Chapter 1 Introduction

In essence, the goal of computer vision research is to formulate a computational method that derives scene description in terms of geometry, motion, appearance, lighting, and object identity from one or more two-dimensional arrays of pixels. Mathematically, due to the projective nature of imaging, the problem is underconstrained as many different scene configurations can produce the same image (figure 1.1), even though, most of the time, only one of these is perceived by the human visual system. In tackling this underconstrained problem, we often seek inspiration from the human vision system. Among others, the phenomenon of illusory figures, as illustrated in figure 1.2, suggests that our vision system tends to interpret two-dimensional (2-D) visual inputs as the projections of three-dimensional (3-D) scenes [44]. By observing our 3-D, physical world, it is possible to derive useful properties, and infer constraints on this 3-D world. The challenge lies in translating these world constraints into constraints applicable at the image level.

Figure 1.1 Possible scene configurations
Perhaps the most commonly used constraint is continuity, which is derived from the "Matter is cohesive" property, as stated by Marr [53]. Numerous techniques have been proposed to translate this constraint into image terms to solve a wide range of computer vision problems, such as perceptual grouping, image segmentation, shape from stereo, shape from shading, image sequence analysis, etc. Despite the diversity of existing techniques, most machine vision solutions to date lack the robustness necessary for use in unsupervised situations. Initialization and parameter tuning are often required. Many algorithms are iterative and produce results that are sensitive to parameter settings.

In this thesis, the goal is to devise a robust methodology for applying the continuity constraint to solve a variety of early vision problems. The approach taken to tackle this fundamental machine vision problem is novel and unconventional:

- a local representation is used to encode global structures such as curves, surfaces, regions and junctions, and their perceived importance (called perceptual saliency), leading to a layered representation of visible curves and surfaces in the scene;
- a consensus-based technique is applied to infer salient features, detect discontinuities, and identify outliers simultaneously;

Figure 1.2 2-D drawing interpreted as 3-D overlapping surfaces and curves
The methodology is grounded in two elements, namely, *tensor calculus* for data representation and non-linear *voting* for data communication, which together provide a unified framework for the robust inference of multiple curves, surfaces, regions and junctions from any combination of points, segments, and surface patch elements. In order to evaluate the efficiency and effectiveness of the method, algorithms have been developed based on the proposed framework to address a number of early vision problems, including perceptual grouping in 2-D and 3-D, shape from stereo, shape from shading, and motion grouping and segmentation.

In this introductory chapter, the motivation behind this thesis is further elaborated in section 1.1. Section 1.2 outlines the proposed approach. Section 1.3 summarizes the methodology and the scope of this thesis. The contributions of the thesis are highlighted in section 1.4.

1.1 Motivation and Goals

1.1.1 The Problem

Despite tremendous progress in vision research and computer technology in the past few decades, the applicability of computer vision remains very limited. The difficulties in automating visual perception stem from the fact that images are 2-D projections of the 3-D world, whose appearance does not only depend on the viewpoint and the geometry of the objects in the scene, but is also determined by other physical properties such as surface reflectance, scene illumination, and imaging sensor characteristics. In other words, machine vision researchers are tackling the difficult problem of finding the inverse of a non-linear function with multiple variables.
While the values associated with an individual pixel are determined by a combination of scene properties and sensor characteristics, the fact that matter is cohesive provides a very strong constraint for recovering these properties and characteristics. In general, the variation of pixel values in a local neighborhood is small. An abrupt intensity change always corresponds to discontinuities of some scene properties. This relation between image and scene properties has been the premise of computational vision from the beginning. When devising solutions to many vision problems, we often face the problem of identifying perceptually important (or salient) and structured information from a noisy data set. From greyscale images, edges are extracted by first detecting changes in intensity, and then linking locations based on the noisy signal responses. From binocular images, surfaces are inferred by first obtaining depth hypotheses for points and/or edges using local correlation measures and then selecting and interpolating appropriate values for all points in the images. Similarly, in image sequence analysis, the estimation of motion and shape starts with local measurements of feature correspondences that give noisy data for the subsequence computation of scene information. As computer vision systems move from controlled laboratory settings to real applications, the need for robust salient structure inference becomes more apparent.

1.1.2 General approaches in computer vision

To overcome the complexity of the computer vision problem, researchers have employed the traditional approach of divide-and-conquer since the early days. Simpler versions of various vision problems have been addressed, ranging from object recognition from scenes with single polygonal object to shape reconstruction from 3-D range data. Most existing solutions take the modular approach, tackling various issues such as matching, interpolation, and handling of noise and outliers independently. In other words, as stated by Marr, "to the Desirable via the Possible" has always been the guid-
The representational framework presented by Marr in [53] (figure 1.3) not only summarizes early efforts towards a unified approach for deriving shape information from images, but also becomes the standard paradigm for solving machine vision problems.

At the heart of Marr’s representational framework is the intermediate representation called the $2\frac{1}{2}$-D sketch, which is a viewer-center description of the visible surfaces. It serves as the main stepping stone toward the recovery of three-dimensional, object-centered description of object shapes and their spatial organization from images. This simplified representation, together with the modular approach to problem solving seems to provide a handle to solve the difficult task of deriving scene description from images. In particular, many aspects of various vision problems can be formulated in the standard functional optimization framework, which can then be solved using well-known mathematical techniques. Accordingly, the main focus of computer vision research has been...
on finding the "right" functional to optimize for each aspect of the vision problem. Complicated situations are handled by incorporating more optimizing criteria into the framework.

1.1.3 Common limitations of current methods

Although the use of the divide-and-conquer approach seems appropriate, the results produced so far are less than ideal. Since imaging is a non-linear process, any solution to the general vision problem should reflect this non-linearity. In the functional optimization framework, this is translated as the optimization of a non-linear function. Unfortunately, all numerical solutions to non-linear functional optimization require initialization and iterations. Moreover, in order to address several issues simultaneously, it is often necessary to combine multiple optimizing criteria. Current approaches either combine different criteria into one optimizing function, or optimize each criterion independently and solve the problem iteratively. In either case, parameters are introduced to weigh the relative importance of each criterion. Usually, the setting of these parameters is not intuitive. The fact that many methods produce parameter-dependent solutions aggravates the situation.

On the other hand, the use of the $2^{1/2}$D sketch has also complicated the matter. As it is a viewer-centered representation, viewpoint dependent elements are unnecessarily introduced into the framework. For instance, in perceptual grouping and image segmentation, each pixel is associated with a partition of the image (figure 1.4). Hence, any single visible surface under occlusion is represented by several regions of the partition. While the use of viewpoint dependent representation is arguably acceptable when dealing with individual images, such a representation, in handling multiple images, is harmful. For instance, in stereo, a depth value is computed for every pixel (figure 1.5).
Therefore, discontinuities occur at the boundaries of overlapping visible surfaces. Since these discontinuities in depth do not correspond to any physical property of a 3-D object, viewpoint-dependent constraints are needed to derive the solution, leading to more complicated and less stable algorithms that give results containing unnecessary information. In motion grouping and segmentation, an optical flow vector is estimated for each pixel in the image (figure 1.6). Again, discontinuities occur at the boundaries where visible surfaces overlap. Similarly, more constraints are needed to resolve these viewpoint-dependent issues, whereas the results obtained are often the fragmented version of the desired one.

![Figure 1.4 Partitioning of a Greyscale Image](image)

**Figure 1.4 Partitioning of a Greyscale Image**

### 1.1.4 Desirable solutions

Based upon the performance of the existing methods, we argue that the standard paradigm from which these methods are derived needs some adjustment. In particular, the fine-grain divide-and-conquer approach to problem solving only provides partial solutions to the vision problem. As the cost of integrating individual solutions is so high, only limited progress has been made to solving the general vision problem. On the other
Figure 1.5 2\(^{1/2}\)-D sketch for stereo images

(a) Input images

(b) The "wedding cake" description of the scene

*Figure 1.5 2\(^{1/2}\)-D sketch for stereo images*
Figure 1.6  $2^{1/2}$-D sketch for motion segmentation

(a) Input images

(b) Segmentation into regions
hand, since the use of the $2^{1/2}$-D sketch as intermediate representation unnecessarily introduces viewpoint dependent elements, it not only doesn’t facilitate, but in fact hinders, the recovery of visible curves and surfaces from images. Hence, desirable solutions to the general computer vision problem are unlikely to be derived from the current paradigm.

While it is almost certainly impossible to have machine vision deliver a completely invariant scene description from an image in only one step, as Marr [53] said, we believe the vision problem cannot be solved by tackling all individual issues independently either. Instead, the divide-and-conquer strategy should be applied up to the point where each individual sub-problem remains self-contained. For instance, the handling of outliers and noise should not be considered to be independent of the inference of salient structures such as curves, surfaces, and regions. On the other hand, viewpoint dependent information and object dependent information should not be mixed and represented together.

We argue that in deriving desirable solutions to computer vision problems, the enforcement of each constraint, such as the continuity constraint and the uniqueness constraint, should be considered as a single task. In other words, it is necessary and sufficient to enforce each constraint individually, but all aspects of the constraint must be addressed simultaneously. This is sufficient because different constraints encode different imaging or object properties. This is necessary because different aspects of a constraint are tightly coupled and cannot be addressed independently. For example, salient features, outliers and noise capture different aspects of continuity. Furthermore, we argue that the shape continuity constraint is the strongest one as it determines the perceived importance, or saliency, of each measurement. Other constraints, such as
uniqueness and opacity constraints, are combinatorial and hence are less discriminative as they often can be satisfied by more than one instance.

Accordingly, we reason that the desirable intermediate representation should be an object-centered description of the visible curves and surfaces. The use of a viewpoint independent representation is vital to the proper implementation of constraints, as each constraint is derived either from an imaging characteristic or an object property. Specifically, we believe a layered representation of visible curves and surfaces should be used instead. The layered representations of the scenes illustrated in figures 1.4, 1.5, and 1.6 are shown in figures 1.7, 1.8, and 1.9 respectively. By using such a representation, we avoid overloading the representation with both shape and viewpoint information. The advantage of using a layered representation has been noticed recently, mostly in tackling the image sequence analysis problem [1, 11, 12, 41, 85, 86].

The above discussion of the issues, in a sense, defines the goal of this thesis. More precisely, this thesis addresses the problem of deriving an integrated solution for applying the continuity constraint to infer a layered description of visible curves and surfaces of the scene from one or more images. The emphasis is on the robustness, generality and efficiency of the solution. To determine the effectiveness of the method, this thesis also attempts to tackle a number of early vision problems using the proposed methodology. In particular, solutions for perceptual grouping, shape from shading, shape from stereo and motion grouping and segmentation are developed.
Figure 1.7 Layered Representation of Visible Surfaces - Image Segmentation

(a) input image

(b) layered description in terms of curves, regions, and corners
Figure 1.8 Layered Representation of Visible Surfaces - Stereo
Figure 1.9  Layered Representation of Visible Surfaces - Motion Segmentation
1.2 Our Approach

From the previous discussion about the general vision problem, two broad issues need to be addressed in order to develop robust methods for inferring visible curves and surfaces:

- What is an appropriate representation for a layered description of visible curves and surfaces?
- How to compute a layered description of visible curves and surfaces from one or more images?

The first issue is related to a representation of visible curves and surfaces that is general enough to account for all possible visible curves and surfaces and discontinuities, and yet specific enough to be encoded computationally. The second issue is related to the algorithmic aspect of the problem, where the task is to robustly infer a layered description of visible curves and surfaces from images.

1.2.1 Data Representation

There are two ways to represent visible curves and surfaces. Global representations use parametric functions to capture the shape of curves or surfaces where as local representations describe curves and surfaces by specifying their local geometric properties. While global representations can provide a more abstract description of the scene, the best parametric description of curves or surfaces is hard to obtain computationally. The situation is also complicated by the presence of curve or surface discontinuities. As demonstrated by numerous attempts to infer curves and surfaces by explicit functional optimization, parametric model misfits are often indistinguishable from errors caused by outliers and curve or surface discontinuities. On the other hand, local representations are more general as they describe different shapes in a uniform matter. We hence pro-
pose to use a local representation to describe layers of visible curves and surfaces. We believe that once the local properties of the curves and surfaces are obtained, the derivation of the global curve and surface representation will be simplified.

To determine the appropriateness of a representation, we need to devise some evaluation criteria from the property of the desirable description. Based upon the "Matter is cohesive" property, a visible curve or surface can be characterized geometrically by the fact that it is smooth almost everywhere, except at locations where discontinuities occur. Therefore, an appropriate local representation should be able to encode the complementary properties of smoothness and discontinuity properly.

In summary, we seek to devise a representation that encodes at every location in the domain space the following information:

1. The saliency of having a geometric feature occur at the location
2. The geometric property of the feature if one occurs at the location

The geometric features that are relevant here are curve segments, surface patches, region boundaries, and curve, surface or region boundary discontinuities. The geometric properties that need to be encoded are tangents for curves, surface normals for surface patches and tangents for region boundary. Note that curve orientation, surface orientation and region boundary are three separable aspects of curves and bounded surfaces, and therefore can be encoded by three different representations. Hence, it is only necessary to devise a representation that encodes a feature, either as an orientation or a discontinuity, and the saliency of the feature.

While it is intuitive to use a unit length vector to represent an orientation, there is no obvious choice for representing discontinuity. In fact, hardly any attempt was made
to derive a local representation for discontinuity. In our case, the challenge stems from
the requirement that the data representation for discontinuity must be compatible with
that for an orientation, so that they can be combined into one. Interestingly, it is this
seemingly tough requirement that leads us to devise a simple, yet effective, representa-
tion for visible surfaces.

The derivation of our curve and surface representation stems from the observation
that discontinuities occur at locations where multiple salient structures such as curves,
surfaces or region boundaries meet. Therefore, curve orientation discontinuities are sig-
nified by the presence of multiple curve segments. Surface orientation discontinuities
are signified by the presence of multiple partial surface patches. A similar argument ap-
plies to region boundary discontinuities. In other words, whereas there is only one ori-
ettation associated with a location within a smooth curve segment, or a surface patch
or a region boundary, there are multiple orientations associated with any location with
a discontinuity. Hence, the desirable data representation is the one that can encode one
or more orientations. It turns out that a 2nd-order symmetric tensor possesses precisely
this property. Indeed, the tensor formalism was first developed for capturing variations
of orientations in the study of fluid dynamics. Recently, Knutsson [48] and Westin [87]
have used 2nd-order tensor as data representation to solve a number of signal process-
ing problems in computer vision and have obtained promising results. Figure 1.10 de-
picts a geometric illustration of a 2nd-order symmetric tensor. We use the shape of the
tensor to encode uncertainty of orientation, and the size of the tensor to encode feature
saliency.

Note that due to symmetry, the 2nd-order tensor representation of an orientation \([x
y z]^t\) is the same as that of the orientation \([-x -y -z]^t\). To capture the polarities of orien-
tations, which is a 1st-order orientation information, we make use of the 1st-order ten-
sors, which are commonly known as vectors. Our data representation hence includes a
1st-order tensor for encoding polarity saliency, and a 2nd-order tensor for encoding ori-
entation saliency.

1.2.2 Computational Methodology

Having decided to use tensors to represent visible curves and surfaces locally, we es-
sentially eliminate the possibility of using any popular scalar functional optimization
techniques. Instead, we need to explore other methods so as to find one that is compat-
ible with the use of the tensorial representation.

Recall that we use tensors to capture the variations of orientations. We hence argue
that statistical methods are among the best choices. In particular, for each location, we
proposed to collect a large number of orientation estimations obtained by fitting simple
curve/surface/boundary model in a local neighborhood. By analyzing the consistency
of the orientation estimations and the amount of support, we can determine the feature
type and the saliency of the feature simultaneously. This non-linear voting technique is
related to the Hough Transform, and was first proposed by Guy and Medioni [26, 28,
29] for perceptual grouping and surface reconstruction. They have developed a vector
voting technique that infers, from a sparse set of data, multiple curves in 2-D and surfaces in 3-D. While they have obtained a number of impressive results, the use of vectors has limited the development of their methodology. Some ad-hoc measures were devised to select the relevant first order orientation information from the second order orientation information collected. By replacing the data representation with tensors, we extend their methodology significantly into a unified framework for the inference of salient features in many early vision problems.

1.3 Overview of the Proposed Method

An overall illustration of our method, summarizing its different components, is shown in figure 1.11. The methodology is grounded in two elements: tensor calculus for data representation, and non-linear voting for data communication. Each input site propagates its information in a neighborhood. The information is encoded in a tensor, and is determined by a predefined voting field. Each site collects the information cast there and analyzes it, forming a saliency map for each feature type. Salient features are located at local extrema of these saliency maps, which can be extracted by non-maximal suppression.

The use of tensor voting provides a unified framework for salient feature inference, in which input data are encoded in a sparse tensor field, inference rules are encoded in a dense tensor field, and the output are represented by a dense tensor field, as outlined in figure 1.12. The method is non-iterative, requires no initialization and threshold, and the only parameter is scale. Results obtained are associated with saliency values, which can be used as evaluation criteria when applying other constraints.
Figure 1.11 Salient Feature Inference Engine

Figure 1.12 The Unified Framework of Tensor Voting
The methodology discussed in this thesis has been implemented and applied to tackle a number of early vision problems, as presented in chapter 5. Two examples of the results produced by our method on the perceptual grouping and shape from stereo are given in figure 1.13 and figure 1.14 respectively. The figures show the resulting descriptions consisting of overlapping layers of surfaces or curves together with the inferred junctions.

1.4 Contribution of this Thesis

This thesis addresses the important problem of deriving a scene description in terms of visible surfaces from one or more images. The emphasis is on the robustness, generality and efficiency of the solution. The contributions of this thesis are the following:

- Identification of the continuity constraint as the most basic one in solving machine vision problems.

- Development of a tensorial representation for describing curve, surface or boundary orientations, discontinuities, and outlier simultaneously.

- Development of a tensor voting technique for the proper enforcement of the continuity constraint to infer multiple curves, surfaces, regions and junctions from noisy, irregularly clustered data set.

- Development of algorithms for applying tensor voting to solve a number of early vision problems, including curve and region inference in perceptual grouping, shape from shading, shape from stereo, and image sequence analysis.
Figure 1.13 Perceptual Grouping from dots

(a) input

(b) detected salient structures

region
curves
labeled junctions
Figure 1.14 Result on Shape from Stereo

(a) input images

(b) inferred surfaces and junctions in disparity space

(c) a rectified view of the inferred salient structures

*Figure 1.14 Result on Shape from Stereo*
1.5 Outline and Notations

In this chapter, we have presented the definition and the motivation of the problem and have given an overview of the proposed approach and its implementation. In chapter 2, a more elaborate description of related research in the literature is given. The discussion focuses on general computational approaches for early vision. Individual methods are only cited as references. In chapter 3, we present the elements of our salient feature inference engine. In chapter 4, the inference of curves, surfaces, and regions using the salient feature inference engine is illustrated. In chapter 5, we derive solutions based upon this tensorial framework for solving a number of early vision problems. In chapter 6, we show how salient feature inference can also be used in a few real applications. Chapter 7 concludes this thesis and discusses future research issues.

In this thesis, scalars are denoted by italic letters, e.g. $l$, tensors are denoted by bold capital letters, e.g. $T$, vectors are denoted by bold lower-case letters, e.g. $e$. Unit length vector for $e$ is denoted as $\hat{e}$.