to answer these issues. Since they are coupled to some fundamental problems of visual perception such as perceptual organization, attention, reference frames, and recognition, it will be necessary to address these, too. The suggestions presented in the Appendix are based on some simple observations. The essence of them can be easily grasped by glancing at the accompanying figures. The text alternates the presentation of such observations with the discussion of the suggested implications. We begin by challenging the notion that objects have well defined boundaries.

B.1 Introduction

The natural world is usually conceived as being composed of different objects such as chairs, dogs, or trees. This conception carries with it a notion that objects occupy a region of space, and have an "inside". By default, things outside this region of space are considered "outside" the object. Thus, the lungs of a dog are inside the dog, but the chair occupies a different region and is outside the object dog. When an object

[1]This Appendix is adapted from [Subirana-Vilanova and Richards 1991].

is projected into an image, these simple notions lead to what appears to be a clear disjunction between what is considered figure, and what is ground. Customarily, the figure is seen as the inside of the imaged shape as defined by its bounding contours (i.e. its silhouette). The region outside this boundary is ground. This implies that the points of an image are either figure or ground. Such a view is reinforced by reversible figures, such as Rubin's vase-face or Escher's patterns of birds and fish. This view carries the notion that, at any instant, the attended region has a well defined boundary.

Here, we show that such a simple disjunctive notion of attention and its reference frame is incorrect, and in particular, we show that the assignment of an attentional boundary to a region of an image is ill-posed. If the region of attention has an ill-defined boundary, presumably there are some regions of the image that are receiving "more attention" than others. We present observations that support this conclusions and show that the result is due not only to processing constraints but also to some computational needs of the perceiver. In particular, a fuzzy boundary leaves room to address regions of the image that are of more immediate concern, such as the handle of a mug (its outside) that we are trying to grasp, or the inside surface of a hole that we are trying to penetrate.

The ambiguity in defining a precise region of the image as the subject of attention arises in part because many objects in the world do not have clearly defined boundaries. Although objects occupy a region of space, the inside and outside regions of this space are uncertain. For example, what is the inside of a fir tree? Does it include the region between the branches where birds might nest, or the air space between the needles? If we attempt to be quite literal, then perhaps only the solid parts define the tree's exterior. But clearly such a definition is not consistent with our conceptual view of the fir tree which includes roughly everything within its convex hull. Just like the simple donut, we really have at least two and perhaps more conceptualizations of inside and outside. For the donut, the hole is inside it, in one sense, whereas the dough is inside it in another. But the region occupied by the donut for the most part includes both. Similarly for the fir tree, or for the air space of the mouth of a dog when it barks. Which of these two quite distinct inclusions of inside should be associated with the notion of object? Or, more properly, what is the shape of the atentional region and its reference frame in this case?

We begin, in the next Section, by presenting some demonstrations that clarify how figural assignments are given to image regions. Along the way, we use these demonstrations to suggest an operational definition of the attentional reference frame. In the following two sections we suggest that outside is more salient than inside; or not; or what? In section B.5 we review the notion of "hole" and in Sections B.6 and B.7 that of attentional reference frame. In Sections B.8 and B.9 we discuss the implications of our findings to visual perception. In Section B.11, we suggest that typically "near is more salient than far" and point to other similar biases. We end in Section B.12 with a summary of the new findings presented.

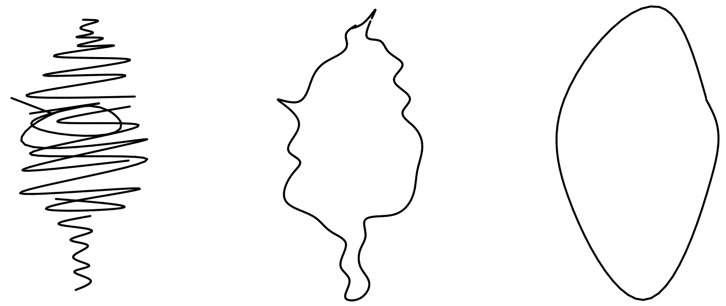


Figure B.1: Fir tree at several scales of resolution. What is inside the tree?

B.2 Fuzzy Boundaries

Typically, figure-ground assignments are disjunctive, as in the Escher-drawings. However, when the image of a fractal-like object is considered, the exact boundary of the image shape is unclear, and depends upon the scale used to analyze the image. For the finest scale, perhaps the finest details are explicit, such as the needles of a spruce or the small holes through which a visual ray can pass unobstructed. But at the coarsest scale, most fractal objects including trees will appear as a smooth, solid convex shape. Any definition of the region of attention and its frame must address this scale issue. Consider then, the following definitions:

Definition 1 (Region of Attention): *The region of attention is that collection of structures (not necessarily image-based) which currently are supporting the analysis of a scene.*

Definition 2 (Attentional Frame): *The attentional frame, for a region of attention, is a coordinate frame within which the structures can be organized.*

By these definitions, we mean to imply that the perceiver is trying to build or recover the description of an object (or scene) in the world, and his information-processing capability is focused on certain regions in the scene that are directly relevant to this task.

The precise regions of the scene that are being analyzed, and their level of detail will be set by the demands of the goal in mind. Such a definition implies that the regions of the scene assigned as attentional frame may not have a well-defined, visible contour. Indeed, by our definition these regions do not have to be spatially coherent!

It has long been known that humans concentrate the processing of images in certain regions or structures of the visual array (e.g. [Voorhis and Hillyard 1977]). Attention has several forms: one of them, perhaps the most obvious, is gaze. We can not explore a stationary scene by swinging our eyes past it in continuous movements. Instead, the eyes jump with a saccadic movement, come to rest momentarily and then jump to a new locus of interest (see [Yarbus 1967]).

These observations suggest the following:

Claim 1 *The region of the image currently under directed attention may not have a well-defined boundary earmarked by a visible image contour.*

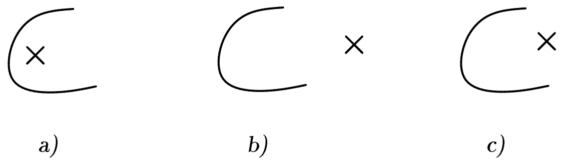


Figure B.2: The notion of "what is the attentional frame" does not require that the figure be a region enclosed by a visible contour. In (a) the x is seen to lie within the C, and is associated with the figure, whereas in (b) the x lies outside the figure. In (c) the answer is unclear.

In support of this claim consider the C of Figure B.2. Although the image contour by itself is well-defined, the region enclosed by the C is not. The region "enclosed" by the "C" is a legitimate processing chunk. For example, if one asks the question does the "X" lie inside the C, our immediate answer is yes for case (a), and no for case (b). To make this judgement, the visual system must evaluate the size of the interior region of the C. Thus, by our definition, the concept "inside of C" must lead to an assignment of certain pixels of the display as the region of attention. Without an explicit contour in the image, however, where should one draw the boundary between the region under attention? For example, should we choose to close the attentional region with a straight line between the two endpoints? Another possibility would be to find a spline that completes the curve in such a way that the tangent at the two endpoints of the C is continuous for the complete figure. These findings agree with a model in which the boundary is something more closely approaching a "blurred" version of the C, as if a large Gaussian mask were imposed on a colored closed C. We contend that such "fuzzy" attentional boundaries occur not only within regions that are incompletely specified, such as that within the incomplete closing of the C, but also within regions that appear more properly defined by explicit image contours.

To further clarify our definition of the attentional region and its frame, note that it is not prescribed by the retinal image, but rather by the collection of image structures in view. Any pixel-based definition tied exclusively to the retinal image is inadequate, for it will not allow attentionnal (and processing) assertions to be made by a sequence of fixations of the object. Rather, a structure-based definition of attentional frame presumes that the observer is building a description of an object or event, perhaps by recovering object properties. The support required to build these object properties is what we define as the attentional region. This support corresponds closely to Ullman's incremental representations [Ullman 1984] upon which visual routines may act, and consequently the operations involved in attentional assertions should include such procedures as indexing the sub-regions, marking these regions, and the setting of a coordinate frame. We continue with some simple observations that bear on these problems.

B.3 Outside is More Salient than Inside

When binary attentional assignments are made for an image shape with a well-defined, simple, closed contour, such as an "O", the assignment is equivalent to partitioning the image into two regions, one lying inside the contour, the other outside. For such a simple shape as the "O", the immediate intuition is that it is the inside of the contour which is given the attentional assignment, and this does not include any of the outside of the contour (see [Hoffman and Richards 1984] for example, where shape descriptors depend on such a distinction). By our definition, however, the attentional region might also include at the very least a small band or ribbon outside the contour, simply because contour analysis demands such. As a step toward testing this notion, namely that a ribbon along the outer boundary of the shape should also be included when attentional assignments are made, we perturb the contour to create simple textures such a those illustrated in Figure B.3 (bottom row).

In this Figure, let the middle star-pattern be your reference. Given this reference pattern, which of the two adjacent patterns is the most similar? We find that the pattern on the left is the most similar[2]. Now look more closely at these two adjacent patterns. In the left pattern, the intrusions have been smoothed, whereas in the right pattern the protrusions are smooth. Clearly the similarity judgement is based upon the similarity of the protrusions, which are viewed as sharp convex angles. The inner discrepancy is almost neglected.

The same conclusion is reached even when the contour has a more part-based flavor [Hoffman and Richards 1984], rather than being a contour texture, as in Figure B.4. Here, a rectangle has been modified to have only two protrusions [3]. Again, subjects will base their similarity judgments on the shape of the convex portion of the protrusion, rather than the inner concavity.

This result is not surprising if shape recognition is to make any use of the fact that

[2]The results are virtually independent of the viewing conditions. However, if the stars sustain an angle larger than 10 degrees, the preferences may reverse. A detailed experiment has not been made and the observations of the reader will be relied upon to carry our arguments

[3]A collection of similar shapes could be used in a formal experiment.

most objects in nature can be decomposed into parts. The use of such a property should indeed place more emphasis upon the outer portions of the object silhouette, because it is here that the character of a part is generally determined, not by the nature of its attachment. Almost all attachments lead to concavities, such as when a stick is thrust into a marshmallow. Trying to classify a three-dimensional object by its attachments is usually misguided, not only because many different parts can have similar attachments, but also because the precise form of the attachments is not reliably visible in the image. Hence the indexing of parts (or textures) for shape recognition can proceed more effectively by concentrating on the outer extremities.

Another possible justification for such observations is that, in tasks such as grasping or collision avoidance, the outer part is also more important and deserves more attention because it is the one that we are likely to encounter first[4].

The outer region of a shape is thus more salient than its inner region. This implies that the region of scene pixels assigned to the attentional region places more weight on the outer, convex portions of the contour than on its interior concave elements (or to interior homogeneous regions), and leads to the following claim:

Claim 2 *The human visual system assigns a non-binary attentional function to scene pixels with greater weight given to regions near the outside of shapes, which become more salient.*

Note that this claim simply refers to "regions near the outside", not to whether the region is convex or concave. In Figure B.4, the outer portion of the protrusion contains a small concavity, which presumably is the basis for the figural comparison.

Exactly what region of the contour is involved in this judgement is unclear, and may depend upon the property being assessed. All we wish to claim at this point is that whatever this property, its principal region of support is the outer portion of the contour. The process of specifying just which image elements constitute this outer contour is still not clear, nor is the measure (nor weight) to be applied to these elements. One possibility is an *insideness* measure. Such a measure could be easily

[4]For example, in Figure B.3 (bottom), the center and left stars would "hurt" when grasped, whereas the right star would not because it has a "smooth" outside.

computed as a function of the distance of the image elements to the "smoothed" version of the contour (a circle in Figure B.3 and something close to a rectangle in Figure B.4). In this context, the smoothed contour corresponds to the notion of frame curves as used in Chapter 4.

This leads us to the following definition of frame curve which has to be read bearing in mind claim 1:

Definition 3 (Frame Curve): *A frame curve is a virtual curve in the image which lies along "the center" of the attentional region's boundary.*

In general, the frame curve can be computed by smoothing the silhouette of the shape. This is not always a well-defined process because the silhouette may be ill-defined or fragmented, and because there is no known way of determining a unique scale at which to apply the smoothing. Figure B.5 (center) shows a frame curve for an alligator computed using such scheme. On the right, the regions of the shape that are "outside" the frame curve have been colored; note that these regions do not intersect, and correspond closely to the outer portions of the different parts of the shape. As mentioned above, these outer portions are both more stable and more likely to be of immediate interest.

Our interpretation of the bias towards the inside suggests, implicitly, a part perceptual organization scheme using the boundary of the shape as has been just discussed. The frame curve can be used to compute a part as follows:

Definition 4 (Outside Rule): *A part can be approximated as the portion of the boundary of a shape (and the region enclosed by it) which lies outside (or inside) the frame curve. Such portion of the boundary should be such that there is no larger portion of the boundary which contains it.*

Note that Claim 2 supports Claim 1 because the attentional function mentioned in Claim 2 is to be taken to represent a fuzzy boundary for the attentional region. The frame curve should not be seen as a discrete boundary for this attentional region (perhaps only at a first approximation). Indeed, we contend that a discrete boundary is not a realistic concept.

Furthermore, the outside rule should not be seen as a definition of part (see [Hoffman and Richards 1984] for a detailed definition) instead, it is just a way in which parts can be computed. In fact, it resembles closely the way in which Curved Inertia Frames computes the different regions of the object.

B.4 Inside is More Salient than Outside

Consider once more the three-star patterns of Figure B.3. Imagine now that each of these patterns is expanded to occupy 20 degrees of visual angle (roughly your hand at 30 centimeters distance). In this case the inner protrusions may become more prominent and now the left pattern may be more similar to the middle reference pattern. (A similar effect can be obtained if one imagines trying to look through these patterns, as if in preparation for reaching an object through a hole or window.) Is this reversal of saliency simply due to a change in image size, or does the notion of a "hole" carry with it a special weighting function for attentional assignments?

For example, perhaps by viewing the central region of any of the patterns of Figure B.3 as a "hole", the specification of what is outside the contour has been reversed. Claim 2 would then continue to hold. However, now we require that pixel assignments to "the attentional region" be gated by a higher level cognitive operator which decides whether an image region should be regarded as an object "hole" or not.

B.5 When a Hole Is Not a Hole

Consider next the star patterns in Figure B.6, which consist of two superimposed convex shapes, one inside the other. Again with the middle pattern as reference, typically subjects will pick as most similar the adjacent pattern to the right. This is surprising, because these patterns are generally regarded as textured donuts, with the inner-most region a hole. But if this is the case and our previous claim is to hold, then the left pattern should have been most similar. The favored choice is thus as if the inner star pattern were viewed as one object occluding another. Indeed, if we now force ourselves to take this view, ignoring the outer pattern, then the right patterns are again more similar as in Figure B.3. So in either case, regardless of whether we view the combination as a donut with a hole, or as one shape occluding part of another, we still use the same portion of the inner contour to make our similarity judgement. The hole of the donut thus does not act like a hole. The only exception is when we explicitly try to put our hand through this donut hole. Then the inner-most protrusions become more salient as previously described for "holes".

These results lead to the following (re-visited) claim:

Claim 2 (revisited): *Once the attentional region and its frame are chosen, then (conceptually) a sign is given to radial vectors converging or diverging from the center of this frame (i.e. the focal point). If the vector is directed outward (as if an object representation is accessed), then the outer portion of the encountered contours are salient. If the vector is directed inward to the focal point (as if a passageway is explored), then the inner portion of the contour becomes salient*[5]

In the star patterns that we discussed in Section B.2 (see Figure B.3) the attention was focused primarily on the stars as whole objects. That is, there is a center of the figure that appears as the "natural" place to begin to direct our attention. The default location of this center, which is to become the center of a local coordinate frame, seems to be, roughly, the center of gravity of the figure [Richards and Kaufman 1969], [Kaufman and Richards 1969], [Palmer 1983]. Attention is then allowed to be directed to locations within this frame. Consider next the shapes shown in the top of Figure B.7. Each ribbon-like shape has one clear center on which we first focus our attention. So now let us bend each of the ribbons to create a new frame center which lies near the inner left edge of each figure (Figure B.7, lower). Whereas before the left pattern is regarded as more similar to the middle reference, now the situation

[5]There is an interesting exception to the rule: if the size of the contours is very big (in retinal terms) then the inside is always more salient (as if we where only interested in the inside of large objects).

is starting to become confused. When the ribbons are finally closed to create the donuts of Figure B.6, the favored similarity judgement is for the right pattern. The primary effect of bending and closing the ribbon seems to be a shift in the relation between the attentional frame and the contours. Following the center of gravity rule, this center eventually will move outside the original body of the ribbon. This suggests that the judgments of texture similarity are dependent on the location of the attentional coordinate frame.

Typically, as we move our gaze around the scene, the center of the coordinate frame will shift with respect to the imaged contours, thus altering the pixel assignments. A shift in the focus of attention without image movements can create an effect similar to altered gaze. More details regarding these proposed computations will be given in the next sections. What is important at the moment is that the saliency of attentional assignments will depend upon the position of the contour with respect to the location of the center of the attentional coordinate frame. The reader can test this effect himself by forcing his attention to lie either within or outside the boundary of the ribbons. Depending upon the position chosen, the similarity judgments change consistently with Claim 2.

As we move our attention around the scene, the focus of attention will shift, but the frame need not. But if the frame moves, then so consequently will the assignment of scene pixels to (potentially) active image pixels. The visual system first picks a (virtual) focal point in the scene, typically bounded by contours, and based on this focal point, defines the extent of the region (containing the focal point) to be included as the attended region. If all events in the selected region are treated as one object or a collection of superimposed objects, then the radially distant (convex) portions of the contours drive the similarity judgments and are weighted more heavily in the figural computations. On the other hand, if the choice is made to regard the focal point as a visual ray along which something must pass through (such as a judgement regarding the size of a hole), then the contours that lie radially the closest are given greater weight (i.e. those that were previously concave). This led us to the revised version of claim 2, namely that the attentional coordinate frame has associated with it either an inward or outward pointing vector that dictates which portion of a contour will be salient (i.e. outer versus inner). We have argued that the orientation of this vector is task-dependent.

Here we must introduce a contrary note. We do not propose that the attentional frame is imposed only upon a 3D structure seen as an object. Such a view would require that object recognition (or a 2 1/2 sketch [Marr 1982]) had already taken place. Rather, our claim is that the attentional coordinate frame is imposed upon a (frontal plane) silhouette or region prior to recognition (see claim 3 below), and is used to support the object recognition process, such as by indexing the image elements to a model. Hence, because a commitment is made to a coordinate frame and the sign of its object-associated vectors (inward or outward), proper object recognition could be blocked if either the location of the frame or the sign of its radial vectors were chosen improperly.

B.6 What's an Attentional Frame?

Our least controversial claim is that the image region taken as the attentional frame depends upon one's goal. Reversible illusory patterns, such as the Escher drawings or Rubin's face-vase support this claim. The more controversial claim is that the image region taken as attentional region does not have a boundary that can be defined solely in terms of an image contour, even if we include virtual contours such as those cognitive edges formed by the Kanizsa figures. The reason is two fold: First, the focal position of our attentional coordinate frame with respect to the contours determines that part of the contour used in figural similarity judgments, implying that the region attended has changed, or at the very least has been given altered weights. Second, whether the focal position is viewed as part of a passageway or alternatively simply as a hole in an object affects the figural boundary. In each case, the region is understood to lie within an object, but the chosen task affects the details of the region being processed. This effect is also seen clearly in textured C-shaped patterns, and becomes acute when one is asked to judge whether X lies inside the C, or if Y will fit into the C, etc. The virtual boundary assigned to close the C when making such judgments of necessity will also depend in part upon the size of the second object, Y. To simply assert that the attentional window is that region lying inside an image contour misses the point of what the visual information processor is up to.

A simple experiment from the Rock laboratory demonstrates that the attentional window is not simply the entire region of the display, but rather a collection of scene elements. Some of the elements within this "attended" region may be ignored, and thus, not be part of the structures at which higher level visual operations are currently being applied. [Rock and Gutman 1981] showed two overlapping novel outline figures, one red and one green, for a brief period, e.g. one second. Subjects were instructed to rate figures of a given color on the basis of how much they liked them (this attracts attention to one of the figures). They later presented subjects with a new set of outline figures and asked subjects whether they had seen these figures in the previous phase of the experiment, regardless of their color. They found that subjects were very good at remembering the attended shapes but failed on the unattended ones. This experiment agrees with the model presented here, in which only the attended set of structures is being processed. Thus, the attended region is, clearly, not a region of pixels contained in the attended figure because the attended figure was partly contained in such a region and did not yield any high-level perception.

In order to show that what they were studying was failure of perception and not merely selective memory for what was being attended or not attended, [Rock and Gutman 1981] did another experiment. They presented a series, just like in the previous case but with two familiar figures in different pairs of the series, one in the attended color and one in the unattended color. They found that the attended familiar figure was readily recognized but that the unattended familiar figure was not. It is natural that if the unattended figure is perceived and recognized it would stand out. Failure of recognition therefore supports the belief that the fundamental deficit is of perception. The extent of such deficit is unclear; it may be that the level of processing reached for the unattended figures is not complete but goes beyond that of figures not contained in the attended region.

Therefore, an operational definition of "what is an attentional region?" seems more fruitful in trying to understand how images are interpreted. Our definition is in this spirit, and leads to a slightly different view of the initial steps involved in the processing of visual images than those now in vogue in computational vision. This is the subject of the next two sections.

B.7 Figure/Ground and Attentional Frames

In classical perceptual psychology "figure" has a well-defined meaning that is most closely associated with those image regions defined by "occluding" (as opposed to ground, which corresponds to a "partly occluded surface"). Therefore, the classical definition of figure (versus ground) is in terms of three properties: (1) it is perceived as closer to the observer, (2) it has the shape defined by the bounding contour, and (3) it occludes the ground[6].

Our latest claim introduces a complementary, mutually exclusive state to the attentional frame within which figural processing is presumed to occur. When the attentional vector is pointing outward, as in "object mode", this implies that the contour regions associated with the inward state of this vector should be assigned to a separate state.

Consider the following experiment of [Rock and Sigman 1973] in which they showed a dot moving up and down behind a slit or opening, as if a sinusoidal curve was being translated behind it. The experiments were performed with slits of different shapes, so that in some cases the slit was perceived as an occluded surface and in others as an occluding one. They found that the perception of the curve is achieved only if the slit is perceived as an occluded region and not when it is perceived as an occluding region. Using their terms, the "correct" perception is achieved only if the slit is part of ground but not when it is part of the figure. Using our terms, the attentional window has not changed but rather its *attributes* have, because the slit was viewed as a passageway between objects, and not as an object with a hole. Note the difference between ground and attentional region.

In support of our view, another experiment by [Rock and Gilchrist 1975] shows that the attentional window need not correspond to the occluding surface. In this second experiment, they showed a horizontal line moving up and down with one end remaining in contact with one side of an outline figure of a face. Consequently, the line in the display changes in length. When the line is on the side of the face most observers see it changing size, adapting to the outline, while when it is on the other

[6] S. Palmer pointed out to us the importance of the classical definition of figure/ground.

side of the contour, it is seen with constant length but occluded by the face. This has been described as a situation in which no figure-ground reversal occurs. However, in our terms the attentional window has changed because the attended region changes. In the first case the region of attention corresponds to the occluding surface, and in the second to the occluded one. Thus, attentional window need not correspond to the occluding surface, even when the surfaces that are occluded are known. Again, this conclusion is consistent with our definition of the attentional frame and its subsumed region.

B.8 Against Ground; Or Not?; Or What?

Consider now a figure-ground assignment in a situation where you are looking at the edge between two objects that do not occlude each other. For example, the grass in the border of a frozen lake or the edge of your car's door. What is ground in this case? Clearly, in these examples there is not a well-defined foreground and background. Is figure the grass or is it the lake? These examples have been carefully chosen so that depth relations are unclear between objects. In these situations one simply can not assign figure-ground. What is puzzling is that the number of occasions where this happens is very abundant: a bottle and a cap, objects in abstract paintings, the loops of a metallic chain etc.

Our proposal on attentional reference frames does not suffer from this problem, since depth is treated as a figure-ground attribute which need not have a well-defined meaning in all cases. In other words, our notion of attentional frames can be used to explain more perceptual phenomena than the classical notion of figure-ground.

In addition, our observations are difficult to explain in terms of figure-ground. What is the figure and what is the ground in a fir tree? Is the air part of the figure? Our observations require that figure be modified so that it has a fuzzy boundary. This can be seen as an extension of the classical definition. In other words, the insideness measure mentioned above can be translated into a figureness measure. However, this interpretation would leave little role to ground. Furthermore, in some cases the ground is what is capturing the attentional frame. In these latter cases the insideness measure would translate into a groundness measure leaving little role to the figure.

Therefore, figure-ground and attentional frames are different concepts. Attentional frames can easily incorporate into them an attribute that measures figureness, hence capturing the essence of figure-ground. In contrast, without an explicit reference to attention, one can not explain our observations with the classical notion of figure-ground.

B.9 Convexity, Perceptual Organization, Edges, and Frames: Which Comes First?

There have been many proposals on what are the different steps involved in visual perception and it is not the main goal of this research to make yet another such proposal. Nevertheless, our findings have some relevant implications to what should be the nature of these steps which we will now discuss.

We suggest that the attentional frame and objects are not strongly coupled. The attentional window is simply the image-based structures which support some high-level processing, regardless of whether the region is assumed to be an object in the foreground, an occluded object, a hole, a passageway *or none of the above*. Rather, we have shown several examples where the assumptions or role of the region is transformed onto an attribute (such as figure-ground) of the attentional frame that governs both which portions of the contours are included in the processing and the type of processing to be done in it. Curiously, this suggests that a cognitive judgement proceeds and selects that portion of an image or contour to be processed for the task at hand.

But how can a cognitive judgement anticipate where attention will be directed without some preliminary image processing that notes the current contours and edges? We are thus required to postulate an earlier, more reflexive mechanism that directs the eye, and hence the principal focus of attention, to various regions of the image. Computational studies suggest that the location of such focus may involve a bottom-up process such as the one described in [Subirana-Vilanova 1990]. Subirana-Vilanova's scheme computes points upon which further processing is directed using either image contours or image intensities directly. Regions corresponding to each potential point can also be obtained using bottom-up computations[7]. There is other computational evidence that bottom-up grouping and perceptual organization processes can correctly identify candidate interesting structures (see [Marroquin 1976], [Witkin and Tenenbaum 1983], [Mahoney 1985], [Harlick and Shapiro 1985], [Lowe 1984, 1987], [Sha'ashua and Ullman 1988], [Jacobs 1989], [Grimson 1990], [Subirana-Vilanova 1990]).

Psychological results in line with the Gestalt tradition [Wertheimer 1923], [Koffka 1935], [Köhler 1940] argue for bottom-up processes too. However, they also provide evidence that top-down processing is involved. Other experiments argue in this direction, such as the one performed by [Kundel and Nodine 1983] in which a poor copy of a shape is difficult, if not impossible to segment correctly unless one is given some high level help such as "this image contains an object of this type". With the hint, perceptual organization and recognition proceed effortlessly. Other examples of top-down processing include [Newhall 54], [Rock 1983], [Cavanagh 1991], [Friedman-Hill, Wolfe, and Chun 1991],C.M. Mooney and P.B. Porter's binary faces, and R.C. Jones' spotted dog. The role of top-down processing may be really simple, such as controlling or tweaking the behavior of an otherwise purely bottom-up process; or perhaps it involves selecting an appropriate model or structure among several ones computed bottom-up; or perhaps just indexing. In either case the role of top-down processing can not be ignored. Indeed, here we claim that the setting up of the attentional coordinate frame is an important early step in image interpretation.

Our observations suggest that perceptual organization results in regions that are closed or convex (at a coarse scale) as discussed (see also section 5). This corroborates computational studies on perceptual organization which also point in that

---

[7] The result of these computations may affect strongly the choice of reference frames. For example, if the inner stars in Figure B.6 are rotated so as to align with the outer stars (creating convexities in the space between the two), our attention seems more likely to shift to the region in-between the two stars and in this case the similarities will change in agreement with claim 2.

Another way of increasing the preference for the "in-between" reference frame in Figures B.6 and B.7 is by coloring the donut black and leaving the surrounding white (because in human perception there is a bias towards dark objects).

direction, demonstrating the effectiveness and the viability of perceptual organization schemes which limit themselves to finding convex or "enclosed" regions (or at least favor them) [Jacobs 1989], [Huttenlocher and Wayner 1990], [Subirana-Vilanova 1990], [Clemens 1991], [Subirana-Vilanova and Sung]. It is still unclear if this is a general limitation of the visual system, a compound effect with inside and outside, or rather specific to shape perception. There are, however, several areas that may bring some more light onto the question. One of them is the study of the gamma effect: When a visual object is abruptly presented on a homogeneous background, its sudden appearance is accompanied by an expansion of the object. Similarly, a contraction movement is perceived if the object suddenly disappears from the visual field. Such movements were observed a long time ago and were named "gamma" movements by [Kenkel 1913], (see [Kanizsa 1979] for an introduction). For non-elongated shapes, the direction of movement of the figure is generally centrifugal (from the center outward for expansion and from the periphery toward the center for contraction). For elongated shapes, the movement occurs mainly along the perceptual privileged axes. It is unclear whether the movements are involved in the selection of figure or if, on the contrary are subsequent to it. In any case they might be related to a coloring process (perhaps responsible for the expansion movements) involved in figure selection that would determine a non-discrete boundary upon which saliency judgments are established (see also [Mumford, Kosslyn, Hillger and Herrnstein 1987]). If this is true, studying the effect on non-convex shapes (such as those on Figure B.7) may provide cues to what sort of computation is used when the figures are not convex, and to the nature of the inside/outside asymmetry.

Another area that may be interesting to study is motion capture which was observed informally by [Ramachandran and Anstis 83]: When an empty shape is moved in a dynamic image of random dots it "captures" the points that are inside it. This means that the points inside the shape are perceived as moving in the same direction of the shape even though they are, in fact, stationary (randomly appearing for a short interval). This can be informally verified by the reader by drawing a circle on a transparency and sliding it through the screen of a connected TV with noise: The points inside the circle will be perceived as moving along with the circle. The results hold even if the circle has some gaps and it has been shown that they also hold when the shapes are defined by subjective contours [Ramachandran 86].

There is no clear study of what happens for non-convex shapes such as a C. What portions are captured? Informal experiments done in our laboratory seem to confirm that the boundary of the captured region is somewhat fuzzy for unclosed shapes like a C which supports the notion of a fuzzy boundary. In addition, the shape for the captured region seems to have convexity restrictions similar to the ones suggested for the inside-outside relations. It is unclear if both mechanisms are related but the similarity is intriguing. This seems a very promising direction for future research.

Further evidence for the bias towards convex structures is provided by an astonishing result obtained recently by [Cumming, Hurlbert, Johnson and Parker 1991]: when a textured cycle of a sine wave in depth (the upper half convex, the lower half concave) is seen rotating both halfs may appear convex[8], despite the fact that this challenges rigidity[9] (in fact, a narrow band between the two ribbons is seen as moving non-rigidly!).

It is also of interest to study how people perceive ambiguous patterns or tilings [Tuijl 1980], [Shimaya and Yoroizawa 1990] that can be organized in several different ways. It has been shown that in some cases the preference for convex structures can overcome the preference for symmetric structures that are convex [Kanizsa and Gerbino 1976]. The interaction between convex and concave regions is still unclear, especially if the tilings are not complete.

Studies with pigeons[10] [Herrnstein, Vaughan, Mumford and Kosslyn 1989] indicate that they can deal with inside-outside relations so long as the objects are convex but not when they are concave. It is unclear if some sort of "inside-outside" is used at all by the pigeons. More detailed studies could reveal the computation involved, and perhaps whether they use a local feature strategy or a global one. This, in turn, may provide some insights into the limitations of our visual system.

[8] The surface can be described by the equation $Z = sin(y)$ where Z is the depth from the fixation plane. The rotation is along the Y-axis by $+/-$ 10 degrees at 1 Hz.

[9] This observation will be relevant later because it supports the notion that a frame is set in the image before structure from motion is recovered (see claim 3 and related discussion).

[10] The pigeon visual system, despite its reduced dimensions and simplicity, is capable of some remarkable recognition tasks that do not involve explicit inside/outside relations. See [Herrnstein and Loveland 1964], [Cerella 1982], [Herrnstein 1984] for an introduction.

B.10 Against Frame Alignment

As described in the previous sections, our proposal implies that the establishment of a frame of reference is required prior to recognition. In other words, without the frame, which is used to set the saliency of the different image regions, recognition can not proceed. We have pinned down three aspects of it: its location, its size and its inside and outside. Previous research on frames has focused on the orientation of such a frame (relevant results include, to name but a few [Attneave 1967], [Shepard and Metzler 1971], [Rock 1973], [Cooper 1976], [Wiser 1980], [Schwartz 1981], [Shepard and Cooper 1982], [Jolicoeur and Landau 1984], [Jolicoeur 1985], [Palmer 1985], [Corballis and Cullen 86], [Maki 1986], [Jolicoeur, Snow and Murray 1987], [Parsons and Shimojo 1987], [Robertson, Palmer and Gomez 1987], [Shepard and Metzler 1988], [Corballis 1988], [Palmer, Simone and Kube 1988], [Georgopoulos, Lurito, Petrides, Schwartz and Massey 1989], [Tarr and Pinker 1989]), on the influence of the environment ([Mach 1914], [Attneave 1968], [Palmer 1980], [Palmer and Bucher 1981], [Humphreys 1983], [Palmer 1989]), on its location ([Richards and Kaufman 1969], [Kaufman and Richards 1969], [Cavanagh 1978], [Palmer 1983], [Cavanagh 1985], [Nazir and O'Reagan 1990]), and on its size ([Sekuler and Nash 1972], [Cavanagh 1978], [Jolicoeur and Besner 1987], [Jolicoeur 1987], [Larsen and Bundsen 1987]). Exciting results have been obtained in this directions but it is not the purpose to review them here.

The shape of the frame, instead, has received very little attention. The frame alignment approach to recognition suggests that in some cases, a curved frame might be useful (see also [Palmer 1989]). In particular, it suggests the recognition of elongated curved objects, such as the ones shown in Figure B.7, by unbending them using their main curved axis as a frame to match the unbended versions. If human vision used such a scheme, one would expect no differences in the perception of the shapes shown on the top of Figure B.7 from those on the bottom of the same figure. As we have discussed, our findings suggest otherwise, which argues against such a mechanism in human vision.

B.11 Related Effects: What Do You Want to be More Salient?

The shapes used so far in our examples have been defined by image contours. The results, however, do not seem to depend on how such contours are established and similar results seem to hold when the shapes are defined by motion or other discontinuities. Thus, the results seem to reflect the true nature of shape perception. In this section we will suggest that similar biases in saliency occur in other dimensions of visual perception. What all of them have in common is that they require the establishment of an attentional frame of reference at an early stage, and that the nature of the frame depends on the task at hand. In particular, we will suggest that: top is more salient than bottom, near is more salient than far and outward motion is more salient than inward motion.

Top is more salient than bottom; or not.

Consider the contours in Figure B.8, the center contour appears more similar to the one on the right than to the one on the left. We suggest that this is because the top of the contours is, in general, more salient than its bottom. We can provide functional justification similar to that given in the inside-outside case: the top is more salient because, by default, the visual system is more interested in it, as if it were the part of a surface that we contact first. Just like with our inside-outside notion, the outcome can be reversed by changing the task (consider they are the roof of a small room that you are about to enter). Thus, there is an asymmetry on the saliency of the two sides of such contour (top and bottom) similar to the inside/outside one discussed in the previous sections.

Near is more salient than far; or not.

When looking for a fruit tree of a certain species it is likely that, in addition we are interested in finding the one that is closer to us. Similarly, if we are trying to grasp something that is surrounded by other objects, the regions that are closer to our hand are likely to be of more interest than the rest of the scene. We suggest that when three-dimensional information is available, the visual system emphasizes the closer regions of the scene. Evidence is shown in Figure B.9 in which we show a stereo pair with surfaces similar to the silhouette of the star of Figure B.3.

At a first glance, most see two of the three surfaces of Figure B.3 as being more similar. The preference, as in the previous case, can be reversed if we change the task: imagine, for example, that you are flying above such surfaces and are looking for a place to land. Your attention will change to the far portions of the surfaces and with it your preferred similarities. Therefore, attention and the task at hand play an important role in determining how we perceive the three-dimensional world. Note also, that, as in the previous examples, a matching measure based on the distance between two surfaces can not account for our observations. For in this case, such distance to the center surface is the same for both bounding surfaces.

Expansion is more salient than contraction; or not.

Is there a certain type of motion that should be of most interest to the human visual system? Presumably, motion coming directly toward the observer is more relevant than motion away from it. Or, similarly, expanding motion should be more salient than contracting motion. Evidence in support of this suggestion is provided by a simple experiment illustrated in Figure B.10[11]. Like in the previous cases, two seemingly symmetric percepts are not perceived equally by the visual system. This distinction, again, seems to bear on some simple task-related objectives of the observer.

So, what's more salient? How does perception work?

Inside/outside, near/far, expansion/contraction and top/bottom are generally not correlated. If saliency were determined independently for each of these relations, then conflicts could arise in some cases. For example, the inside of an object may be

[11]In a pool of 7 MIT graduate students, all but one reported that their attention was directed first at the expanding pattern.

near or far, in the top or in the bottom of the image. Will, in this case, the outside regions on the bottom be more salient than those that are inside and on the top?

This is an important issue that will not be addressed here. A more detailed understanding of how attention and perceptual organization interact with the early vision modules is required. In any case, it would be interesting to find a modular division showing how these processes may interact. Unfortunately, this is a no-win situation. Either the modules are too few to be interesting or the division is easily proven to be wrong. Nevertheless, it may be useful to give a proposal as precise as possible to illustrate what has been said so far. Figure B.11 is it.

Like in [Witkin and Tenenbaum 1983], our proposal is that grouping is done very early (before any 2 1/2 D sketch-like processing), but we point out the importance of selecting a coordinate frame which, among other things, is involved in top-down processing and can be used to index into a class of models. Indexing can be based on the coarse description of the shape that the frame can produce, or on the image features associated with the frame. As shown in Figure B.11, this frame may later be enhanced by 3D information coming from the different early vision modules. Like in [Jepson and Richards 91], we suggest that one of the most important roles of the frame is to select and articulate the processing on the "relevant" structures of the image (see also footnote 7). This leads us to the last claim of the Appendix:

Claim 3 *An attentional "coordinate" frame is imposed in the image prior to constructing an object description for recognition.*

In fact, the version of Curved Inertia Frames presented in Chapter 3 computes frame curves prior to an object description. In addition, Curved Inertia Frames can locate convex structures to support an object description for recognition.

B.12 What's New

The fact that figure and ground reversals are attention related has been known for some time [Rubin 1921][12]. However, there appears to be no precise statement of the relation between "figure" and notions of "inside" and "object", nor has it been noted previously that contour saliency depends on *inside/outside, near/far, expansion/contraction and top/bottom* relations, and changes when the *task* is changed, such as viewing a region as something to pass thru, rather than as a shape to be recognized.

These new observations support an operational definition of *reference frames* which are based on attention. We have suggested that occlusion be treated as an attribute of the attentional frame. A key ingredient is *the processing focus*, not an image region typically defined as "figure". Clearly, any proposal that relates an attentional window to object fails: due to the existence of fuzzy boundaries. The idea that the processing focuss has a non-discrete boundary has not been suggested previously. This leads to the concept of *frame curve* which can be used for *shape segmentation* in conjunction with Inside/Outside relations.

Our findings also demonstrate that the task at hand controls *top-down processing*. Existing evidence for top-down processing shows its role in increasing the speed and performance of recognition (by providing hints, such as restricting the set of models to be considered). However, a qualitative role of top-down processing (such as determining whether we are looking for an object or a hole), not dependent on the image, like the one presented here, suggests new directions for inquiries.

Finally, we have shown that "matching to model" will not correspond with human perception unless inside/outside, top/bottom, expansion/contraction and near/far relations are factored early in the recognition strategy. We have also discussed several ways in which the role of *convexity* can be studied in human vision, such as inside/outside relations, gamma movements and motion capture. Our observations

[12][Rubin 1921] showed subjects simple contours where there was a two way ambiguity in what should be figure (similar to the reversible figure in the top of Figure B.4). He found that if one region was found as figure when shown the image for the first time then, if on subsequent presentations the opposite region was found as figure, recognition would not occur.

provide new insight into the nature of the attention and perceptual organization processes involved in visual perception. In particular, they indicate that a *frame* is set prior to recognition, (challenging, among other things, the early role of rigidity in motion segmentation) and agree with a model in which recognition proceeds by the successive processing of *convex* chunks of *image structures* defined by this frame.

Note that the notions of fuzzy boundaries and frame curve reinforce the idea that discontinuities should not be detected early on but only after a frame has been computed (see also Chapter 3).

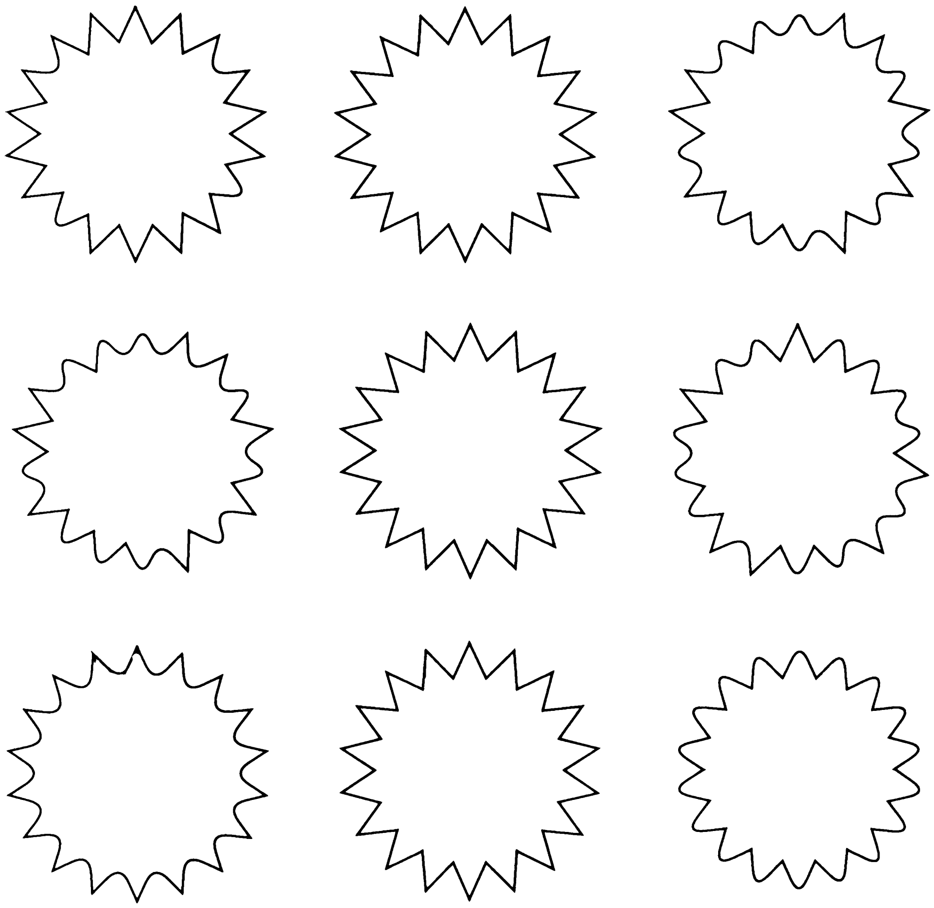


Figure B.3: *Top row:* Use the middle pattern as reference. Most see the left pattern as more similar to the reference. This could be because it has a smaller number of modified corners (with respect to the center) than the right one, and therefore, a pictorial match is better. *Second row:* In this case, the left and right stars look equally similar to the center one. This seems natural if we consider that both have a similar number of corners smoothed. *Third row:* Most see the left pattern as more similar despite the fact that both, left and right, have the same number of smoothed corners with respect to the center star. Therefore, in order to explain these observations, one can not base an argument on just the number of smoothed corners. The position of the smoothed corners need be taken into account, i.e. preferences are not based on just pictorial matches. Rather, here the convexities on the outside of the patterns seem to drive our similarity judgement.

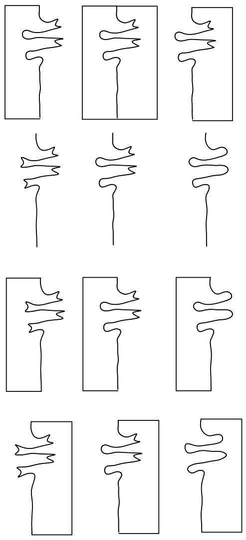


Figure B.4: *Top:* Reversible figure. *Second Row:* The contour (shown again in the center) that defined the previous reversible figure is modified in two similar ways (left and right contours). *Third and fourth row:* When such three contours are closed a preference exists, and this preference depends for most on the side used to close the contour. Use the center shape as reference in both rows. As in the example of the previous Figure most favor the outer portions of the shape to judge similarity. A distance metric, based solely on a pictorial match and that does not take into account the relative location of the different points of the shape, can not account for these observations.

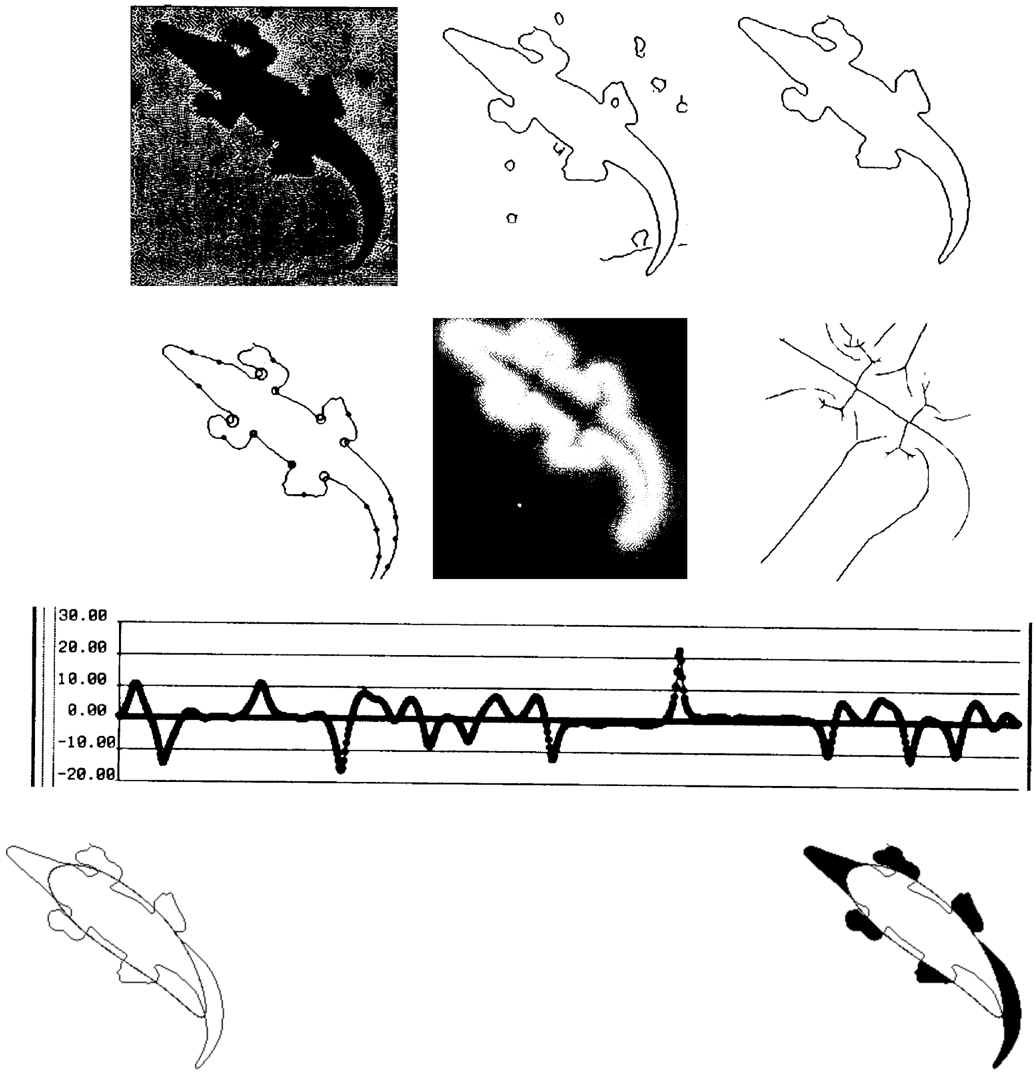


Figure B.5: *Top left:* Alligator image. *Top center:* Edges of alligator image. *Top right:* Alligator edges (the longer ones). *Second row left:* Alligator parts defined as negative minima of curvature [Hoffman and Richards 1984] (curvature plot bellow and extrema overlaid as circles on alligator edges) images illustrate the part decomposition on the alligator image using the negative-minima rule. *Second row center:* Distance transform. *Second row right:* Extrema of distance transform. *Bottom left:* Alligator with *frame curve* superimposed which has been computed using a standard smoothing algorithm. *Bottom right:* The different parts of the alligator are shaded using the *outside rule.*

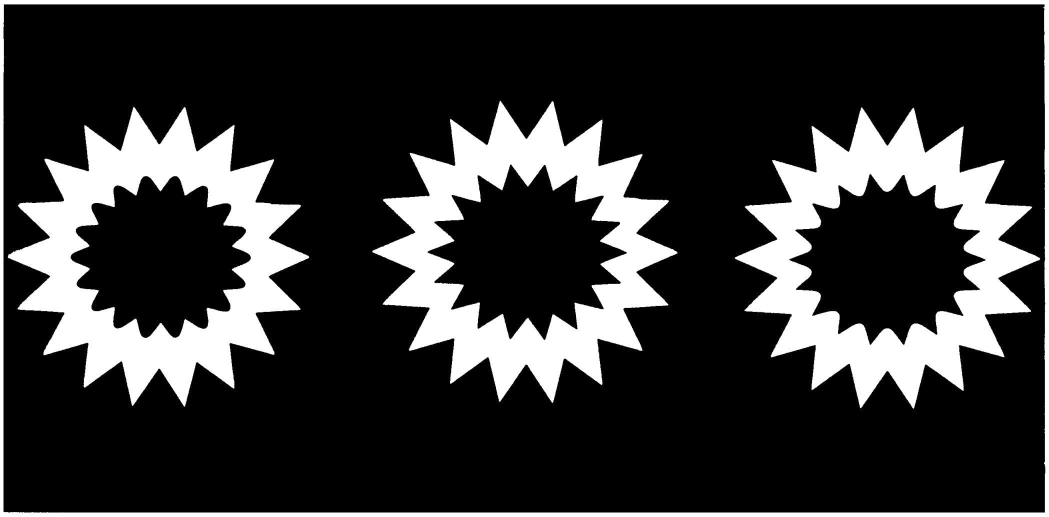


Figure B.6: Star patterns with "holes" treat the inside ring of the shape as if this ring was an occluding shape, i.e. as if it was independent from the surrounding contours, even if one perceives a donut like shape.

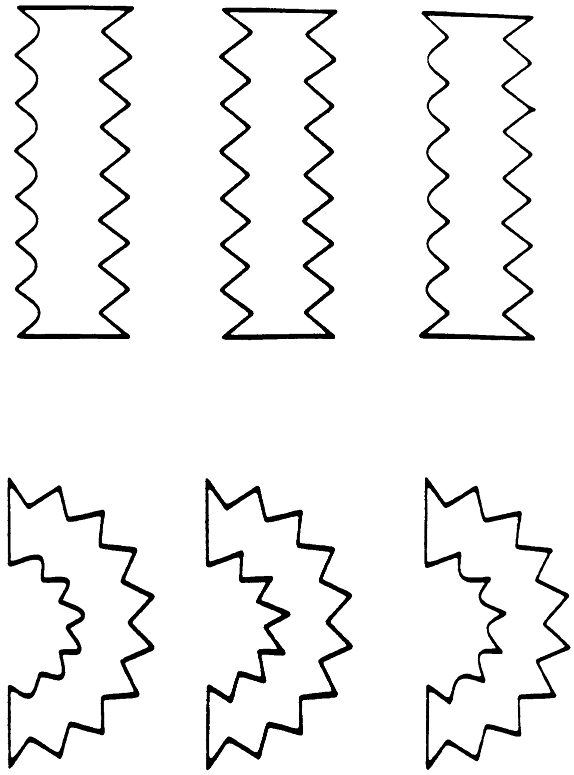


Figure B.7: *Top:* Using the middle shape as a reference, most see the left shape as more similar. *Bottom:* If this same shape is bent, the situation becomes confused.

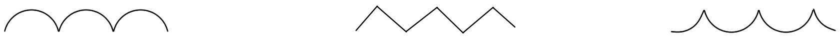


Figure B.8: Top is more salient than bottom: Using the middle pattern as reference, most see the right contour as more similar.

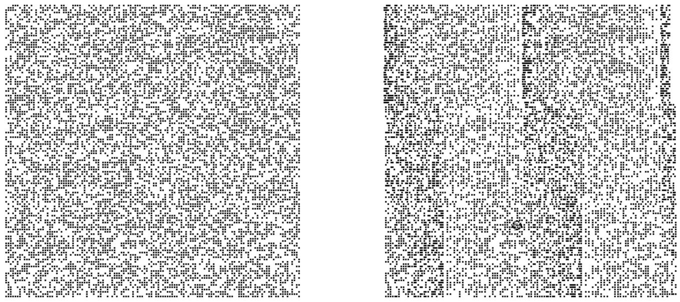


Figure B.9: This random dot stereo diagram illustrates that close structures are more salient in visual perception than those that are further away.

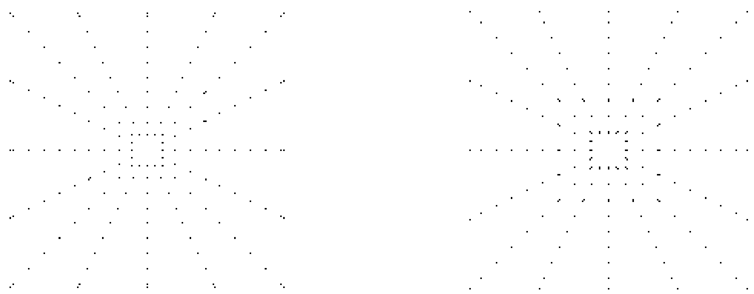


Figure B.10: When the dots in this two similar figures are moving towards the center on one of them, and towards the outside on the other, attention focuses first on the expanding flow. This provides evidence that motion coming directly toward the observer is more salient than motion away from it.

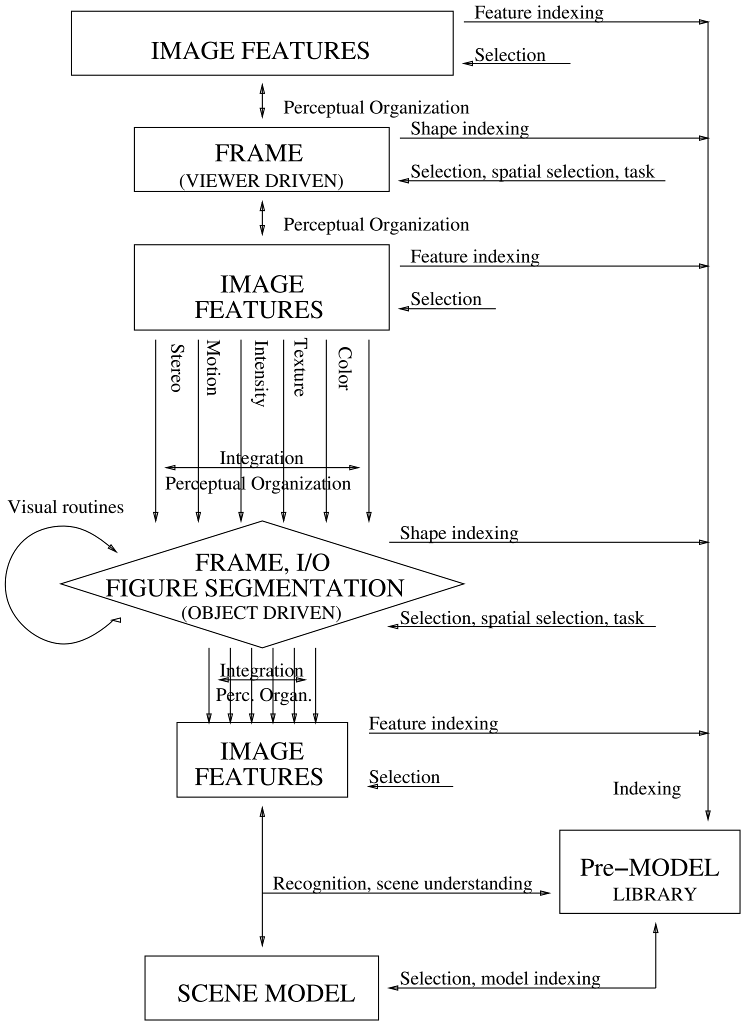


Figure B.11: A modular description of visual perception that illustrates some of the concepts discussed here. The different frames and image features depicted may share data structures; this accounts for some implicit feed-back in the diagram. (The Figure emphasizes the order in which events take place, not the detailed nature of the data structures involved.) It is suggested that perception begins by some simple image features which are used to compute a frame that is used to interpret these image features. The frame is an active structure which can be modified by visual routines. In this diagram, shape appears as the main source for recognition but indexing plays also an important role. Indexing can be based on features and on a rough description of the selected frame.

Bibliography

[1] J-F. Abramatic and O. Faugeras. Sequential convolution techniques for image filtering. *IEEE Trans. Acoust., Speech, Signal Processing*, 30(1):1–10, 1982.

[2] C. Arcelli, L.P. Cordella, and S. Levialdi. From local maxima to connected skeletons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-3(2):134–143, 1981.

[3] H. Asada and M. Brady. The curvature primal sketch. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(1):2–14, 1985.

[4] F. Attneave. Triangles as ambiguous figures. *American Journal of Psychology*, 81:447–453, 1968.

[5] F. Attneave and R.K. Olson. Discriminability of stimuli varying in physical and retinal orientation. *Journal of Experimental Psychology: Human Perception and Performance*, 74:149–157, 1967.

[6] N. Badler and R. Bajcsy. Three-dimensional representations for computer graphics and computer vision. *Computer Graphics*, 12:153–160, 1978.

[7] S.C. Bagley. Using models and axes of symmetry to describe two-dimensional polygonal shapes. Master's thesis, Massachusetts Institute of Technology, 1985.

[8] H. Baird. *Model-based image matching using location.* The MIT Press, Cambridge, MA, 1985.

[9] R. Bajcsy. Computer description of textured surfaces. In *Proceedings IJCAI*, pages 572–579, 1973.

[10] R. Bajcsy and S. Kovacic. Multiresolution elastic matching. Technical Report MS-CIS-87-94, Department of Computer and Information Science, School of Engineering and Applied Science, University of Philadelphia, 1987.

[11] H.B. Barlow. The past, present and future of feature detectors. In D.G. Albrecht, editor, *Recognition of pattern and form. Proceedings of a conference held at the University of Texas at Austin, 1979*, pages 4–32. Springer-Verlag, 1982.

[12] H.B. Barlow and B.C. Reeves. The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Research*, 19:783–793, 1979.

[13] J. Beck. Textural segmentation. In Jacob Beck, editor, *Organization and Representation in Perception.* Erlbaum, Hillsdale, NJ, 1982.

[14] P.J. Besl and R.C. Jain. Three-dimensional object recognition. *ACM Computing Surveys*, 17(1):75–145, 1988.

[15] C.E. Bethell-Fox and R.N. Shepard. Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, 1:12–23, 1988.

[16] D. Beymer. Finding junctions using the image gradient. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 720–721, Ann Arbor, MI, 1991.

[17] I. Biederman. Human image understanding: Recent research and a theory. *Computer Vision, Graphics, and Image Processing*, 32:29–73, 1985.

[18] I. Biederman and G. Ju. Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, 20:38–64, 1988.

[19] T.O. Binford. Visual perception by a computer. In *Proceedings of the IEEE Conf. on Systems and Controls, Miami*, 1971.

[20] H. Blum. A transformation for extracting new descriptors of shape. In Walthen Dunn, editor, *Models for the perception of speech and visual form*, pages 362–380. The MIT Press, Cambridge, MA, 1967.

[21] H. Blum and R.N. Nagel. Shape description using weighted symmetric axis features. *Pattern Recognition*, 10:167–180, 1978.

[22] R.C. Bolles and R.A. Cain. Recognizing and locating partially visible objects: The local-feature-focus method. *International Journal of Robotics Research*, 1(3):57–82, 1982.

[23] G. Borgefors. Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, 34:344–371, 1986.

[24] E.G. Boring. *A history of experimental psychology.* Appleton Century Crofts, Inc., 1964. Second Edition. Originally published 1929.

[25] Bovik, Clark, and Geisler. Multichannel texture analysis using localized spatial filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-12:56–65, 1990.

[26] M. Brady. Criteria for the representations of shape. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human and Machine Vision*, pages 39–84. Academic Press, New York, 1983.

[27] M. Brady. Seeds of perception. In *Proceedings of the Third Alvey Vision Conference, (Cambridge University, 15–17 September)*, pages 259–266. The University of Sheffield Printing Unit, 1987.

[28] M. Brady and H. Asada. Smoothed local symmetries and their implementation. *International Journal of Robotics Research*, 3(3):36–61, 1984.

[29] M. Brady and G. Scott. Parallel algorithms for shape representation. In Ian Page, editor, *Parallel Architectures and Computer Vision.* OUP, 1988.

[30] M. Brady and A. Yuille. An extremum principle for shape from contour. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(3):288–301, 1984.

[31] T.M. Breuel. Geometric aspects of visual object recognition. Technical Report AI-TR-1374, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992.

[32] C. Broit. *Optimal registration of deformed images.* PhD thesis, University of Pennsylvania, 1981.

[33] R.A. Brooks. Symbolic reasoning among 3-D models and 2-D images. *Artificial Intelligence*, 17:285–348, 1981.

[34] R.A. Brooks, G. Russell, and T. Binford. The acronym model based vision system. In *Proceedings IJCAI*, pages 105–113, 1979.

[35] D.J. Burr. Elastic matching of line drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-3(6):708–712, 1981.

[36] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8:679–698, 1986.

[37] P.A. Carpenter and M.A. Just. Eye fixations during mental rotation. In J.W. Senders, D.F. Fisher, and R.A.Monty, editors, *Eye movements and the higher psychological functions*, pages 115–133. Erlbaum, Hillsdale, NJ, 1978.

[38] P. Cavanagh. Size and position invariance in the visual system. *Perception*, 7:167–177, 1978.

[39] P. Cavanagh. Local log polar frequency analysis in the striate cortex as a basis for size and orientation invariance. In D. Rose and V. Dobson, editors, *Models of the visual cortex*, pages 85–95. Wiley, 1985.

[40] P. Cavanagh. Reconstructing the third dimension: Interactions between color, texture, motion, binocular disparity, and shape. *Computer Vision, Graphics, and Image Processing*, 37:171–195, 1987.

[41] P. Cavanagh. What's up in top-down processing? In Andrei Gorea, Yxes Fregnac, Zoi Kapoula, and John Findlay, editors, *Representations of vision. Trends and tacit assumptions in vision research. A collection of essays based on the 13th european conference on visual perception organised by Andrei Gorea and held in Paris at the Cité des Sciences ed de l'Industrie in September 1990*, pages 295–304. Cambridge University Press, Cambridge, U.K., 1991.

[42] J. Cerella. Mechanisms of concept formation in the pigeon. In D.J. Ingle, M.A. Goodale, and R.J.W. Mansfield, editors, *Analysis of visual behavior*, pages 241–260. The MIT Press, Cambridge and London, 1982.

[43] F.H. Cheng and W.H. Hsu. Parallel algorithm for corner finding on digital curves. *Pattern Recognition Letters*, 8:47–53, 1988.

[44] R.T. Chin and C.R. Dyer. Model-based recognition in robot vision. *ACM Computing Surveys*, 18(1):68–108, 1986.

[45] J.J. Clark. Singularity theory and phantom edges in scale space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-10:720–727, 1988.

[46] D. Clemens. *Region-based feature interpretation for recognizing 3D models in 2D images*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1991.

[47] J.H. Connell. Learning shape descriptions: generating and generalizing models of visual objects. Technical Report AI TR No. 853, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1985.

[48] J.H. Connell and M. Brady. Generating and generalizing models of visual objects. *Artificial Intelligence*, 31:159–183, 1987.

[49] L.A. Cooper. Demonstration of a mental analog to an external rotation. *Perception and Psychophysics*, 1:20–43, 1976.

[50] M.C. Corbalis. Recognition of disoriented shapes. *Psychological Review*, 95:115–123, 1988.

[51] M.C. Corbalis and S. Cullen. Decisions about the axes of disoriented shapes. *Mem. Cognition*, 14:27–38, 1986.

[52] T.R. Crimmins. A complete set of fourier descriptors for two-dimensional shapes. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-12:848–855, 1982.

[53] B.G. Cumming, A.C. Hurlbert, E.B. Johnson, and A.J. Parker. Effects of texture and shading on the kde. In *The Association for Research in Vision and Ophthalmology. Annual Meeting Abstract Issue. Vol 32, NO. 4*, page 1277, Bethesda, Maryland 20814-3928, 1991.

[54] J.B. Deregowski. Real space and represented space: Cross-cultural perspectives. *Behavioral and Brain Sciences*, 12:51–119, 1989.

[55] S.E. Dreyfus. *Dynamic programming and the calculus of variations*. Academic Press, New York, NY, 1965.

[56] G. Dudek. Shape description and classification using curvature scale-space measurements. In Springer-Verlag, editor, *Proceedings of a NATO Workshop Shape in Picture*, 1993.

[57] M.J. Eccles, M.P.C. McQueen, and D. Rosen. Analysis of the digitized boundaries of planar objects. *Pattern Recognition*, 9:31–41, 1977.

[58] S. Edelman. *Reading and writing of cursive script: a computational study*. PhD thesis, Weizmann Institute of Science, 1988.

[59] F. Etesami and Jr J.J. Uicker. Automatic dimensional inspection of machine part cross-sections using fourier analysis. *Computer Vision, Graphics, and Image Processing*, 29:216–247, 1985.

[60] C.R. Feynman. Modeling the appearance of cloth. Master's thesis, Massachusetts Institute of Technology, 1983.

[61] M.M. Fleck. Local rotational symmetries. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 332–337, 1986.

[62] M.M. Fleck. Boundaries and topological algorithms. Technical Report AI-TR-1065, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1988.

[63] I. Fogel and D. Sagi. Gabor filters as texture discriminators. *Biological Cybernetics*, 61:103–113, 1989.

[64] P. Foldiak. Learning invariance from transformation sequences. *Neural Computation*, 3:194–200, 1991.

[65] D. Forsyth and A. Zisserman. Mutual illumination. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 466–473, 1989.

[66] W.T. Freeman. *Steerable filters and local analysis of image structure.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1992.

[67] W.T. Freeman and E. H. Adelson. Steerable filters for early vision, image analysis and wavelet decomposition. In *Third International Conference on Computer Vision*, pages 406–415. IEEE Computer Society, 1990.

[68] S. Friedberg. Finding axes of skewed symmetry. *Computer Vision, Graphics, and Image Processing*, 34:138–155, 1986.

[69] S.R. Friedman-Hill, J.M. Wolfe, and M.M. Chun. Further evidence for top-down activation of orientation categories in visual search. In *ARVO Annual Meeting*, 1991.

[70] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shifts in position. *Biological Cybernetics*, 36:193–202, 1980.

[71] J.M. Gauch. The multiresolution intensity axis of symmetry and its application to image segmentation. Technical Report TR89-047, Dept. of Computer Science, Univ. of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3175, 1989.

[72] J.M. Gauch and S.M. Pizer. The intensity axis of symmetry and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(8):753–763, 1993.

[73] D. Geiger and T. Poggio. An optimal scale for edge detection. In *Proceedings IJCAI*, pages 745–748, Milan, Italy, 1987.

[74] Stuart Geman and Don Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6:721–741, 1984.

[75] M.A. Gennert. Detecting half-edges and vertices in images. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 552–557, 1986.

[76] A.P. Georgopoulos, J.T. Lurito, M. Petrides, A.B. Schwartz, and J.T. Massey. Mental rotation of the neuronal population vector. *Science*, 243:234–236, 1989.

[77] C.R. Giardina and F.P. Kuhl. Accuracy of curve approximation by harmonically related vectors with elliptical loci. *Computer Vision, Graphics, and Image Processing*, 6:277–285, 1977.

[78] G. Giraudon and R. Deriche. On corner and vertex detection. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 650–655, Lahaina, Maui, Hawaii, 1991.

[79] B. Gluss. *An elementary introduction to dynamic programming: a state equation approach.* Allyn and Bacon, Inc., Boston, MA, 1975.

[80] N.H. Goddard. The perception of articulated motion: recognizing moving light displays. Technical Report 405, Dept. of Computer Science, University of Rochester, 1992.

[81] E. Goldmeier. Similarity in visually perceived forms. *Psychological Issues*, 8:Number 29, 1972. (Originaly published 1936).

[82] G.H. Granlund. Fourier preprocessing for hand print character recognition. *IEEE Computer*, 21:195–201, 1972.

[83] R.L. Gregory. *The intelligent eye.* McGraw-Hill Book Company, New York, St. Louis and San Francisco, 1970.

[84] W. Eric L. Grimson. *From Images to Surfaces.* MIT Press, Cambridge, Mass., 1981.

[85] W.E.L. Grimson. *Object Recognition By Computer: The Role Of Geometric Constraints.* The MIT Press, Cambridge and London, 1990.

[86] A. Guzman. Decomposition of a visual scene into three-dimensional bodies. In *Proc. AFIPS 1968 Fall Joint Computer Conference*, pages 291–304, 1968.

[87] J.R. Hamerly and R.M. Springer. Raggedness of edges. *J. Opt. Soc. Am.*, 71(3):285–291, 1981.

[88] L.G.C. Hamey. Computer perception of repetitive textures. Technical Report CMU-CS-88-149, Carnegie-Mellon University, 1988.

[89] A. R. Hanson and E. M. Riseman. Segmentation of natural scenes. *CVS*, 1978.

[90] R.M. Haralick and L.G. Shapiro. Image segmentation techniques. *Computer Vision, Graphics, and Image Processing*, 29:100–132, 1985.

[91] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the Fourth Alvey Vision Conference (Manchest er University, 31st August–2nd September)*, pages 147–152. The University of Sheffield Printing Unit, 1988.

[92] D. Heeger. Optical flow from spatiotemporal filters. In *Proceedings of the Second International Conference on Computer Vision*, pages 181–190, 1988.

[93] S.S. Heide. A hierarchical representation of shape from smoothed local symmetries. Master's thesis, Massachusetts Institute of Technology, 1984.

[94] H. Helson. Fundamental principles in color vision. *Journal of Exp. Psychol.*, 23:439–471, 1938.

[95] R.J. Herrnstein. Objects, categories, and discriminative stimulus. In H.L. Roitblat, T.G. Bever, and H.S. Terrace, editors, *Animal Cognition: Proceedings of the Frank Guggenheim Conference*. Lawrence Erlbaum Associates, Hillsdale N.J, 1984.

[96] R.J. Herrnstein, W. Vaughan Jr., D.B. Mumford, and S.M. Kosslyn. Teaching pigeons an abstract relational rule: Insideness. *Perception and Psychophysics*, 46(1):56–64, 1989.

[97] R.J. Herrnstein and D. Loveland. Complex visual concept in the pigeon. *Science*, 46:549–551, 1964.

[98] D.D. Hoffman and W.A. Richards. Parts of recognition. In S. Pinker, editor, *Visual Cognition*, pages 2–96. The MIT Press, Cambridge, MA, 1984.

[99] J. Hollerbach. Hierarchical shape description of objects by selection and modification of prototypes. Technical Report AI TR No. 346, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1975.

[100] B.K.P. Horn. Image intensity understanding. *Artificial Intelligence*, 8(2):201–231, 1977.

[101] S. L. Horowitz and T. Pavlidis. Picture segmentation by a direct split and merge procedure. In *Proceedings Int. Conf. on Pattern Recognition*, pages 424–433, August 1974.

[102] M.K. Hu. Visual pattern recognition by moment invariants. *IRE Trans. Infrom. Theory*, IT-8:179–187, 1962.

[103] G.W. Humphreys. Reference frames and shape perception. *Cognitive Psychology*, 15:151–196, 1983.

[104] A. Hurlbert and T. Poggio. Synthetizing a color algorithm from examples. *Science*, 239:482–485, 1988.

[105] D.P. Huttenlocher and S. Ullman. Recognizing rigid objects by aligning them with an image. *Massachusetts Institute Technology AI Memo 937*, 1987.

[106] D.P. Huttenlocher and P. Wayner. Finding convex edge groupings in an image. TR 90-1116, Department of Computer Sciences, Cornell University, Ithaca, New York, 1990.

[107] Y.K. Hwang and N. Ahuja. Path planning using a potential field representation. UILU-ENG-88. 2251, Coordinated Science Laboratory, University of Illinois, Urbana-Champaign, Illinois, 1988.

[108] D.W. Jacobs. The use of grouping in visual object recognition. A.I. Technical Report No. 1023, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1988.

[109] D.W. Jacobs. Grouping for recognition. A.I. Technical Report No. 1117, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[110] R.C. James. The dalmatian dog. This photograph was made by R.C. James and has been reproduced in a number of different publications. A binary version can be found in [Marr 82] and a non-discretized one in [Gregory 1970].

[111] A. Jepson and W. Richards. Integrating vision modules. *IEEE Systems and Cybernetics*, 1991. Special Issue on Assimilation. Editor: A. Jain. To appear.

[112] P. Jolicoeur. The time to name disoriented natural objects. *Mem. Cognition*, 13(4):289–303, 1985.

[113] P. Jolicoeur. A size-congruency effect in memory for visual shape. *Mem. Cognition*, 15(6):531–543, 1987.

[114] P. Jolicoeur and D. Besner. Additivity and interaction between size ratio and response category in the comparison of size-discrepant shapes. *Journal of Experimental Psychology: Human Perception and Performance*, 13(3):478–487, 1987.

[115] P. Jolicoeur and S.M. Kosslyn. Is time to scan visual images due to demand characteristics. *Mem. Cognition*, 13(4):320–332, 1985.

[116] P. Jolicoeur and M.J. Landau. Effects of orientation on the identification of simple visual patterns. *Canadian Journal of Psychology*, 38(1):80–93, 1984.

[117] P. Jolicoeur, D. Snow, and J. Murray. The time to identify disoriented letters: Effects of practice and font. *Canadian Journal of Psychology*, 41(3):303–316, 1987.

[118] D. Jones and J. Malik. Computational stereopsis–beyond zero-crossings. *Invest. Ophtalmol. Vis. Sci. (Supplement)*, 31(4):529, 1990.

[119] D. B. Judd and G. Wyszecki. *Color in Business, Science and Industry (Third Edition).* Wiley, New York, 1975.

[120] D.B. Judd. Appraisal of land's work on two-primary color projections. *Journal of the Optical Society of America*, 50:254–268, 1940.

[121] B. Julesz. A theory of preattentive texture discrimination based on first-order st atistics of textons. *Biological Cybernetics*, 41:131–138, 1981.

[122] B. Julesz. Texton gradients: the texton theory revisited. *Biological Cybernetics*, 54:245–251, 1986.

[123] B. Julesz and J.R. Bergen. Textons, the fundamental elements in preattentive vision and perception of textures. *Bell Syst Tech J.*, 62:1619–1645, 1983.

[124] C. Kambhamettu, D.B. Goldgof, D. Terzopoulos, and T.S. Huang. Nonrigid motion analysis. In T.Y. Young, editor, *Handbook of Pattern Recognition and Image Processing; Vol. 2: Computer Vision*, page In press. Academic Press, San Diego, CA, 1993.

[125] G. Kanizsa. *Organization in Vision.* Praeger, 1979.

[126] G. Kanizsa and W. Gerbino. Convexity and symmetry in figure-ground organization. In M. Hele, editor, *Vision and Artifact.* Springer, New York, 1976.

[127] M. Kass. Computing visual correspondence. In *From Pixels to Predicates*, pages 78–92. Ablex Publishing Corporation, Norwood, NH, 1986.

[128] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, pages 321–331, 1988.

[129] Michael Kass and Andrew Witkin. Analyzing oriented patterns. In W. Richards, editor, *Natural Computation*, pages 252–264. M.I.T. Press, 1988.

[130] L. Kaufman and W. Richards. Spontaneous fixation tendencies for visual forms. *Perception and Psychophysics*, 5(2):85–88, 1969.

[131] J.R. Kender and R. Kjeldsen. On seeing spaghetti: a novel self-adjusting seven parameter hough space for analyzing flexible extruding objects. Technical Report RC 16577 (#73582), IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY 10598, 1991.

[132] F. Kenkel. Untersuchungen uber den zusammenhang zwischen erscheinungsgrosse und erscheinungsbewegung bei einigen sogennanten optischen tauschungen. *Zeitschrift fur Psychologie*, 67:358–449, 1913.

[133] A.E. Kertesz. Effect of stimulus size on fusion and vergence. *J. Opt. Soc. Am.*, 71(3):289–295, 1981.

[134] O. Khatib. Real time obstacle avoidance for manipulators and mobile robots. *International Journal of Robotics Research*, 5(1):90–98, 1986.

[135] P. Khosla and R. Volpe. Superquadric artificial potentials for obstacle avoidance and approach. In *Proceedings of IEEE Conference on Robotics and Automation*, pages 1778–1784, Philadelphia, Pennsylvania, 1988. IEEE-Computer Society Press.

[136] R. Kimchi and S.E. Palmer. From and texture in hierarchically constructed patterns. *Journal of Experimental Psychology: Human Perception and Performance*, 8(4):521–535, 1982.

[137] H. Knuttson and G.H. Granlund. Texture analysis using two-dimensional quadrature filters. In *Workshop on Computer Architecture for Pattern Analysis and Image Database Management*, pages 206–213, IEE Computer Society, 1983.

[138] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[139] J.J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.

[140] J.J. Koenderink. The structure of images. *Biol. Cybern.*, 50:363–370, 1984.

[141] J.J. Koenderink and A. van Doorn. A description of the structure of visual images in terms of an ordered hierarchy of light and dark blobs. In *Proc. 2nd Int. Conf. on Vis. Psychophysics and Med. Imaging. IEEE cat. no 81CH1676-6*, 1981.

[142] A. Koffka. *The principles of Gestalt psychology.* Harcourt, Brace, New York, 1940.

[143] W. Kohler. *Dynamics in psychology.* Liveright, New York, 1940.

[144] A.F. Korn. Toward a symbolic representation of intensity changes in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(5):610–625, 1988.

[145] F.P. Kuhl and C.R. Giardina. Elliptic fourier features of a closed contour. *Computer Vision, Graphics, and Image Processing*, 18:236–258, 1982.

[146] H.L. Kundel and C.F. Nodine. A visual concept shapes image perception. *Radiology*, 146(2):363–368, 1983.

[147] M.S. Langer and S.W. Zucker. Shape from shading on a cloudy day. Technical Report TR-CIM-91-7, McRCIM, McGill University, Montreal, Canada, 1992.

[148] A. Larsen and C. Bundesen. Size scaling in visual pattern recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 4(1):1–20, 1978.

[149] Y. LeCun, B. Boser, J.S. Benker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. On using neural networks for the recognition of handwriting. *Neural Computation*, 1:541–551, 1989.

[150] Frédéric Leymarie. Tracking and describing deformable objects using active contour models. Report No. TR-CIM 90-0, McGill Research Center for Intelligent Machines, Montréal, Québec, Canada, 1990.

[151] P. Lipson, A.L. Yuille, D. O'Keefe, J. Cavanaugh, J. Taaffe, and D. Rosenthal. Automated bone density calculation using feature extraction by deformable templates. Technical Report 89-14, Harvard Robotics Laboratory, 1989.

[152] D.G. Lowe. *Perceptual Organization and Visual Recognition.* PhD thesis, Standford University, 1984.

[153] D.G. Lowe. *Perceptual Organization and Visual Recognition.* Kluwer Academic Publishers, Boston, 1986.

[154] D.G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:355–395, 1987.

[155] Y. Lu and R.C. Jain. Behavior of edges in scale space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11:337–356, 1989.

[156] E. Mach. *The analysis of sensations.* Chicago: Open Court, 1914.

[157] A.J. Maeder. Polygonal harmonic shape characterization. In Springer-Verlag, editor, *Proceedings of a NATO Workshop Shape in Picture*, 1993.

[158] S.T.F. Mahmood. Attentional selection in object recognition. Technical Report AI-TR-1420, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1993.

[159] J.V. Mahoney. Image chunking: defining spatial building blocks for scene analysis. A.I. Technical Report No. 980, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1987.

[160] J.V. Mahoney. Image chunks and their applications. Technical Report EDL-92-3, Xerox Corporation Palo Alto Research Center, Palo Alto, CA 94304, 1992.

[161] J.V. Mahoney. Image links and their applications. Technical Report EDL-92-2, Xerox Corporation Palo Alto Research Center, Palo Alto, CA 94304, 1992.

[162] R. Maki. Naming and locating the tops of rotated pictures. *Canadian Journal of Psychology*, 40(1):368–387, 1986.

[163] J. Malik and Z. Gigus. A model for curvilinear segregation. *Invest. Ophtalmol. Vis. Sci. (Supplement)*, 32(4):715, 1991.

[164] J. Malik and P. Perona. A computational model of texture perception. Report No. UCB-CSD 89-491, Computer science division (EECS), University of California, Berkeley, CA, 1989.

[165] J. Malik and P. Perona. Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America*, 7(5):923–932, 1990.

[166] P.A. Maragos and R.W. Schafer. Morphological skeleton representation and coding of binary images. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(5):1228–1238, 1986.

[167] G. Marola. On the detection of the axis of symmetry of symmetric and almost symmetric planar images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11(1):104–108, 1989.

[168] G. Marola. Using symmetry for detecting and locating objects in a picture. *Computer Vision, Graphics, and Image Processing*, 46:179–195, 1989.

[169] D. Marr. Analysis of occluding contour. *Proceedings of the Royal Society of London B*, 197:441–475, 1977.

[170] D. Marr. The structure and creation of visual representations. In Duane G. Albrecht, editor, *Recognition of pattern and form. Proceedings of a conference held at the University of Texas at Austin, 1979*, pages 59–87. Springer-Verlag, 1982.

[171] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W.H. Freeman and Company, New York, 1982.

[172] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London*, B(207):187–217, 1980.

[173] D. Marr and H.K. Nishihara. Representation and recognition of the spatial organization of three dimensional structure. *Proceedings of the Royal Society of London B*, 200:269–294, 1978.

[174] J.L. Marroquin. Human perception of structure. Master's thesis, Massachusetts Institute of Technology, 1976.

[175] G. Medioni and Y. Yasumoto. Corner detection and curve representations using cubic b-splines. *Computer Vision, Graphics, and Image Processing*, 39:267–278, 1987.

[176] C.B. Mervis and E. Rosch. Categorization of natural objects. *Ann. Rev. Psychol.*, 32:85–115, 1981.

[177] D.N. Metaxas. *Physics-based modeling of nonrigid objects for vision and graphics.* PhD thesis, Dept. of Computer Science, University of Toronto, 1992.

[178] F. Meyer and S. Beucher. Morphological segmentation. *Journal of Visual Communication and Image Representation*, 1(1):21–46, 1990.

[179] G.S.P. Miller. The motion dynamics of snakes and worms. *Computer Graphics*, 22(4):169–173, 1988.

[180] J. Montes, G. Cristóbal, and J. Bescós. Texture isolation by adaptive digital filtering. *Image and Vision Computing*, 6(3):189–192, 1988.

[181] M.C. Morrone and D.C. Burr. A model of human feature detection based on matched filters. In P. Dario and G. Sandini, editors, *Robots and biological systems*. Academic Press, 1990.

[182] M. Moshfeghi. Elastic matching of multimodality medical images. *Computer Vision, Graphics, and Image Processing*, 53(3):271–282, 1991.

[183] D. Mumford, S.M. Kosslyn, L.A. Hillger, and R.J. Herrnstein. Discriminating figure from ground: The role of edge detection and region growing. *Proceedings of the National Academy of Science*, 87:7354–7358, 1984.

[184] G.L. Murphy and E.J. Wisniewski. Categorizing objects in isolation and in scenes: What a superordinate is good for. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(4):572–586, 1989.

[185] L.R. Nackman and S.M. Pizer. Three-dimensional shape description using the symmetric axis transform i : Theory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-7(2):187–203, 1985.

[186] P.J. Nahin. Silhouette descriptor for image pre-processing and recognition. *Pattern Recognition*, 6:85–95, 1974.

[187] T.A. Nazir and J.K. O'Reagan. Some results on translation invariance in the human visual system. *Spatial Vision*, 5(2):81–100, 1990.

[188] R. Nevatia and T.O. Binford. Description and recognition of curved objects. *Artificial Intelligence*, 8:77–98, 1977.

[189] S.M. Newhall. Hidden cow puzzle picture. *American Journal of Psychology*, 65(110), 1954.

[190] J.A. Noble. Morphological feature detection. In *Proceedings of the Second International Conference on Computer Vision*, pages 112–116, Dec 1988.

[191] S. Palmer. Symmetry, transformation, and the structure of perceptual systems. In Jacob Beck, editor, *Organization and Representation in Perception*. Erlbaum, Hillsdale, NJ, 1982.

[192] S.E. Palmer. Structural aspects of visual similarity. *Mem. Cognition*, 6(2):91–97, 1978.

[193] S.E. Palmer. What makes triangles point: Local and global effects in configurations of ambiguous triangles. *Cognitive Psychology*, 12:285–305, 1980.

[194] S.E. Palmer. On goodness, gestalt, groups, and garner. *Paper presented at Annual Meeting of Psychonomic Society. San Diego, California. Unpublished manuscript*, 1983.

[195] S.E. Palmer. The role of symmetry in shape perception. *Acta Psychologica*, 59:67–90, 1985.

[196] S.E. Palmer. Reference frames in the perception of shape and orientation. In B.E. Shepp and S. Ballesteros, editors, *Object perception: Structure and process*, pages 121–163. Hillsdale, NJ: Lawrence Erlbaum Associates, 1989.

[197] S.E. Palmer and N.M. Bucher. Configural effects in perceived pointing of ambiguous triangles. *Journal of Experimental Psychology: Human Perception and Performance*, 7:88–114, 1981.

[198] S.E. Palmer, E. Simone, and P. Kube. Reference frame effects on shape perception in two versus three dimensions. *Perception*, 17:147–163, 1988.

[199] L.M. Parsons and S. Shimojo. Percieved spatial organitzation of cutaneous patterns on surfaces of the human body in various positions. *Journal of Experimental Psychology: Human Perception and Performance*, 13(3):488–504, 1987.

[200] T. Pavlidis. A vectorizer and feature extractor for document recognition. *Computer Vision, Graphics, and Image Processing*, 35:111–127, 1986.

[201] A. Pentland. Automatic extraction of deformable part models. Technical Report 104, Vision Sciences, Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1988.

[202] A. Pentland. Parallel part segmentation for object recognition. Technical Report 108, Vision Sciences, Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1988.

[203] P. Perona and J. Malik. Scale space and edge detection using anisotropic diffusion. In *Proceedings of a Workshop on Computer Vision, Published by IEEE Computer Society Press, Washington, DC.*, pages 16–22, Miami Beach, FL, 1987.

[204] P. Perona and J. Malik. Detecting and localizing edges composed of steps, peaks and roofs. In IEEE Computer Society, editor, *Third International Conference on Computer Vision*, pages 52–57, 1990.

[205] C.G. Perrott and L.G.C. Hamey. Object recognition: a survey of the literature. Technical Report 91-0065C, School of MPCE, Macquarie University, NSW 2109 Australia, 1991.

[206] E. Person and K.S. Fu. Shape discrimination using fourier descriptors. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-7:170–179, 1977.

[207] E. Persoon and K.S. Fu. Shape discrimination using fourier descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(3):388–397, 1986.

[208] G. Petter. Nuove ricerche sperimentali sulla totalizzazione perceptiva. *Rivista di psicologia*, 50:213–227, 1956.

[209] S.M. Pizer. Multiscale methods and the segmentation of medical images. Technical Report TR88-051, Dept. of Computer Science, Univ. of North Carolina, Chapel Hill, NC 27599-3175, 1988.

[210] S.M. Pizer, C.A. Burbeck, and J.M. Coggins. Object shape before boundary shape: scale-space medial axes. In Springer-Verlag, editor, *Proceedings of a NATO Workshop Shape in Picture*, 1993.

[211] S.M. Pizer, J.J. Koenderink, L.M. Lifshits, L. Helmink, and A.D.J. Kaasjager. An image description for object definition based on extremal regions in the stack. In *9th IPIMI Conference*, pages 24–37, Washington, DC, 1986.

[212] T. Poggio, E.B. Gamble, and J. Little. Parallel integration of vision modules. *Science*, 242:436–440, 1988.

[213] J. Ponce and M. Brady. Towards a surface primal sketch. A.I. Memo No. 824, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1985.

[214] P.B. Porter. Another picture puzzle. *American Journal of Psychology*, 67:550–551, 1954.

[215] P. Raghavan. *Randomized Algorithms*. Lecture Notes, 1990.

[216] V.S. Ramachandran. Capture of stereopsis and apparent motion by illusory contours. *Perception and Psychophysics*, 39(5):361–373, 1986.

[217] V.S. Ramachandran and S.M. Anstis. Displacement thresholds for coherent apparent motion in random dot-patterns. *Vision Research*, 23(12):1719–1724, 1983.

[218] K. Rao. Shape description from sparse and imperfect data. IRIS 250, Institute for Robotics and Intelligent Systems, University of Southern California, Los Angeles, CA, 1988.

[219] K. Rao and R. Nevatia. Descriptions of complex objects from incomplete and imperfect data. In *Proceedings Image Understanding Workshop*, pages 399–414, Palo Alto, CA, 1989. Morgan and Kaufman, San Mateo, CA.

[220] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Detection of interest points using symmetry. In IEEE Computer Society, editor, *Third International Conference on Computer Vision*, pages 62–65, 1990.

[221] I. Rentschler, M. Hubner, and T. Caelli. On the discrimination of compound gabor signals on textures. *Vision Research*, 28(2):279–291, 1988.

[222] G. Rhodes, S. Brennan, and S. Carey. Identification and ratings of caricatures: implications for mental representations of faces. *Cognitive Psychology*, 19:473–497, 1987.

[223] C.W. Richard Jr. and H. Hemami. Identification of three-dimensional objects using fourier dercriptors of the boundary curve. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-4:371–378, 1974.

[224] W. Richards. How to play twenty questions with nature and win. Technical Report AI Memo No. 660, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1982.

[225] W. Richards. Personal communication. 1989.

[226] W. Richards and L. Kaufman. "center-of-gravity" tendencies for fixations and flow patterns. *Perception and Psychophysics*, 5(2):81–84, 1969.

[227] L.C. Roberson, S.E. Palmer, and L.M. Gomez. Reference frames in mental rotation. *Journal of Experimental Psychology: Human Perception and Performance*, 13(3):368–379, 1987.

[228] I. Rock. *Orientation and form.* Academic Press, New York, 1973.

[229] I. Rock. *The logic of perception.* The MIT Press, Cambridge, MA, 1983.

[230] I. Rock. *Perception.* Sci. American Library, 1984.

[231] I. Rock and J. DiVita. A case of viewer-centered object perception. *Cognitive Psychology*, 19:280–293, 1987.

[232] I. Rock and A. Gilchrist. The conditions for the perception of the covering and uncovering of a line. *American Journal of Psychology*, 88:571–582, 1975.

[233] I. Rock and D. Gutman. The effect of in attention on form perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7:275–285, 1983.

[234] I. Rock, F. Halper, and T. Clayton. The perception and recognition of complex figures. *Cognitive Psychology*, 3:655–673, 1972.

[235] I. Rock and E. Sigman. Intelligence factors in the perception of form through a moving slit. *Perception*, 2:357–369, 1973.

[236] H. Rom and G. Medioni. Hierarchical decomposition and axial representation of 2d shape. Technical report, Institute for Robotics and Intelligent Systems, University of Southern California, Los Angeles, CA, 1991.

[237] E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[238] A. Rosenfeld and J.L. Pfaltz. Distance functions on digital pictures. *Pattern Recognition*, 1:33–61, 1968.

[239] E. Rubin. *Visuell Wahrgenommene Figuren.* Glydendalske, 1921. See [Boring 1964] Pg. 605, or [Rock 83] Pg. 306.

[240] J.M. Rubin and W.A. Richards. Colour vision and image intensities: When are changes material. Technical Report AIM–631, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1981.

[241] T. Ryanm and C. Schwartz. Speed of perception as a function of mode of representation. *American Journal of Psychology*, 69:60–69, 1956.

[242] T. Sanger. Stereo disparity computation using gabor filters. *Biological Cybernetics*, 59:405–418, 1988.

[243] E. Saund. The role of knowledge in visual shape representation. A.I. Technical Report No. 1092, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1988.

[244] B.G. Schunck. Edge detection with gaussian filters at multiple scales. In *Proceedings of a Workshop on Computer Vision, Published by IEEE Computer Society Press, Washington, DC.*, pages 208–210, Miami Beach, FL, 1987.

[245] G.L. Scott, S.C. Turner, and A. Zisserman. Using a mixed wave-diffusion process to elicit the symmetry set. *Image and Vision Computing*, 7(1):63–70, 1989. This paper appeared also in the Procedings of the 4th Alvey Vision Conference, Manchester, England, 1988.

[246] R. Sekuler and D. Nash. Speed of size scaling in human vision. *Psychonomic Science*, 27:93–94, 1972.

[247] J. Serra. *Image analysis and mathematical morphology.* Academic Press Inc., London, 1982.

[248] A. Sha'ashua and S. Ullman. Structural saliency: The detection of globally salient structures using a locally connected network. In *Proceedings of the Second International Conference on Computer Vision*, pages 321–327, 1988.

[249] A. Sha'ashua and S. Ullman. Grouping contours by iterated pairing network. In *NIPS*, 1990.

[250] R.N. Shepard. *Mind sights: Original visual illusions, ambiguities, and other anomalies, with a commentary on the play of mind in perception and art.* W.H. Freeman and Company, New York, 1990.

[251] R.N. Shepard and L.A. Cooper. *Mental Images and their Transformations.* The MIT Press, Cambridge, MA, 1982.

[252] R.N. Shepard and S. Hurwitz. Upward direction, mental rotation, and discrimination of left and right turns in maps. *Cognition*, 18:161–193, 1985.

[253] S. Shepard and D. Metzler. Mental rotation of three dimensional objects. *Science*, 171:701–703, 1971.

[254] S. Shepard and D. Metzler. Mental rotation: Effects of dimensionality of objects and type of task. *Journal of Experimental Psychology: Human Perception and Performance*, 14(1):3–11, 1988.

[255] A. Shimaya and I. Yoroizawa. Automatic creation of reasonable interpretations for complex line figures. In *Proceedings Int. Conf. on Pattern Recognition*, pages 480–484, Atlantic City, New Jersey, 1990.

[256] S.P. Shwartz. The perception of disoriented complex objects. In *Proceedings of the 3rd Conference on Cognitive Sciences*, pages 181–183, Berkeley, 1981.

[257] E.P. Simoncelli and E.H. Adelson. Non-separable extensions of quadrature mirror filters to multiple dimensions. Vision and Modeling Group Technical Report 119, Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[258] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, and D.J. Heeger. Shiftable multi-scale transforms. Vision and Modeling Group Technical Report 161, Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1991. Also in IEEE Transactions on Information Theory. Special Issue on Wavelets, 1992.

[259] A. Singh and M. Shneier. Grey level corner detection: A generalisation and a robust real time implementation. *Computer Vision, Graphics and Image Processing*, 51:54–69, 1990.

[260] J.P. Smith. *Vascular Plant Families.* Mad River Press Inc., Eureka, California, 1977.

[261] J.G. Snodgrass and M. Vanderwart. A standarized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Perception and Performance*, 6(2):174–215, 1980.

[262] F. Solina. Shape recovery and segmentation with deformable part models. Technical Report MS-CIS-87-111, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104-6389, 1987.

[263] H. Solomon. *Geometric Probability.* Society for industrial and applied mathematics, Philadelphia, Pennsylvania, 1978.

[264] K.A. Stevens. Surface perception from local analysis of texture and contour. Technical Report AI TR No. 512, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1980.

[265] T.M. Strat. *Natural object recognition.* PhD thesis, Dept. of Computer Science, Stanford University, 1990.

[266] J.B. Subirana-Vilanova. Curved inertia frames and the skeleton sketch: finding salient frames of reference. In *Proceedings of the Third International Conference on Computer Vision*, pages 702–708. IEEE Computer Society Press, 1990.

[267] J.B. Subirana-Vilanova. The skeleton sketch: finding salient frames of reference. In *Proceedings Image Understanding Workshop*, pages 399–414. Morgan and Kaufman, 1990.

[268] J.B. Subirana-Vilanova. Curved inertia frames: Perceptual organization and attention using convexity and symmetry. A.I. Memo No. 1137, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1991.

[269] J.B. Subirana-Vilanova. On contour texture. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 753–754, Ann Arbor, MI, 1991.

[270] J.B. Subirana-Vilanova. Frame curves and contour texture for non-rigid object recognition. In Springer-Verlag, editor, *Proceedings of a NATO Workshop Shape in Picture*, 1993.

[271] J.B. Subirana-Vilanova. Learning for contour texture and recognition. In *Proceedings of the III IMACS International Workshop on: Qualitative Reasoning and Decission Technologies*, pages 212–222, 1993.

[272] J.B. Subirana-Vilanova and S. Casadei. Parallel visual processing using random networks. In MIT Laboratory for Computer Science, editor, *Proceedings of the III M.I.T. Workshop on Supercomputing*, pages 44–45, Cambridge, MA, 1993.

[273] J.B. Subirana-Vilanova and W. Richards. Figure-ground in visual perception. In *The Association for Research in Vision and Ophthalmology. Annual Meeting Abstract Issue. Vol 32, NO. 4*, page 697, Bethesda, Maryland 20814-3928, 1991.

[274] J.B. Subirana-Vilanova and W. Richards. Perceptual organization, figure-ground, attention and saliency. A.I. Memo No. 1218, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1991.

[275] J.B. Subirana-Vilanova and K.K. Sung. Perceptual organization without edges. In *Proceedings Image Understanding Workshop*. Morgan and Kaufman, 1992.

[276] J.B. Subirana-Vilanova and K.K. Sung. Vector-ridge-detection for the perceptual organization without edges. A.I. Memo No. 1318, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, December 1992. See also proceedings ICCV-93.

[277] J.B. Subirana-Vilanova and K.K. Sung. Ridge detection for perceptual organization without edges. In *Proceedings of the Fourth International Conference on Computer Vision*, pages 57–65. IEEE Computer Society Press, 1993.

[278] K. Sung. A vector signal processing approach to color. AI-TR No. 1349, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992.

[279] M. Tarr and S. Pinker. Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21:233–282, 1989.

[280] D. Terzopoulos. Regularization of inverse problems involving discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(4):413–424, 1986.

[281] D. Terzopoulos and D. Metaxas. Dynamic 3d models with local and global deformations: Deformable superquadrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13:703–714, 1986.

[282] R.S. Thau. Illuminant precompensation for texture discrimination using filters. In *Proceedings Image Understanding Workshop*, pages 179–184, Pittsburgh, Pennsylvania, 1990. Morgan Kaufman Publishers Inc., San Mateo, CA.

[283] F. Tomita, Y. Shirai, and S. Tsuji. Description of textures by a structural analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-4(2):183–191, 1982.

[284] A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[285] Y.F. Tsao and K.S. Fu. Stochastic skeleton modeling of objects. *Computer Vision, Graphics, and Image Processing*, 23:348–370, 1984.

[286] H. Tuijil. Perceptual interpretation of complex line patterns. *Journal of Experimental Psychology: Human Perception and Performance*, 6(2), 1983.

[287] M. Turner. Texture discrimination by gabor functions. *Biological Cybernetics*, 55:71–82, 1986.

[288] S. Ullman. Filling in the gaps: The shape of subjective contours and a model for their generation. *Biological Cybernetics*, 25:1–6, 1976.

[289] S. Ullman. Visual routines. *Cognition*, 18, 1984.

[290] S. Ullman. Aligning pictorial descriptions: An approach to object recognition. *Cognition*, 32(3):193–254, 1989.

[291] C. Uras and A. Verri. Describing and recognising shape through size functions. In Springer-Verlag, editor, *Proceedings of a NATO Workshop Shape in Picture*, 1993.

[292] L.M. Vaina and S.D. Zlateva. The largest convex patches: A boundary-based method for obtaining parts. *Biological Cybernetics*, 62:225–236, 1990.

[293] S. Vallmitjana, J. Bertomeu, I. Juvells, S. Bosch, and J. Campos. Optical edge detection by a holographic filtering method. *Optics Communications*, 68(5):334–338, 1988.

[294] Peter J. Van Otterloo. *A contour-oriented approach to shape analysis*. Prentice Hall, UK, 1991.

[295] H. Voorhees and T. Poggio. Computing texture boundaries from images. *Nature*, 333(6171):364–367, 1988.

[296] S. Van Voorhis and S.A. Hillyard. Visual evoked potentials and selective attention to points in the scene. *Perception and Psychophysics*, 22(1):54–62, 1977.

[297] T.P. Wallace and P.A. Wintz. An efficient three-dimensional aircraft recognition algorithm using normalized fourier dercriptors. *Computer Vision, Graphics, and Image Processing*, 13:99–126, 1980.

[298] D.L. Waltz. Understanding line drawings of scenes with shadows. In P. Winston, editor, *The Psychology of Computer Vision.* McGraw-Hill, New York, 1972.

[299] W.M. Wells-III. Statistical object recognition. Technical Report AI-TR-1398, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1993.

[300] M. Wertheimer. Principles of perceptual organization. In B. Beardslee and M. Wertheimer, editors, *Readings in perception.* Van Nostrand, Princeton, 1958. Originally published in 1923.

[301] M.A. Wiser. *The role of intrinsic axes in the mental representation of shapes.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1980. See also the *Proceedings of 3rd Conference on Cognitive Sciences, Berkeley*, pp. 184-186, 1981.

[302] Andrew P. Witkin. Scale-space filtering. In *Proceedings IJCAI*, pages 1019–1022, 1983.

[303] A.P. Witkin. Shape from contour. Technical Report AI TR No. 589, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1980.

[304] A.P. Witkin. Scale space filtering: a new approach to multi-scale description. In *Proceedings Image Understanding Workshop*, pages 79–95, Pittsburgh, Pennsylvania, 1984. Morgan Kaufman Publishers Inc., San Mateo, CA.

[305] A.P. Witkin and J.M. Tenenbaum. On the role of structure in vision. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human and Machine Vision.* Academic Press, New York, 1983.

[306] G. Wolberg. Image warping among arbitrary planar shapes. Technical Report CUCS-329-88, Dept. of Computer Science, Columbia University, New York, NY 10027, 1988.

[307] A.L. Yarbus. *Eye movements and Vision.* Plenum, New York, 1967.

[308] A. Yuille, D. Cohen, and P. Hallinan. Facial feature extraction by deformable templates. Technical Report CICS-P-124, Center for Intelligent Control Systems, 1989.

[309] C.T. Zahn and R.Z. Roskies. Fourier descriptors for plane closed curves. *IEEE Computer*, 21(3), 1972.

[310] J. Zerubia and D. Geiger. Image segmentation using 4 direction line-processes and winner-take-all. In T. Kohonen, K. Mákisara, O. Simula, and J. Kangas, editors, *Artificial Neural Networks*, pages 1083–1092. Elsevier Science Publishers B.V., 1991.

[311] S. Zhong and S. Mallat. Compact image representation from multiscale edges. In IEEE Computer Society, editor, *Third International Conference on Computer Vision*, pages 406–415, 1990.

[312] X. Zhuang, T.S. Huang, and H.H. Chen. Multi-scale edge detector using gaussian filtering. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 558–563, 1986.

[313] S.W. Zucker, A. Dobbins, and L. Iverson. Two stages of curve detection suggest two styles of visual computation. *Neural Computation*, 1(1):68–81, 1989.