Perception of Shape Properties from Multiple Cues

by

Massimiliano Di Luca

Master, Psychology, Universitá di Trieste (IT), 2000

A dissertation submitted in partial fulfillment of the

requirements for the Degree of Doctor of Philosophy

in the Department of Cognitive and Linguistic Sciences at Brown University

Providence, Rhode Island

August 2006

Abstract of "Perception of Shape Properties from Multiple Cues" by Massimiliano Di Luca, Ph.D., Brown University, August 2006.

The world appears three-dimensional (3D) even though the depth dimension is lost with projection to the retina. The visual system uses different cues that carry information about the 3D aspects of the world, combining the information they convey. Most research is based on the following assumptions: cues specify depth information, they are processed independently, and they combine linearly in order to provide a single depth-map. I present data showing that the visual system does not rely on these assumptions. Cues are informative about only some aspects of the 3D shape of objects. Some cues specify depth, while others carry information about surface orientation, curvature or local shape. My hypothesis is that cues are combined independently for each of the 3D properties, and that the computation is not derived from any unified representation.

I asked participants to make judgments about monocularly viewed computer-generated convex shapes. Participants compared two of these shapes in terms of the magnitude of one 3D property: depth, curvature, or orientation at a given point. One surface was kept constant while the shape of the other was either varied between trials or it was dynamically modified by the participants.

Results indicated that even when shapes defined by either motion, texture, or shading were perceived as having the same curvature, they were not necessarily perceived as having the same depth or orientation at specified points. Three-Dimensional shapes reconstructed from the judgment of different shape properties were significantly different from one another. Since cues carry different information about these 3D properties, I conclude that they must be represented independently. Since properties estimated in single-cue stimuli are predictive of the same property in cue-combined stimuli, cue combination must be independent for each property. I propose a new approach to cue combination that accounts for all of the observed differences.

Acknowledgements

Only when you close an important phase of your life do you realize how many people love you. As I leave for Germany, I realize how much I will miss these people and I am deeply grateful for their support and help. Many, many faces and memories are popping up in my mind at this moment, and I am sorry if I cannot remember everybody, but I will do my best.

My first thanks go to my advisor, Fulvio Domini. His abilities, enthusiasm, and support made all the difference in my academic career. He gave me the opportunity to come to Brown. His dedicated work helped me finish the program. I hope we will continue working together on new projects and hope to someday publish these and the other results we have collected during these years.

Many thanks also go to the other two members of my dissertation committee, Michael Tarr and William Warren. They went well beyond their duties in helping me with this dissertation. Frequently, I interacted pleasantly with them both personally and academically. I have learned a lot by their examples and I will continue to admire their skills and professionalism even when far away.

I would like to thank my Italian supporters: friends and colleagues from Trieste who helped me in the preparation of this work. First, I am grateful to Corrado Caudek, who served as my second advisor in undergraduate studies and continues to deserve that title. I hope he will somehow be my second advisor in the future as well. Unfortunately, for logistical reasons, I had to excuse him from my thesis committee. I look forward to future occasions for us to work together. The other Italian supporters are Carlo Fantoni and Sara Rigutti. We had many chances to chat about this thesis and I thank them for their suggestions and friendship.

Special thanks go to Daniel Rothman, who, as he pointed out during my defense, appeared as the first name without a PhD. This now recurs in the acknowledgments. Daniel was once my splendid Research Assistant but has become my motivator. Without him, I probably would not have finished this dissertation in this decade. He provided me with resources in the form of participants, scheduling, and assistants from the VenLab. He also reviewed this manuscript with dedicated care. But most importantly he is a good friend. I wish him all the best as he starts his academic career so that among other things, he can have a PhD as well.

I also appreciate Jon Shotola, whom I met (officially) when I needed a Teaching Assistant. He proved to be irreplaceable both as a teacher and as a friend. He frequently helped me write my proposal and later, accompanied by Kathleen, we digest it at Tortilla Flats.

And since I mention the Flats, I want to thank with all my heart Kathryn Good,

who is more than a friend, and has been like family for my wife and me during our sojourn in Providence. She helped us to cope with American culture and the English language and guided us deftly through our personal challenges.

I thank Samir Fatima, a helpful friend who recently became my neighbor. He provided encouragement, good movies and beer.

I am also grateful to Mona Delgado, who helped my wife and me when I first arrived and now helps once again. In the period of frantic confusion preceding commencement, I would not have been able to safely navigate Brown's administrative shoals, which so much resemble the labyrinthine bureaucracies of Italy.

I would like to thank the members of the Brown's academic community, particularly Sheila Blumstein, for whom I had the honor to work as a Teaching Assistant. She is not only a thoughtful teacher and researcher, but also a wonderful person, infinitely capable of listening and counseling. Her brightness and true empathy will forever remain among my cherished memories. I also wish to acknowledge Steven Sloman, whom I thank for the friendly encouragement during my five years in Providence. Only now do I realize that at one particular moment I would have abandoned the program without his support.

My gratitude goes also to the current and former members of the 3-D Shape Perception Laboratory, especially Erika Scilipoti, Rajesh Shah, Hadley Tassinari and Quoc Vuong, the last of whom will remain my colleague at the Max Planck Institute. The encouragement of other colleagues and friends over the years has been invaluable and they deserve my heartfelt thanks: Joon-Yoon and Evan Kim, Sharon Goldwater, Justin and Valerie Owens, Jonathan Cohen, Joanna Tai, Emily Myers and Paul Allopenna, Wendy Adams, Chun-Cha Kung, Massimo Ciaramita, Melissa and Tim Bud, Socrates Dimitriadis, Elena Tenembaum, and Hamid-Reza Sarajian.

My gratitude extends to my fellow Italian graduate students, who shared in my pain and passion: Simone, Mario, Mauro, Roberto, Stefano, Monica, Marco, Evelyn, Massimo, and Antonella. I also must thank the Holy Ghost Church and the Italian community of Federal Hill.

Special thanks go to my family: my grandparents, both of whom sadly passed shortly after my arrival in Providence, my parents, and my sister.

Ultimately, I am deeply grateful to my beloved wife Chiara. She made all the difference. She supported, encouraged and took good care of me. She was courageous enough to come to a foreign country and learn a foreign language for me, and now she will do it again. She walked beside me every step on the road to this dissertation. I could not have done it without her. This work is hers as much as it is mine.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> “The education of our space-perception consists largely of two processes: reducing the various sense-feelings to a common measure and adding them together into the single all-including space of the real world” James (1890, Vol 2 p 268-9)

On my desk there is an apple, a green shiny apple with some yellow spots. It is illuminated by the light streaming in from the window and has the prototypical shape of an apple, the shape that you expect an apple to be. Although its overall shape reminds one of a sphere, the apple is more round at the top and shrinks toward the bottom, the big concavity on the top deforms the sphere in a substantial way, and there are many bumps distributed mostly on the two rims at the top and bottom. I can naively say that I can see the three-dimensional shape of the apple very easily. Although effortless, 3D shape perception is a remarkable and complex ability of the human visual system. The visual system processes the information in the optic array captured by the two retinae and estimates the aspect of the world lost in optical projection[1].

The classical approach to the perception of 3D shape is based on the laws of inverse optics; it aims to reconstruct the Euclidean shape of the apple. In this approach, the visual system picks up “depth-cues” in the retinal image [2] and analyzes them in different modules. For the apple on my desk, there are many sources of information (for a complete list see Cutting & Vishton, 1995), including the stereoscopic disparity of the

[1] The term estimate refers to the result of the computation of a scene property

[2] For reasons that should be clear to the reader the term “cue” will be used in the rest of the text.

little spots, their relative velocity as I move my head, their shape and distribution on the retina, the shape of contours, the different luminance, the position and shape of highlights. Each of the depth-cues provides slightly different information about shape and according to the classical approach the difference is due only to the amount of noise in the pick-up; it is implicitly assumed that cues provide the same type of information about shape, the relative depth of its parts. Each cue is analyzed in isolation by a module that computes an estimate of shape in the form of a "depth map". So, because each cue has the same informative power, it is analyzed in a modular way, and it is used to recover the Euclidean shape of objects, this way of analyzing the information is called "shape-from-x" (Bülthoff, 1991). Subsequently, all depth maps are combined using a weighted average according to the reliability of cues in order to reduce the noise in the measurement and achieve a unique representation of shape. Other geometric properties can be computed from this single representation. In this framework, the visual system's task is to reconstruct the metric shape of the object, that is to obtain a veridical perception of the Euclidean shape.

There are findings which challenge many of the assumptions that the classical approach makes and in this work I will provide further evidence against the approach. My key observation is that cues are not equally informative about different aspects of the shape of an object. Each cue is produced by some particular geometrical aspect of the scene and therefore is informative about this property, but it might not be informative for other ones. As a result, a given property is not equally specified by cues, and computations based on any one cue may not be equivalent for recovering different shape properties.

I believe that the key to understanding shape perception is the analysis of the information conveyed by the cues. Researchers have failed to identify the relevant information that the visual system uses to estimate shape. Most of the initial research attempted to expand the number of identified sources of information. As a result, the number of identified cues increased but without actually increasing the understanding of the problem. At a certain point, as Epstein (in Cutting & Vishton, 1994) said, "the most curious fact about psychological approaches to the study of layout is that its history is little more than a plenum of lists."

In this work, I will analyze a short list of cues too, but I will not only attempt to create a taxonomic distinction. I propose a new way of looking at cues in order to understand how they are combined and how the visual information is represented in the brain. If this approach is proven to be correct, not only would lists of this kind be useful, they would also become necessary to understand cue combination and shape perception. Previous works of this type were limited in two ways (an exception is Bülthoff & Mallot, 1990): they analyzed one cue (i.e. Todd & Mingolla, 1983), and they proved either the inferiority of one property (Koenderink, Doorn, & Kappers, 1996) or the superiority of a second one (i.e. A. Johnston & Passmore, 1994c). This work not only shows the relative influence of cues on three of geometric properties that have been widely investigated in the literature, for the first time provides an explanation of the human performance based on a single mechanism.

Studies involving shape cues have often led to conflicting results. One application of the present approach is a reformulation of the apparently incoherent and contradictory results on cue combination. Many of the observed differences have been attributed to dedicated rules of interaction for each cue pair (e.g. Bülthoff & Mallot, 1988, 1990). I propose, instead a simple general mechanism to combine cues that can account for the different results. This mechanism simply differentiates cues depending on the informational contribution of the cues in isolation. Once this factor has been accounted for, all the cues interact lawfully as we will see below.

This work is organized as follows [3]. In Chapter 2, I will analyze the informational contribution of three cues: motion, texture and shading. I will describe how the image signals are generated and their mathematical relation to different geometric properties of shape. In particular, the shape properties I will consider include the relative depth of different points on the surface, local surface orientation and local curvature. For each cue, a description of the most common theoretical approaches to the problem will be provided along with a summary of the psychophysical investigations that involve perceptual estimates of the properties mentioned. The emphasis will be kept on the informational limitations of each cue.

Chapter 3 summarizes the literature on cue combination. First, it will be shown that the classical approach to the problem of cue combination is based on the estimate

[3] Some sections are preceded by a quote; it does not serve any purpose except my and perhaps your enjoyment.

of depth maps. I will discuss also why this assumption is needed for the classical models to work and what its implications are. Then, a new approach to cue combination, the Intrinsic Constraint model, will be introduced as an example of a non-Euclidean model of perception that does not need such an assumption. This model will serve as a tool for understanding the experimental results.

In Chapter 4, I will use the preceding analysis of cues to simulate an optimal observer. The goal of this simulation is to support the idea that cues have an informational contribution that depends on the shape property estimated.

In Chapters 5 through 8, I will describe a series of experiments aimed to demonstrate that the visual system does not consider these cues to be equivalent in specifying different aspects of shape. The effect that the cues described have on perceiving different properties of shape will be explored experimentally. Participants viewed computer-generated shapes created using motion, texture, or shading cues either in isolation or combined. The shapes were first matched in terms of one property and then compared for other properties. The judgment was based on three different aspect of the shapes: curvature, orientation, or slant. The task given to the participant was to either choose which of two shapes had the greater magnitude of the property or to modify the shapes so that the properties were perceived as equal. The goal of these experiments is to demonstrate task significantly influences the responses of the participants. The results show that the shape from which the property were judged depended on the cues in the stimulus and the task given to the participant. Changing the task changes the properties judged on the same visual information. The shape from which the judgments appear to be based on changes considerably with this modification. These results are inconsistent with the existence of a unique Euclidean representation of perceived shape. I propose that cues specify different properties depending on the type of judgment required and therefore cue combination is determined independently for each property by the relative contribution of cues for that property.

In Chapter 9, I will integrate the different topics covered to demonstrate that cues are differentially informative about geometric properties of shape and that these properties are represented independently by the visual system. As a consequence, cue combination is different for each property and the resulting representations are not necessarily consistent. Perceived properties of shape are not computed from a unique representation, but are obtained directly. For this reason, different tasks are not based on the same property and therefore can provide inconsistent response patterns.

On my desk there is an apple and I have the impression that I can see its shape very clearly and consistently. Many researchers believe that the consistency in shape perception is due to the Euclidean properties of shape captured by our perception. Certainly our experience of the apple is consistent with these theories: the apple appears to be a single entity and all its properties appear to be perceived veridically. It seems that I perceive the shape of the apple as it really is in the Euclidean world. This work shows that appearances can be misleading, the perceptual world is not veridical nor single. A closer analysis of the perceived shape of the apple indicates that it is not constant, shape changes with the task. The geometric properties constituting the shape of the apple are perceived independently and the task determines which ones are used. The visual world is composed of its properties, that are not integrated to achieve consistency but remain somehow independent. Perception is therefore a description of the environment based on a vocabulary of properties. This work indicates that the brain does not speak the Euclidean language. Each description of the environment is created independently from the visual information and it coexists with the others. Contrarily to my experience, rather than the shape of the apple I PERCEIVE THE SHAPE PROPERTIES OF THE APPLE.

Chapter 2

Cues as Signals

"No visual pattern is only itself" Arnheim (1954, p. 63)

As mentioned at the beginning of this work, the optic array produced by viewing an object contains several sources of information about the object's shape. These sources of information are called cues [1]. Some of the cues, defined as "pictorial", are available in a single static image; other cues, defined as "dynamic", are defined by systematic transformations of the projection; and others depend on the differences between the stimulation of the eyes and are therefore binocular cues (see Cutting & Vishton, 1995).

The visual system uses cues in the retinal image to compute an estimate of a "property" of the environment. The first step in the chain of events that lead to perception is the detection of the cue in the optic array. Contrary to common language a cue does not only describe the form of information about shape, a cue is actually a pattern in the optic array. The visual system has detectors distributed across the optic array that are tuned to the type of pattern. The response of the detectors across the optic array is the information the visual system uses to estimate the shape and should be considered the "input" available to the module. The patterns in the image and the measurements that follow are affected by errors. Imperfections in the optics

[1] One of the first terms used to define such information was 'sign' because it referred to the implicit nature of the specification of the feature. 'Sign' was mainly used by Berkeley (1709) to indicate surrogates that stand in place of some aspects of the world. The term 'cue' introduced later refers to the implicitness of information in the stimulus. 'Cue' was introduced by James (1890) from theater documents of the sixteenth century where the abbreviation Q for the Italian word 'quando' (=when) indicates the triggering of an action in response to information only hinted at. Subsequently, Titchener (1906) /cite thought that information could cue perception as well, and nowadays cues to depth are considered to be properties of the image that elicit the perception of three-dimensional features of the world. (Cutting, 1986, p 40-41 261-262)

of the eye, limitations in the neural detectors, and neuronal noise all contribute to increase the level of noise in the measurement. For this reason, the detection of a cue is often referred to as a "measurement" of a noisy "image signal".

Although often overlooked, it is important to distinguish two types of error that can occur: noise in the measurement versus sensitivity to the image signal. This nomenclature has been borrowed from K. A. Stevens (1981) who distinguishes different ways in which an estimate can be erroneous and states which are independent (precision, sensitivity and accuracy). Here I'd like to especially consider precision, which is affected by the amount of noise in the measurement. Sensitivity, the ability to detect a signal, is partially taken into consideration, but cannot be analyzed extensively because there are few investigations that analyze this issue. The problem of accuracy will be addressed experimentally in Chapter 4.

Even in presence of measurement noise it is possible to estimate a property of the environment, but this limits precision. Insufficient sensitivity has more extreme and qualitatively different consequences. Even though the image signal is there, if the visual system is not able to measure it, then it is not possible to make an estimate. This situation is in turn different from one in which the visual system possesses the ability to measure an image signal, but that signal is simply not present in the stimulus (as in the case of single cue experiments). Here the magnitude of the measurement is close to zero, but there still is an error due to noise.

Insufficient sensitivity and precision can explain why the structure from motion theorem proposed by Ullman (1979) cannot be applied to human perception. Even though it is true that the position and displacement of 4 points in 3 views is sufficient to determine the structure in 3D, this solution cannot be achieved by the visual system (i.e. Todd & Bressan, 1990). The precision required in the measurement of the positions and velocities of the points exceed the visual system's limits. Moreover, the visual system uses only velocity signals from two views to estimate structure from motion probably because it is insensitive to acceleration signals available in three views (Todd, Tittle, & Norman, 1995).

The second step to make an estimate of shape from the retinal information is the interpretation of the image signals. Image signals are created by retinal projections of 3D shape, but what the visual system does to invert this projection, and what aspects of shape can be recovered still remains unknown (this is part of the quest of this dissertation). In this work I consider especially three metric aspects of shape: depth, slant, and curvature. Different results indicate that these may not be the right surface descriptors for the visual system. However, the cue combination literature uses mostly tasks that require the evaluation of different aspects of metric shape. I chosen these properties only as a tool to investigate perceived shape without assumptions about their ecological validity.

The difficulty in the computation of a property is different than the effect of noise in the measurement discussed above, but it is also affected by this noise. In fact, a property might not be estimated correctly in two cases:

- if the measurement contains a level of noise that makes the computation impossible, as seen above, or
- if the measurement is done correctly, but the computation is impossible because mathematically underdetermined.

Even if the visual system can pick up some relevant information from the image with sufficient accuracy, this signal might be insufficient to make an estimate if taken alone. It might well be that the mapping from the 3D world to the image signals cannot be inverted because it is an ill-posed problem (Clark & Yuille, 1990; Backus & Banks, 1999). The visual input is "inadequate" for a correct and unbiased estimate because the information available is ambiguous and must be supplemented by the beholder (Berkeley, 1709). In this case, the estimate of a property by the visual system is sometime described as a matter of guesswork (Helmholtz, 1910) or generation of "perceptual hypotheses" (Gregory, 1968). Koenderink (2001) refers to the assumptions made by the observer as the "beholder share"(Gombrich, 1974), the part of the process where the observer can intervene. He underscores that the more informative the pattern is, the less one has to guess.

I believe that this problem is extremely important to an understanding of human perception. As we will see below, for some signals there exists a simple mapping from the magnitude of the measurement to a property of the world. For other signal, the estimate requires a more complex measurement or a more complex computation. Gibson (1979) stated that the world provides sufficient information for an active observer to determine the perceivable properties of the environment without supplemental information or guessing. Empirical research with an arbitrarily reduced stimulation is questionable because it forces the visual system to operate with less information than in the normal environment. The information can be manipulated experimentally only when "the essential invariant be isolated and set forth" (p305 Gibson, 1979). Therefore it is not possible to conclude anything about the behavior of the visual system from the usual psychophysical experiments. I believe that he was right in saying that there is sufficient information for the perception of properties, especially if it is provided by multiple cues. In particular, it should be defined what properties of the environment are perceived from each of these cues. I do not agree on Gibson's refusal of single cue experiment, because although they do not provide the correct information for veridical perception, they can be used to analyze the information provided by the cues.

But let's go in order. To understand how perceived shape is achieved it is necessary to analyze what patterns to measure, the image formation process, and how the signals are measured and used by the visual system. This will give us an idea of what information about shape is carried by each of the cues. It is important to spell out that this analysis indicates only what the relations between image signals and shape are from a mathematical standpoint. It would still remain to be determined whether the human visual system uses the measure, whether the measurement is precise enough, and whether the computation performed is similar to the one hypothesized. To test this relation between signal and estimate, it is necessary to establish whether the measure responsible for the estimate is the hypothesized one or a different one (that is either related to the a different surface property or a different relation exists). Moreover, it is necessary to establish which 3D property is recovered from the information (K. A. Stevens, 1981).

An example that illustrates the problem we are dealing with is Gibson (1950)'s proposal that the measurement of texture density is the key in the specification of slant. There are different questions that must be addressed. First of all, there is the question of whether slant can be reliably perceived. It has to be determined that the slant is a perceptually meaningful property of the environment. High reliability in the judgments are usually taken as indication that a property has perceptual importance (Lappin & Craft, 2000). The presence of independent aftereffects of a property are also taken as an indication of perceptual coding (as the independence of the slant aftereffect from distance Berends, Liu, & Schor, 2005). Second, it has to be proved that relation between texture density and slant is the fundamental one by ruling out every other type of information in the stimulus. To be able to prove this relation, it is necessary to isolate the density pattern in the image and make it independent from other possible measurements. This has been proven to be quite difficult to do experimentally (see K. A. Stevens, 1981, for a review). Only after more than twenty years, was it established that the density gradient is not used by the visual system. Third, it has to be determined whether the visual system can detect the texture density with sufficient reliability.

2.1 Motion

"Never mistake motion for action" Ernest Hemingway

The relative motion between an observer and the world creates a pattern of optic flow that can be a source of information for the perception of objects (Gibson, 1979). In fact, classical approaches have proven that the visual system makes use of such information to estimate 3D shape (Wallach & O'Connell, 1953; Ullman, 1979). However, these first formulations stated that the visual system uses *all* the information available in the stimulus, and that performs a correct mathematical analysis producing a veridical estimate of motion and shape of the projected object by making few assumptions (Koenderink, 1986). The various models that embrace this approach differ in terms of the type of geometrical description of the perceived shape and the assumptions used by the visual system. Several assumptions have been proposed, among others there are: rigidity (Ullman, 1979), smoothness of the flow field (Hildreth, 1984), fixed-axis motion (Hoffman, 1986), and rotation as a constant angular velocity (Hoffman, 1985).

As mentioned above, evidence has been accumulating against this approach. The visual system does not use all the information contained in the stimulus. For example, the visual system makes use of the velocity information alone and does not require the second order temporal derivatives of the optic flow for the perception of shape

(Todd, Akerstrom, Reichel, & Hayes, 1988; Todd & Bressan, 1990; Todd & Norman, 1991; Liter, Braunstein, & Hoffman, 1993; Domini & Braunstein, 1998). In fact, the perception of shape from motion information is the same even when only two views are presented to the subjects and not three as required by the theorem (Todd & Bressan, 1990; Todd & Norman, 1991; Todd et al., 1995). Perceptual performance improves very little, or not at all, if additional frames are added to an apparent motion sequence composed of only two frames (Liter et al., 1993; Norman, 1993; Hildreth, 1984). In this case, the information available to the visual system is not sufficient to recover the correct Euclidean shape (for a review see Norman, 1993). Infinite 3D structures related by an Affine stretch along the depth dimension are consistent with the stimulus. Consistent with this limitation, Todd et al. (1995) proposed a theory about shape representation based on Affine rather than Euclidean geometry.

If forced to choose a unique estimate from these infinite solutions, the visual system relies on the use of a heuristic process (Domini & Caudek, 1999) that does not guarantee a mathematically correct solution, but provides a good approximation in most conditions (Braunstein, 1994). The perceptual result is systematically related to some stimulus variables even if it does not match the Euclidean three-dimensional structure. Different patterns in the optic array can be the key information used by the visual system: the extension of the projection (Caudek & Proffitt, 1993), the range of velocities in the image plane (Liter et al., 1993), or some other local component of the optic flow like local deformation (Domini & Caudek, 1999; Todd & Perotti, 1999). This will be shown in more detail in the next section. A new line of research developed from these ideas, that examines the most reliable sources of information in the optic flow are, and the relationship between the information and the perceived shape is (i.e. Domini & Caudek, 1999).

2.1.1 Image formation

The relative motion between an observer and a three-dimensional surface rotating about an axis can be described as illustrated in Figure 2.1. A coordinate system (x, y, z) can be located at the viewing point with the z axis corresponding to the viewing direction. To simplify the analysis, the situation considered in all the subsequent text is limited to a rotation ω about a vertical axis passing through the point $(0, 0, d)$. The point $P = (x, y, z)$ can be also expressed as $P = (x, y, z_r - d)$ to simplify the calculations, where $z_r = z - d$ is the distance of the point in depth from the axis of rotation. The point P projects to $(u, v) = (\frac{xf}{z}, \frac{yf}{z})$ in the image plane Φ at a distance of f from the origin. The projected point can also be expressed in terms of the horizontal and vertical visual angles $(\alpha_u, \alpha_v) = (\arctan \frac{u}{f}, \arctan \frac{v}{f})$ that for small portions of the visual field becomes simply $(\alpha_u, \alpha_v) \approx (\frac{u}{f}, \frac{v}{f}) \approx (\frac{x}{z}, \frac{y}{z})$.

After the rotation, the coordinates of the point P becomes $P' = (x', y', z') = (x \cos \omega - z_r \sin \omega, y, d - z_r \cos \omega + x \sin \omega)$ that for small rotations can be approximated to $P' = (x', y', z') \approx (x + z_r \omega, y, z + x\omega)$, that in visual angle is $(\alpha'_u, \alpha'_v) \approx (\frac{x+z_r\omega}{z+x\omega}, \frac{y}{z+x\omega})$.

In the image plane Φ, the projected velocity can be expressed as the temporal derivative of the horizontal coordinate $\overset{\bullet}{u}$ that corresponds to $\overset{\bullet}{u} \approx \frac{(x')f}{z'} - \frac{(x)f}{z} \approx \frac{(x+z_r\omega)f}{z+x\omega} - \frac{(x)f}{z}$. If expressed in visual angles, this relationship becomes $\overset{\bullet}{\alpha}_u \approx \frac{x-z_r\omega}{z+x\omega} - \frac{x}{z}$. For small stimuli, the term $x\omega$ is negligible because it represents the perspective effect and the equation simply becomes

$$\overset{\bullet}{\alpha}_u \approx -z_r\omega \tag{2.1}$$

This equation indicates that the velocity measured in the retinal image is proportional to the distance between the point and the axis $(z-d)$ and the angle of rotation ω.

Now I will use the relation between image transformation and 3D transformation to derive the image signals associated with different properties of shape. Let's first consider zeroth order information. If equation 2.1 is applied to two points P_1 and P_2, it is possible to express the depth separation as being related to the difference in angular speed according to $z_2 - z_1 \approx \frac{\overset{\bullet}{\alpha}_{u2}}{\omega} - \frac{\overset{\bullet}{\alpha}_{u1}}{\omega}$. So, the relative velocity between points is related to the magnitude of rotation and the depth separation according to:

$$\Delta Z \approx \frac{\overset{\bullet}{\alpha}_{u2} - \overset{\bullet}{\alpha}_{u1}}{\omega} \tag{2.2}$$

This relation does not allow to estimate depth separation from image signals unless the visual system provides a value for the angle of rotation. The visual system may

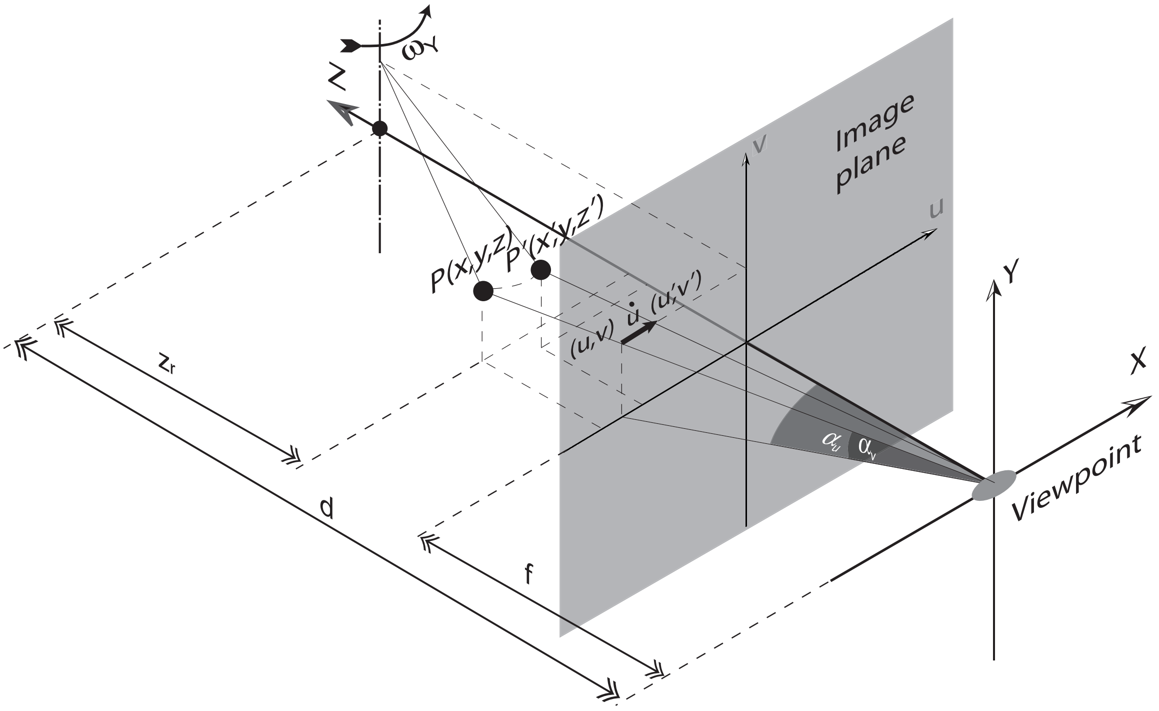


Figure 2.1: The rotation of a point P around a vertical axis creates a pattern of motion in the image plane Φ. The velocity with which the projected point moves on the image plane can be used to estimate the point's distance in depth from the axis of rotation

assume it, or estimate it from the image signals. In this case, the observer needs only to measure the velocity signals to obtain an estimate of the relative depth of points.

This analysis is similar to the conclusions drawn by Perotti, Todd, Lappin, and Phillips (1998). They state that when an object rotates rigidly in depth under *orthographic* projection, the formula $z = -V\sqrt{\frac{x}{accel}}$ can be applied to estimate depth directly from the values of speed and acceleration (up to a reflection in depth)[2]. The estimate of Euclidean distance in depth between points requires the measurement of the second order components of the optic flow. If this information is not detectable,

[2] in perspective projection $\dot{\alpha}_v$ can be used to solve this ambiguity

as discussed above, the visual system could recover the relative depth of point only up to an Affine transformation in depth (see Koenderink, 1990; Todd & Bressan, 1990).

To analyze the first order geometrical properties of surfaces, namely the local orientation of the surface, we will consider that a smooth surfaces S can be locally approximated by planar patches defined by $z = g_x x + g_y y + z_0$. Here g_x and g_y are the horizontal and vertical depth gradients and z_0 is the distance of the point from the origin. If this formula is expressed in terms of visual angle by substituting $(\alpha_u, \alpha_v) \approx (\frac{u}{f}, \frac{v}{f}) \approx (\frac{x}{z}, \frac{y}{z})$, we obtain $z = g_x \alpha_u z + g_y \alpha_v z + z_0$. I will approximate the optical projection of points as composed of a parallel projection to the image plane Φ and a perspective projection to the origin. This state of things can be rendered mathematically by substituting f for z on the right side of the equation so to obtain $z \approx g_x \alpha_u f + g_y \alpha_v f + z_0$. The velocity field in this case is described by the formula

$$\overset{\bullet}{\alpha}_u \approx \omega(d - z0) - \omega g_x \alpha_u f - \omega g_y \alpha_v f \qquad (2.3)$$

These three factors evidence the classical decomposition of the linear velocity field in its spatial derivatives:

$$\overset{\bullet}{\alpha}_u \approx \overset{\bullet}{\alpha}_{u0} + \overset{\bullet}{\alpha}_{uu}\, \alpha_u + \overset{\bullet}{\alpha}_{uv}\, \alpha_v \qquad (2.4)$$

where the term $\overset{\bullet}{\alpha}_{u0}$ is the traslatory component and the velocity gradients $\overset{\bullet}{\alpha}_{uu}$ and $\overset{\bullet}{\alpha}_{uv}$ can be grouped in the term $def = \sqrt{\overset{\bullet}{\alpha}_{uu}{}^2 + \overset{\bullet}{\alpha}_{uv}{}^2} = \sqrt{(\omega g_x f)^2 + (\omega g_y f)^2}$ (Domini & Caudek, 1999; Liter & Braunstein, 1998; Todd & Perotti, 1999).

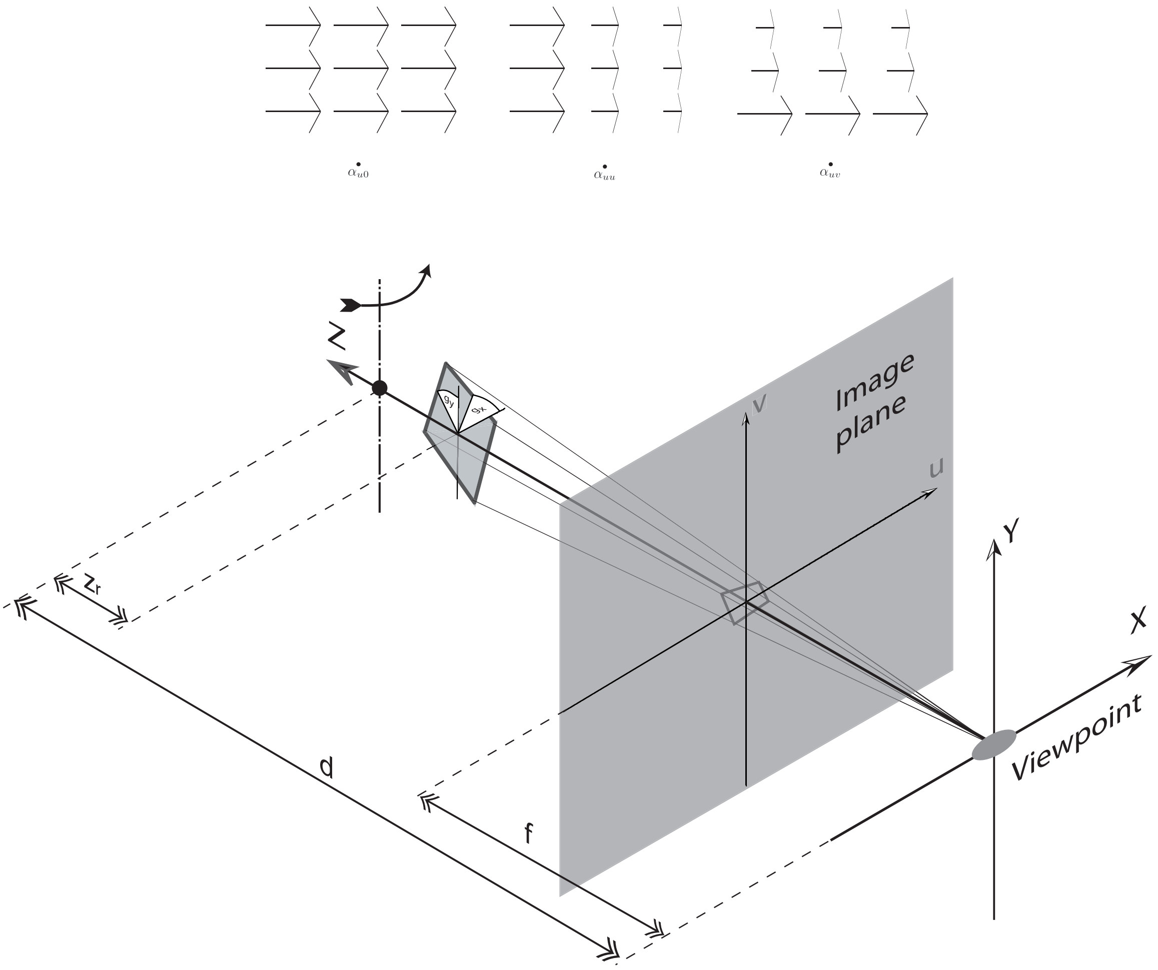


Figure 2.2: A patch of the surface S that rotates around a vertical axis generates a linear velocity field described by the three components $\dot{\alpha}_{u0}$, $\dot{\alpha}_{uu}$ and $\dot{\alpha}_{uv}$ depicted above

This formulation of the velocity field in terms of its spatial derivatives, allows to compute the "tilt" $\tau = \frac{g_x}{g_y}$ of patch. In fact, by multiplying each term of the fraction for ωf the formula can be expressed as $\tau = \frac{\omega f g_x}{\omega f g_y} = \frac{\dot{\alpha}_{uu}}{\dot{\alpha}_{uv}}$, a quantity of the velocity field that uniquely specifies the tilt of the surface.

On the other hand, the "slant" of the surface, defined as $\sigma = \sqrt{g_x^2 + g_y^2}$, is not univocally specified by the velocity field because, since ω is unknown, the relation with the velocity field component def can be expressed only as

$$def = \sqrt{(\omega g_x f)^2 + (\omega g_y f)^2} = f\omega\sqrt{g_x^2 + g_y^2} = |f\omega_y\sigma| \tag{2.5}$$

In this case, the information provided by def is ambiguous if the amount of rotation is unknown (as we've seen above there are infinite solutions). In (Domini & Caudek, 1999) the authors proposed that the ambiguity of the velocity field could be solved by selecting, among the infinite pairs of slant and angular velocities compatible with a given *def*, the most likely one. In particular, they have shown that, if the *a-priori* distributions of slant and angular velocity are uniform and limited, then the a posteriori probability distribution $p(\omega_y, \sigma|def)$ has a maximum for the pair of values $\omega^* = k\sqrt{def}$ and $\sigma^* = \frac{1}{k}\sqrt{def}$. So, if it is possible to separate σ and ω with an heuristic process or it is possible to estimate ω either with a global analysis of the information across the stimulus (as proposed in Di Luca, Domini, & Caudek, 2004), then it is also possible to use local measurements of the optic flow to estimate of slant.

Moving from first order descriptors (local orientation) to second order descriptors of local shape, it is reasonable to expect that since the local orientation of a smooth surface is related to the first spatial derivatives, second order properties of shape will be related to second spatial derivatives of the optic flow. In fact, if the shape under analysis is expressed as a Monge surface with the formula $z = S(x, y)$ then the optic flow is $V(x, y) = S(x, y)\omega$ where the second order spatial derivatives are $V_{xx} = \omega S_{xx}$, $V_{xy} = \omega S_{xy}$, and $V_{yy} = \omega S_{yy}$.

The shape can be described locally using the values of the principal curvatures. The principal curvatures are the maximal and minimal directional curvature measured at a given point. Interestingly, the direction where these two extreme values are found are necessarily orthogonal to each other. The direction with maximal curvature is tilted by a certain amount with respect to the coordinate axes x and y. This angle can be easily calculated from the optic flow pattern according to the formula

$$\cot(2\alpha_{\kappa_M}) = \frac{1}{2}\left(\frac{S_{xx} - S_{yy}}{S_{xy}}\right) = \frac{V_{xx} - V_{yy}}{2V_{xy}} \tag{2.6}$$

So in the same manner as for tilt, the direction of maximal curvature can be obtained from a image measurement without any free parameters. The relation between the magnitude of each of the principal curvatures and the retinal motion, on the other hand, is expressed by the formula

$$\kappa_x = \frac{1}{\sqrt{1 + S_x^2 + S_y^2}}\left(\frac{S_{xx}}{1 + S_x^2}\right) = \frac{\omega^2 V_{xx}}{(\sqrt{\omega^2 + V_x^2 + V_y^2})(\omega^2 + V_y^2)} \tag{2.7}$$

Notice that as happened with slant, in this case the term ω must also be estimated from the equation in order to make an estimate of the magnitude of curvature.

For example, Mamassian, Kersten, and Knill (1996) noticed tat the curvature ratio is an invariant relationship, a description of the local shape of the surface. He called this quantity the "shape characteristic" (figure 2.3) and showed that its value for a slanted patch with principal curvatures aligned to the coordinate axis is related to the velocity field by the equation $\frac{\kappa_x}{\kappa_y} = \frac{S_{xx}}{S_{yy}}\left(\frac{1+S_y^2}{1+S_x^2}\right) = \frac{V_{xx}}{V_{yy}}\left(\frac{\omega^2+V_y^2}{\omega^2+V_x^2}\right)$. He also showed that when the normal to the surface is parallel to the viewing direction the shape characteristic can obtained from the velocity gradient, eliminating the rotation term. In fact, in this case the ratio of curvature can be computed as

$$\frac{\kappa_x}{\kappa_y} = \frac{V_{xx}}{V_{yy}} \quad . \tag{2.8}$$

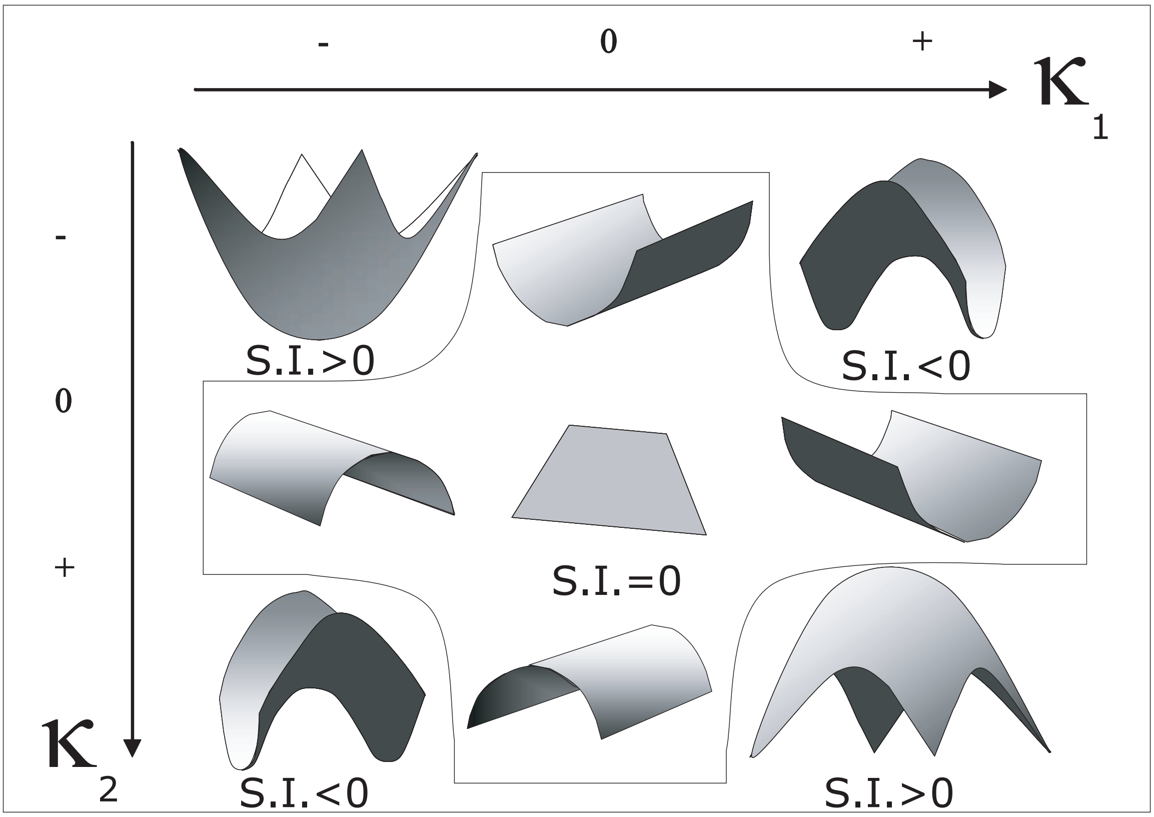
For non frontoparallel patches, the resulting error due to using this formula is small for a wide range of slants (see Dijkstra, Snoeren, & Gielen, 1994). So, the shape characteristic can be estimated from the image signal with no free parameters for most patches.

Similarly, the "shape index" defined by (Koenderink, 1990) as $S = \frac{-1}{\pi} \arctan \frac{k_M + k_m}{k_M - k_m}$ can be used as a descriptor of the local shape as whown in figure 2.3. The shape index can be computed from the optic flow as $S = \frac{-1}{\pi} \arctan \frac{V_{xx}(1+V_y^2)+V_{yy}(1+v_x^2)}{V_{xx}(1+V_y^2)-V_{yy}(1+v_x^2)}$. If the patch is locally frontoparallel the equation simplifies to

$$S = \frac{-1}{\pi} \arctan \frac{V_{xx} + V_{yy}}{V_{xx} - V_{yy}} \quad , \tag{2.9}$$

therefore there is a direct relation between the second order derivatives of speed and the shape index.

The other factor that is commonly used to describe local shape is called curvedness and is defined as $C = \sqrt{\frac{k_m^2 + k_M^2}{2}}$ (Koenderink, 1990). Curvedness captures the magnitude of curvature at a point. Similarly to what has been said about depth and orientation, the absolute magnitude of curvature is not uniquely specified by the



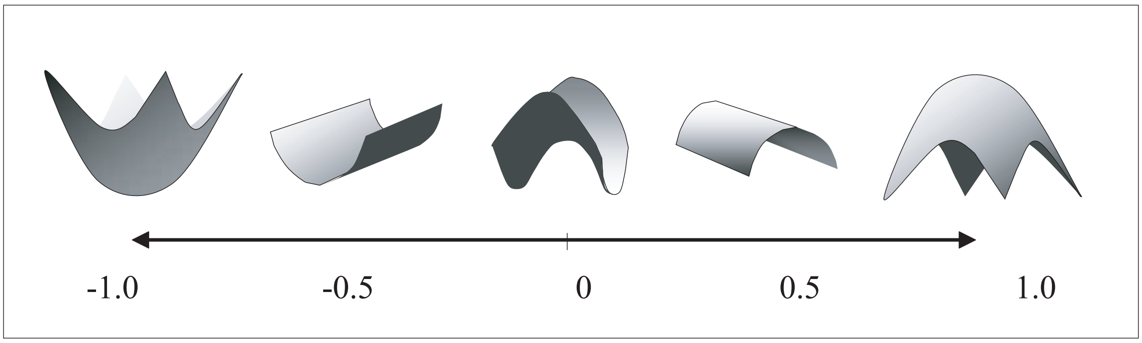


Figure 2.3: Top: Values of the shape characteristic for the two principal curvatures. The sign of the shape characteristic is related to the sign of the two curvatures: negative for hyperbolic shapes and positive for elliptic ones. It is 0 if the local shape is flat in at least one direction. Bottom: The value of the shape index, in the same manner as the shape characteristic, is zero if either one of the curvature is null, but also when the two curvatures have same magnitude but opposite sign. Then it acquire a positive sign for convex surfaces and negative for concave ones.

velocity field as in the case of curvedness. Differently, since shape characteristic and shape index are a ratio of principal curvatures, they are independent from the amount of rotation and can be therefore estimated using only local information in the optic flow.

In this section I analyzed the process of image formation in order to understand the relation between image signals and geometric properties of shape. From this analysis it is evident that local depth, slant, and amount of curvature can be estimated from local measurements of the optic flow if two conditions can be satisfied: image signals (velocity difference, deformation, second order spatial derivatives) can be detected and measured, and the amount of rotation ω can be accounted for either by heuristic process or by a non local computation involving the whole object. On the other hand, qualitative shape can be estimated if second order spatial derivatives can be measured, without knowing the rotation.

2.1.2 Psychophysics

Hogervorst and Eagle (1998) report some estimates for the uncertainty in the estimate of speed. The authors report that it is often assumed that the uncertainty in the retinal position of a feature is negligible, since in optimal condition the Weber fraction of a three-lines bisection task is under 2% (Westheimer & McKee, 1979). Errors in the measurement of retinal velocities are directly reflected on the estimate of depth. (De Bruyn & Orban, 1988)offer an estimate of the human sensitivity for speed measurement. For speeds up to $64^o s^{-1}$, the measurement noise is characterize by $\sigma_s = 0.049 + 0.035S[^o s^{-1}]$.

The data from Perotti et al. (1998) and Lappin and Craft (2000) shows that observers do not obtain reliable information about the metric curvedness of surfaces. However, they are sensitive to the qualitative shape of the surface defined by the ratio of the two principal curvatures. This means that the visual system is sensitive to the second order derivatives required for the estimates of local shape, but it does not posses the ability to account for the amount of rotation as it was said above.

2.2 Texture

A textured surface S projects an image in which the texture pattern has systematic distortions that depend on the viewing direction, distance and other aspect of the projection. This deformation of texture contains information about these properties. Gibson (1950) introduced the term “texture gradient” referring to the systematics of the patterns of optical texture across optic array. Texture gradients, in fact, contain information because the pattern of texture on physical surfaces is repetitive or regular. To the extent that this holds, the differences in patterns of optical texture across the optic array correspond to the projective effect of the arrangement of the surfaces. It is possible to use these variation of texture in different locations to infer variations in depth and surface orientation.

Texture information can be generally divided in to two categories. The first type of information is due to the projective effect on the *distribution of texture* across the surface. The second type of information is related to the *shape* of the individual texture elements. Let's analyze these two components in order.

Texture distribution across the image does not require the presence of identifiable features; the information is contained in the statistical pattern of the whole texture. The projection has different effects on the texture distribution and these effects are not completely independent and easy separable, so it is rather hard to identify and isolate the different texture distortions. For this reason, definitions are not precise in this regard. Two types of nomenclatures will be introduced here, and in the next section the information contained in the subdivision will be spelled out.

K. A. Stevens (1981) distinguished two independent effects of projective transformations of texture: scaling of texture size due to differences in viewing distance, and compression of texture due to orientation of the surface in depth with respect to the line of sight.

Cutting and Millard (1984) adopted a different view in the categorization of the information provided by the texture. They identified three independent constraints that characterize the regularities of many textured surfaces: texture elements are approximatively the same size, are evenly spaced and are relatively flat. When a surface is viewed in perspective each of these constraints produces an image gradient (Gibson, 1950) along different dimensions.

This subdivision, however does not fully capture the information available to the observes (see Gärding, 1992; Knill, 1998). Here we will define gradients as:

- Scaling gradient: if the size of the elements is constant on the surface, the size of the projected elements is inversely proportional to their distance. The scaling component of the perspective projection can be used to derive depth and surface orientation. It is an isotropic transformation of the texture (Witkin, 1981) and according to K. A. Stevens (1981) the slant and curvature of the surface can be derived from the local depth map obtained directly from the size of the elements.

- Compression gradient: if the texels are flat with respect to the surface, the difference in viewing direction form the surface normal produces a projective foreshortening. Compression of texture carries different information from the size gradient, it specifies the local surface orientation, but to derive depth would require integration of orientation into a coherent surface.

- Density gradient: Density is a scalar property of an area of the image, the spacing between elements varies with direction. If the distance between features is constant across the surface, the density on the image is related to surface distance. K. A. Stevens (1981) argues that the difference in density would not be a useful measure for computing distance or surface orientation because it varies with both depth and orientation. To see this point, it is possible to anticipate the analysis of the image in the next paragraph and express the factors influencing

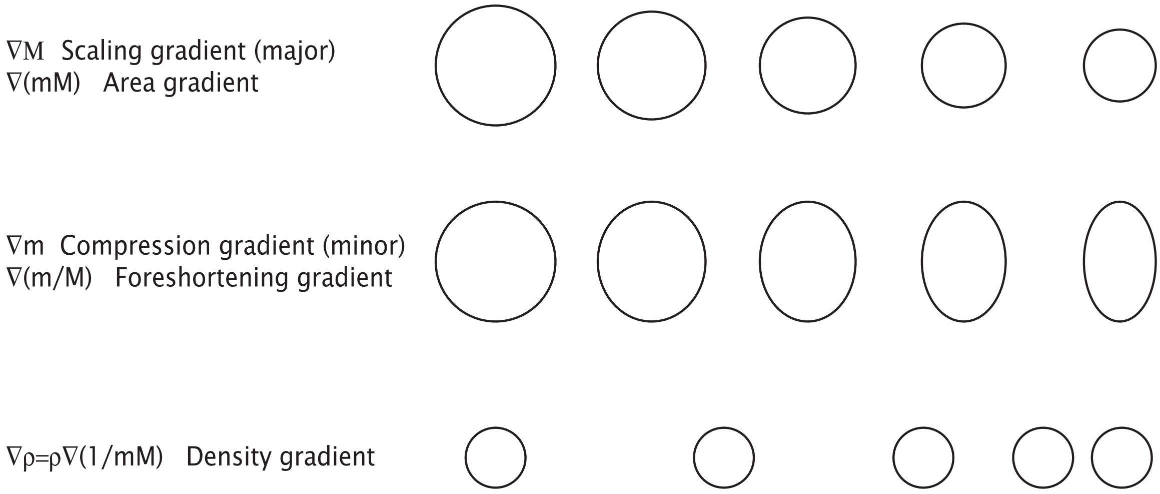

∇M Scaling gradient (major)
∇(mM) Area gradient
∇m Compression gradient (minor)
∇(m/M) Foreshortening gradient
∇ρ=ρ∇(1/mM) Density gradient

Figure 2.4: The different types of texture gradients, nomenclature and the mathematical formulation according to Garding (1992). The horizontal axis of the ellipses represents the magnitude of m and the vertical axis the magnitude of M (see further).

the density of texture elements on the image as $\varrho = \varrho_s z^2 / \cos\sigma$, where ϱ_s is the density of texture on the surface (K. A. Stevens, 1979). Both z and σ contribute to this value at every point. Similarly, the gradient based on this quantity is determined by both slant and curvature.

A second type of information available from the optical projection of a textured shape is the effect that the projection has on *single elements* composing the texture, what Gärding (1992) calls first-order information. Most of the psychophysical work has relied on textures composed of discrete elements called texels. It is clear that this type of information is the more relevant and easier to detect. It carries information about different aspects of shape. The perspective effects on texels can be described using the same framework described above.

- The geometric effect that creates the compression gradient, also generates the difference between the actual shape of texels and their projected shape. The amount of compression from a point in a particular direction on the surface depends on the angle subtended between the line of sight and the direction analyzed. Elements that are circular on the surface project into ellipses when the surface is viewed at an angle. It is the most prominent effect of the different orientation of the surface relative to the line of sight.

- The size of texture elements corresponds to the scaling component of the perspective projection. It carries information about depth.

2.2.1 Approaches

There have been several attempts to relate 3D properties of surfaces with image patterns (or texture measurement K. A. Stevens, 1981) especially for the ground plane. The first, classical analysis considered only distance and orientation as possible 3D properties, but recently curvature has been introduced as well.

The information contained in the texture pattern of an image is complex because it is probabilistic in nature and relies on making assumptions about the original pattern of texture on the object's surface. Two general assumptions used extensively are texture "isotropy" (the texture has the same characteristics in every direction) and "homogeneity" (the properties of the texture are constant across the surface) (Malik & Rosenholtz, 1997; Rosenholtz & Malik, 1997). Statistical variation in texture can be described using these terms to quantify the level of noise in the incorrect interpretation of variation in optical texture as due to 3d structure.

One way to operationalize isotropy is to measure “compactness”: the amount of area contained in a closed contour divided by the square root of the perimeter length (Ikeuchi, 1984). This measure can be used to derive the 3D shape of object because by finding the maximally compact shape in the projection. This shape is assumed to be the “simplest” version of the texture. The transformation that maps the simplest texture to the projected texture indicates the slant of the surface at every point of the image.

A related approach that does not require that individual texture elements are explicitly identified was proposed by (Witkin, 1981). It assumes that orientations of luminance edges in nature have an isotropic distribution. Deviation from this equal distribution is due to the the projective effect of the orientation of the surface. The direction where the edges are less frequent on the image corresponds to the tilt, the ratio between e maximal and minimal occurrence of borders relates to slant.

Similar methods use the same logic to infer surface orientation from the power spectrum at low and high luminance frequencies (Jau & Chin, 1990) or the second order spatial moments of the local spectra (Gärding, 1992) or of the texels (Knill, 1998).

Local parallelism of image features (K. A. Stevens, 1981) or spatial frequencies (Super & Bovic, 1995) can be also used as a constraint in feature based models to be substituted to the isotropy constraint. Locally parallel orientations, can only be preserved in the image projection when the Gaussian curvature is zero, therefore these method find a perfect solution only for plane or cylindrical surfaces.

An assumption that can be made (and that has been used extensively to make measurements on textures) is *homogeneity*: the statistical distribution of texture properties on the surface. (Aloimonos & Swain, 1988) assume uniformity of space in their model to derive slant from the gradient of density. This approach does not assume isotropy and is therefore more robust. (Gärding, 1992) and (Malik & Rosenholtz, 1997) used homogeneity more directly to find areas in the image that can be matched in the image spectrum up to a geometrical transformation. The type of transformation needed to map the areas is used to constrain the difference in orientation and distance.

The problem of the homogeneity constraint is the measurement of statistical properties of the texture (K. A. Stevens, 1984). Deviations from homogeneity can be calculated only over large regions of the image. Therefore, this constraint can be applied only to surface with small changes in orientation. Stevens argues that this spatial limitation create difficulties in using gradients as cues. The local properties might be more informative. Isotropy is less affected by spatial restriction than homogeneity is, because it is a *local* constraint. In situation containing high amount of noise, the use of this constraint is thus more efficient. However, the limits of this constraint are due to the existence of many natural anisotropic textures (like wood grain). Size, density and compression gradients are useful only by applying the constraint of homogeneity to different properties of the texture. Isotropy is the constraint used for local compression of shape or orientation (including asymmetries of the spectra).

Another assumption common to most of these approaches, that is seldom modeled mathematically and will be mention here just for completeness, is that texture is *flat* and coplanar with the surface it belongs. There is a class of natural surfaces for which this constraint does not hold. Surfaces that have a textured relief are constituted by regular perturbation perpendicular to the global surface.

2.2.2 Image formation

Two processes give rise to the information in the image. The first is the process that generated the texture on the surface and it is reflected on the regularities of the pattern. The second is the optical projection which map surface texture to image textures. In the following analysis, from Gärding (1992), we will see the effect of these processes.

When a surface pattern is projected on the optic array, the local metric structure is distorted systematically. The distortion of the structure is purely geometric and independent of the surface pattern. However, the pattern of the surface is what allows a measurement of certain component of the distortion in the optic array.

Gärding (1992) illustrates the geometry of image creation by projecting the surface S on a viewsphere Σ with unit ray (see figure 2.5). He defines the projection map F

of S to Σ as $F(\mathbf{p}) = r(\mathbf{p}) = r(\mathbf{p})\mathbf{p}$ where $\mathbf{p}$ is a unit vector from the focal point to a point of the unit sphere and $r(\mathbf{p})$ is the distance along the visual ray from the focal point through $\mathbf{p}$ to the point $\mathbf{r} = F(\mathbf{p})$ on the surface S.

The local change of metric depends only on the linear part of F_*, the derivative map of F. F_* is the projection of the image on σ to the surface S; it maps tangent vectors of Σ at $\mathbf{p}$ to tangent vectors of S at $F(\mathbf{p})$. For example a small segment $d\mathbf{x}$ in the image maps to the segment $d\mathbf{u} = F_* d\mathbf{x}$ on the surface. The length ratio between the two segments du/dx depends on the surface distance and orientation. Defining the tilt $\mathbf{t}$ as being a unit vector tangent to σ at $\mathbf{p}$ with direction of the gradient of the function $r(\mathbf{p})$, it is possible to create an orthonormal basis $(\mathbf{tb})$ for a tangent plane at $\mathbf{p}$ where $b = \mathbf{p} \times \mathbf{t}$. With this basis, it is possible to describe the linear map F_* as

$$F_* = \begin{pmatrix} r/\cos\sigma & 0 \\ 0 & R \end{pmatrix} = \begin{pmatrix} 1/m & 0 \\ 0 & 1/M \end{pmatrix}$$

where m and M are the inverse eigenvalues of the derivative map F_* ($m < M$) depicted in figure 2.5 and called the “characteristic values” (for the derivation see Gärding, 1992). m corresponds to the ratio of length in the image in the tilt direction and on the surface in the same direction. M is the analogous ratio in the direction orthogonal to the tilt. This description shows that compression of length (the length M in the unforeshortened direction) is inversely proportional to the distance r. Moreover, compression in the tilt direction m is inversely proportional to the distance r divided for the cosine of the slant σ. To summarize this part, by measuring the distortion of local surface geometry it is possible to obtain the quantities $r = 1/M$,

$\sigma = \arccos(m/M)$ and $\mathbf{t} = \pm\mathbf{v}_1$, where $\mathbf{v}_1$ is the eingenvector corresponding to m. An analogous estimate of $\mathbf{t}$ would be to find the direction of maximal magnitude of the directional derivative of the characteristic values values.

By using this description, foreshortening can be defined as the ratio of projected lengths measured in the tilt and perpendicular directions. This projected ratio is related to slant independently from curvature, but to be measured, isotropy is a necessary assumption.

This analysis considered local effect of projection on texture. Gibson suggested that spatial variation of projective distortion are informative for the surface shape as well. The scheme proposed by Gärding (1992) offers a way to describe some of the gradients mentioned in the section above as well. He expresses the "simple distortion gradients" as such: the scaling gradient $\xi_2\nabla M$ (or major gradient), the compression gradient as $\xi_1\nabla m$ (or minor gradient), the foreshortening gradient $\nabla\epsilon = (\xi_1/\xi_2)\nabla(m/M)$, the area gradient $\nabla A = \xi_1\xi_2\nabla(mM)$, and the density gradient $\nabla\rho = \rho_S\nabla(1/(mM))$. These gradients contain unknowns $\xi_1\xi_2\rho_S$ that can be eliminated by normalizing using $(\nabla f)/f$. This normalization allows for estimates without relying on prior information. By defining the common factor $f = \begin{pmatrix} r\kappa_t/\cos\sigma \\ \\ r\gamma_\tau \end{pmatrix}$ where r is the distance, σ is the slant, τ is the tilt, κ_t is the normal curvature in the tilt direction and γ_τ is the geodesic torsion of the surface in the tilt direction. It is possible to express the normalized gradient above as:

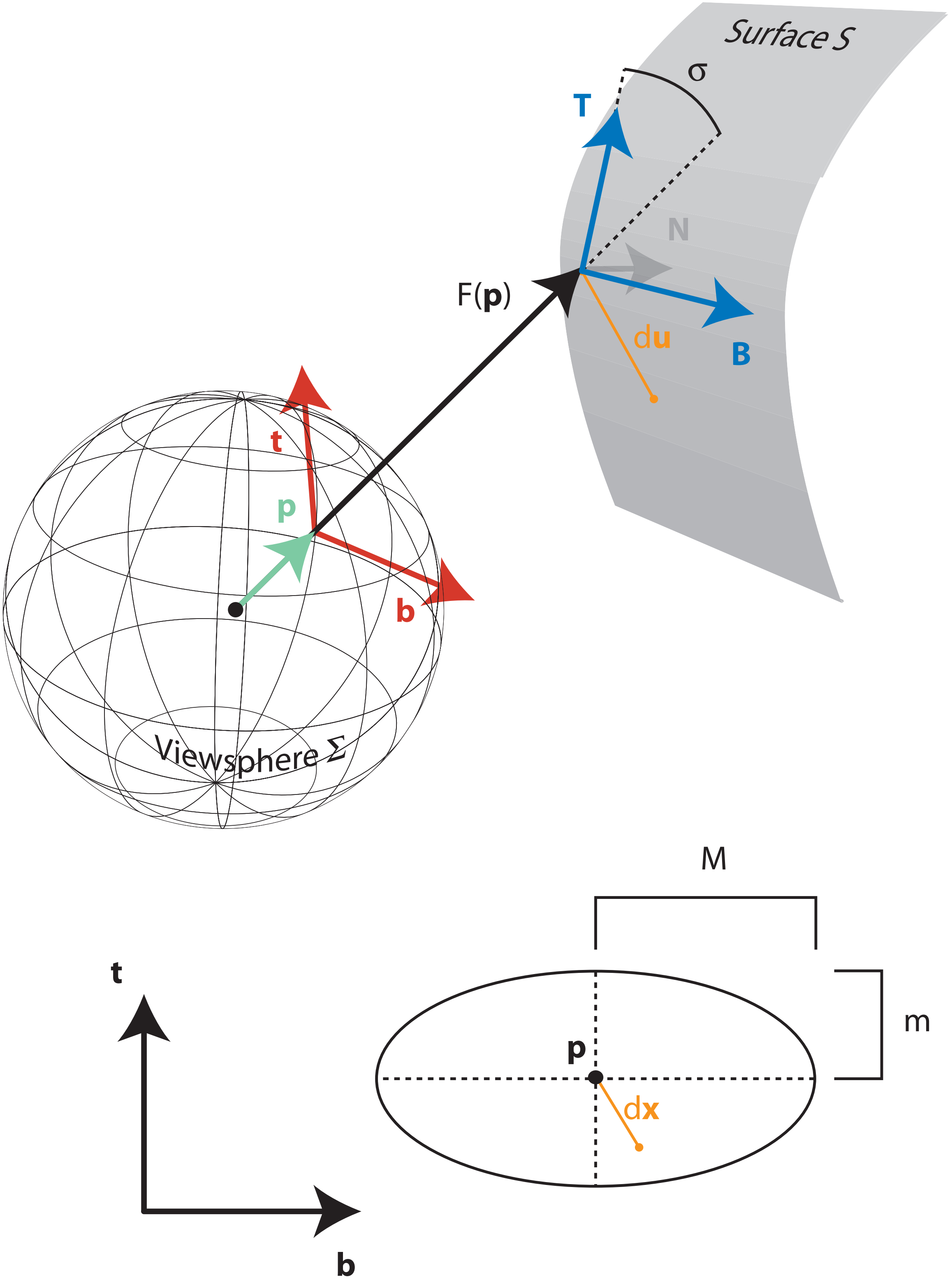


Figure 2.5: Top: Surface geometry according to Garding (1992). The mapping F creates the correspondence between the two orthonormal basis **T B** and **t b**, as well as between the two segments d**x** and d**u**. Bottom: F is visualized as the projection of a unit circle on the image plane Φ tangent to Σ in **p**. The characteristic values m and M correspond to the halflenghts of the ellipse's axis.

$$\frac{\nabla m}{m} = -\tan\sigma \begin{pmatrix} 2 \\ \\ 0 \end{pmatrix} + f \quad , \frac{\nabla M}{M} = -\tan\sigma \begin{pmatrix} 1 \\ \\ 0 \end{pmatrix} \quad , \frac{\nabla \epsilon}{\epsilon} = -\tan\sigma \begin{pmatrix} 1 \\ \\ 0 \end{pmatrix} + f$$

$$\frac{\nabla A}{a} = -\tan\sigma \begin{pmatrix} 3 \\ \\ 0 \end{pmatrix} + f \quad , \frac{\nabla \rho}{\rho} = -\tan\sigma \begin{pmatrix} 3 \\ \\ 0 \end{pmatrix} + f$$

The complete local surface curvature (in terms of normal curvature κ_t twist γ_τ, and Gaussian curvature) cannot be determined by distortion gradients alone. The usefulness of any texture gradient is determined by the information it contains about the surface and possibility to measure it. The compression, foreshortening, area, density gradients all contain similar information about shape, but they differ in the relative weight of the curvature. At a point of known orientation on a curved surface any of the gradients is informative about scaled curvature, whereas for a flat surface each of them determines the orientation uniquely. All the gradients depend on curvature and geodesic torsion, with the exception of the scaling gradient which is independent from local shape. Only scaling gradient $\frac{\nabla M}{M}$ can be used to estimate the orientation σ for surfaces with unknown curvature if spatial invariance is assumed.

2.2.3 Psychophysics

Gibson (1950) claimed that relative texture density is proportional to the relative distance of the corresponding points on the surface ($z_1/z_2 = \varrho_1/\varrho_2$). This formulation is actually incorrect, because texture density is a function of distance, but also of foreshortening as well (see K. A. Stevens, 1981). Other classical formulation establish relation between image patterns and distance. However, the distance is expressed in terms of distance along the ground, therefore from the observer feet.

The perception of slant has been initially attributed to the density gradient (Purdy, 1960; K. A. Stevens, 1979). In fact, the following equation holds: $\tan\sigma = (\nabla\varrho)/3\varrho$, where ϱ is the texture density at one point in the image when surfaces are planar and textures are uniform. Subsequently, K. A. Stevens (1981) proposed a simple model that used the value of local foreshortening m/M to estimate orientation and the major axis M to estimate local depth. This model is in accordance with the analysis made here, but it is also incomplete as my analysis is, because does produces inconsistencies for anisotropic or spatially inhomogenuos textures. Nonewithstanding these limitaion, this idea is supported by different findings.

Cutting and Millard (1984) compared the importance of the scaling (or perspective), foreshortening (or compression) and density gradients for the perception of flat and cylindrical surfaces. Density gradient had only a minimal contribution to perception. The scaling gradient was important for flat surfaces and the foreshortening was important for cylindrical surfaces. Buckley and Frisby (1993) found that the influence of local foreshortening was significantly higher than the other types of information in different cases. Todd and Akerstrom (1987) demonstrated that foreshortening is the predominant information used by the visual system, but not the only one. When the surface was shown in orthographic projection, so that there was no scaling gradient, the perceived depth increased.

J. Hillis, Watt, Landy, and Banks (2004) made a measurement of the sensitivity to texture cues in the estimate of slant. They presented monocularly a Voronoi pattern to the subjects and measured slant discriminability for the texture cue in isolation. From these data it is clear that JND decrease with increase slant, since image signals associated with the same difference in slant increases in magnitude and therefore in detectability (Blake, Bülthoff, & Sheinberg, 1993; Knill, 1998).

Similar data is obtained by Knill (n.d.). He found that slant can be discriminated away from the fronto-parallel only with angles from 29 to 46^o. For slants higher than 70^o, the discrimination thresholds decreased to 1.2 to 3.1^o. This confirms that texture only becomes more as the slant increases.

Gärding (1992) notices the lack of further investigation about the assumption that surfaces are planar even though seldom surfaces on the real word are perfectly planar. The planarity could, however, be assumed locally. Most of the image signals about texture distortions that actually depend on curvature even for infinitesimal patches.

2.3 Shading

> "The outlines and form of any part of a body in light and shade are indistinct in the shadows and in the high lights; but in the portions between the light and the shadows they are highly conspicuous." Leonardo (1888)

Shading patterns can be used as an independent cue for shape perception. The search for a suitable algorithm that can derive shape from shading has been a central problem in computer vision since its beginning. It is well known that the problem of reconstructing a 3D surface from its image is ill-posed, under constrained (B. Horn & Brooks, 1989) and the solution is not unique. Illuminant intensity, surface material and the orientation contribute to the light pattern, and are confounded in a single measurable variable. This problem is particularly evident when an image of an object can appear to depict either a convex or a concave surface (Gibson, 1950), a situation named the "Bas-relief Ambiguity".

2.3.1 Image formation

Most of the studies involving shading used synthetic images of objects (there are also exceptions like Koenderink et al., 1996). Syntectic images are created using a simplified description of the optical laws of reflection called *image irradiance equation* (B. Horn & Brooks, 1989). For a point p on a surface defined as the Monge surface $z = S(x, y)$ this equation determines the amount of light reflected toward the viewer. The local orientation of the surface is described as the normal vector $\mathbf{n} = \frac{(p,q,1)^T}{\sqrt{1+p^2+q^2}}$ and the light direction shining on the surface is $\mathbf{l} = (l_x, l_y, l_z)$. If the surface is matte, so that the reflectance function is Lambertian, the image intensity is determined by the image irradiance equation:

$$I = \rho\lambda(\mathbf{n} \cdot \mathbf{l}) \tag{2.10}$$

where ρ is the albedo of the surface (the ratio between the reflexed and incident flux of light) and λ is the light source Illuminance (flux per unit area of incident light). Notice that if the patch considered points toward the illuminant and therefore $(\mathbf{n} \cdot \mathbf{l}) = 1$, the brightness in the image will be $I = \rho\lambda$. If the surface has uniform reflectivity, if such point is present in the image, by measuring the brightest point one can estimate these parameters.

In the irradiance equation above, it is possible to substitute the values of the elements in the vectors we can write (this formulation will be useful in the Chapter 4): $I = \rho\lambda\frac{pl_x+ql_y+l_z}{\sqrt{1+p^2+q^2}}$. The first directional derivative of the image intensity in the direction $(dx/ds, dy/ds)$ for a distance s can be decomposed in the partial derivatives in the direction of the coordinate axis. $\frac{dI}{ds} = \frac{dI}{dx}\frac{dx}{ds} + \frac{dI}{dy}\frac{dy}{ds}$. The directional derivative can be calculate from the image irradiance equation obtaining: $\frac{dI}{ds} = d(\rho\lambda(\mathbf{n} \cdot \mathbf{l})) = \rho\lambda(d\mathbf{n} \cdot \mathbf{l}) + \rho\lambda(\mathbf{n} \cdot d\mathbf{l})$. Since the light source is assumed to be at a significant distance from the surface, $\mathbf{l}$ is constant and then $(\mathbf{n} \cdot d\mathbf{l})$ is zero. So we obtain

$$dI = \rho\lambda(d\mathbf{n} \cdot \mathbf{l}) \tag{2.11}$$

This equation shows how the derivative of the image intensity in a direction depends on $d\mathbf{n}$, the difference in the normal vectors $\mathbf{n}$ at different points. This difference corresponds to the curvature of the shape in the direction of s. Similarly, higher order derivatives of the shape and of the intensity are related in the same manner $d^2I = d(\rho\lambda(d\mathbf{n} \cdot \mathbf{l})) = \rho\lambda(d^2\mathbf{n} \cdot \mathbf{l}) + \rho\lambda(d\mathbf{n} \cdot d\mathbf{l}) = \rho\lambda(d^2\mathbf{n} \cdot \mathbf{l})$

The intensity on the image depends on the surface normal, and there are many surface normals that yield the same image intensity. Partial derivatives of the image brightness, in a similar way, depend on the partial derivatives of the surface normals, and do not specify uniquely Euclidean shape when the range of all possible surfaces is considered (Bruss, 1982).

B. Horn (1986) showed a possible way to calculate the curvature of the surface from the gradient of luminance. His method aimed at calculating slant change in the direction of the gradient of the reflectance map. He showed that in this case [3] the difference in the orientation (expressed in terms of (p, q) as described at the beginning of this section) can be calculated as:

$$\begin{pmatrix} dp \\ dq \end{pmatrix} = \begin{pmatrix} \frac{dI}{dx} \\ \frac{dI}{dy} \end{pmatrix} ds$$

This equation allows to compute a property of the surface directly from a measurement on the image without the need of additional assumptions. However, it is not possible to determine the global shape of the surface unless the slant at a point is known. From this point the shape of the surface can be extended in one direction along a “characteristic strip”. Many of the assessment in computer vision utilize this methodology. Alternatively, since the image brightness relates to the slant of the surface, it is possible to minimize a functional in order to find a surface that is compatible with the image provided.

[3] when ds is parallel to the gradient of the reflectance map $\nabla R(p, q)$

2.3.2 Psychopshysycs

Different studies of shape from shading often came to the conclusion that shading is one of the weakest cues (Barrow & Tenenbaum, 1978; Mingolla & Todd, 1986; Todd & D., 1989). But shading is another example of a cue that is not a cue to depth, because it provides information useful to estimate only certain geometric properties of shape. It is important to discriminate the various results in terms of the property that was required by the task.

The results of different studies, in fact, start to make more sense once it is acknowledged that depth information is not specified by image signals. In fact, the studies that found lower perceptual performance are mostly based on judgments of perceived depth from shading (Bülthoff & Mallot, 1990). Moreover, the error obtained by reconstructing the surface from judgments other than direct estimation of depth is smaller (Koenderink et al., 1996). Different authors in fact suggest that shading is not actually a cue to depth (Bülthoff & Mallot, 1990; Mamassian et al., 1996). Depth information cannot be obtained directly, but must be obtained by integration of surface orientation, a computation very sensitive to noise in the measurement of the image signals. So, as we saw in the section about the image formation, if slant is perceived accurately, the errors in the reconstruction of the perceived shape should be low. On the contrary, investigations requiring the judgment of local orientation through a direct estimate of slant and tilt in degrees (Mingolla & Todd, 1986) or by using gauge figure on the image (Koenderink, Doorn, & Kappers, 1992) proven that the orientation is perceived consistently (small variability), but still quite inaccurately (large bias). The same result is obtained by Bülthoff and Mallot (1990), which found that shape comparisons lead to a more consistent results, but where the shape is still underestimated.

A. Johnston and Passmore (1994a) suggest that an estimate of depth from a shaded image requires evaluating attributes that are not explicit to the visual system. They found, as support to this, that the Weber fractions for curvature discrimination are comparable to the one obtainable with stereo information, whereas for the discrimination of slant such threshold increases of a factor of 10 (A. Johnston & Passmore, 1994c, 1994b). This result indicates how the property investigated and the task is significantly affecting performance as we will see in more detail in the next chapter. The authors suggest that there are two ways of estimating curvature (A. Johnston & Passmore, 1994a). One possibility is to encode curvature directly from image intensities. The second is to first encode surface orientation and depth and then computing curvature by operating on this information (Carman & Welch, 1992). Their data supports the first of the two hypothesis, the visual system can estimate local shape from image measurement without relying on estimates of slant.

Studies involving the estimate of shape were conducted mainly by investigating shape index and curvedness (Koenderink, 1990; Mamassian et al., 1996). The results indicate that there is a bias to perceive shapes as convex elliptic (Erens, Kappers, & Koenderink, 1993; Mamassian et al., 1996). A small Weber fraction was reported in curvature discrimination tasks (as low as 0.1, A. Johnston & Passmore, 1994a) indicating high consistency. Todd and Mingolla (1983) assert that perceived magnitude of the curvature is underestimated and there is a small correlation between simulated and perceived curvature. However, even though the researchers assert that these results indicate that shading does not provide information about curvature, the actual task was to choose among five profiles of a bump that depict the cross section of the simulated surface. With this task it is in fact possible that participants did not base their judgments on curvature, but on the amount of protrusion, therefore assessing the depth extension of the surface.

Different factors influence the perception of shape from shading information. Illumination conditions and reflectance properties of the surface significantly influence the amount of curvature perceived. Lambertian surfaces lead to underestimation whereas shiny ones generate an overestimation (Todd & Mingolla, 1983; Bülthoff & Mallot, 1990).

Mingolla and Todd (1986) found that perceived orientation does not change when the illuminant is more oblique. Other investigations, in contrast, found that the perceived shape exhibits a systematic bias in the direction of the light source (Christou, Koenderink, & Doorn, 1996; Koenderink et al., 1992, 1996). Perceived slant changes in the direction of the illuminant by a magnitude of 4 degrees (Pentland, 1982; Todd & Mingolla, 1983; Koenderink et al., 1996). This bias is consistent with a regression to the gradient of luminance that is reduced for high values of albedo. Curran and Johnston (1994) found that the discrimination threshold for slant increases with elevation of the light source, whereas the threshold for curvature decreases. In a subsequent study (Curran & Johnston, 1996) they also found that the magnitude of the perceived curvature depends on the angle of illumination. Higher values of perceived curvature are found when the light source is directly above the surface. Deviation from this point cause a decrement of perceived curvature. The authors interpret their results to be in accordance with the assumption of “light from above”. The visual system, in fact, assumes that the light is coming from above in the interpretation of shading patterns (Gibson, 1950; Yonas, 1979; Ramachandran, 1988) unless sufficient information in the stimulus indicates the illuminant position. Explicit reports of the estimates of the direction of illumination for very simple objects is very accurate (Pentland, 1982; Todd & Mingolla, 1983). However, with complex surfaces, the performance decreases significantly and there is no correlation between the estimation illumination direction and perceived surface orientation (Mingolla & Todd, 1986).

2.4 Summary

In this chapter three cues have been analyzed in terms of the relation between geometric properties of shape and image signals. The optic array provides specific information about the shape property with an accuracy that is related to the noise in the measurement. Other properties can be computed, but their estimate will be

	Motion	Texture	Shading
Depth $[mm]$	$\kappa V \left(= \sqrt{\frac{x}{accel(*)}} V\right)$	$(\kappa M(*))$	only by integration
Slant $[^o]$	$\kappa\sqrt{V_s}$	$\arccos \frac{m}{M}$	only by integration
Curvature $[mm^-1]$	κV_{ss}	only if slant is known	κI_s

Table 2.1: Information in the retinal image that can be used as a signal to estimate the shape properties. (*) The visual system is not sensitive to this signal.

necessarily affected by a large amount of noise. The relation between each property and the relative image signals is summarized in table 2.1 (κ is an unknown parameter). This list indicates the image signals that can be measured for the perception of shape. It is important to underline different aspects of these equations.

First, if the appropriate signal can be detected and measured by the visual system it is possible to determine the shape property that is related to it. In this case the estimate of the shape property can be made just by detecting the local pattern in the image. Estimates that are not achieved in this way can be computed by integration or derivation of other measurements across the image. However, the computation is more prone to error and subject to noise as it will be shown in Section 3.5 and Chapter 4.

Second, if the equation contains the multiplication factor κ, then the visual system can estimate the shape parameter only up to this factor. There are three ways the visual system makes an estimated of this parameter: either by making an assumption about what the value is, by computing the value of the parameter using non-local information, or by estimating the parameter from information available in other cues.

Whereas the first possibility is the one classically embraced by constructivist theories of perception (Rock, 1984), the second possibility (non local computation) has received support from my own investigation (Di Luca et al., 2004; Di Luca, Domini, & Caudek, submitted) and the third possibility corresponds to the concept of promotion discussed in the Section 3.1.

Third, the more the equation approximates a linear function, the more accurate the perception of the property will be. In fact, if the equation is non linear with respect to the image measurement, the estimate will contain a statistical bias that is a function of the amount of noise in the measurement (see Daniilidis & Spetsakis, 1996). In fact, if the shape property p is related to the image measurement i by a linear function $p = ki$, the influence of noise in the measurement can be represented by the formula $i = m + \delta_m$ where the random variable δ_m has mean $E(\delta_m) = 0$ and variability $V(\delta_m)$. This variable has an influence on the variability of the estimate $V(p) = kV(\delta_m)$ without affecting the expected value $E(p) = km + kE(\delta_m) = km$. However, if the mapping is not linear, i.e. it is quadratic as in $p = ki^2$, the same noise term modifies the variability $V(p) = 2kmV(\delta_m) + kV(\delta_m)^2$ but also the expected value by an amount that depends on the variability of the random variable $E(p) = km^2 + kV(\delta_m)^2$.

Fourth, higher order derivatives of a signal are more affected by noise than lower order ones. The estimate of properties based on lower order image signals would be more precise than the one based on higher order estimates.

The idea that all the information provided by the cue is expressed in terms of depth and that cues are equivalent is classically incorporated in many approaches of 3D shape perception and computer vision more or less explicitly. From the analysis of cues provided in this Chapter and summarized in Table 2.1 it is clear that cues do not provide interchangeable information about the shape of surface. Each cue provides information that is qualitatively different, so the classic assumption of equivalent estimation of depth is fundamentally flawed. Motion has lower level of noise in the estimate of depth than in the estimates of slant and curvature. Texture specifies slant and curvature when slant is known (therefore wit less accuracy), depth can only be obtained by integration. Shading specifies curvature, and other properties must be obtained by integration. The informativeness of each cue in a particular situation must be addressed before further analysis can be made.

The different information does not depend completely on the image formation process. The image signals are not informative only about the shape property that they are related too. Even if a signal can be described in terms of the shape property that generates it, it is not necessarily true that this mapping can be reversed. For example, even if shading information is created by the orientation of the surface at each point, we demonstrated that curvature is the shape property that can be recovered more easily, not slant.

Only an analysis like the one provided in this chapter allows to specify what are the relations between cues and properties. In the investigation of shape perception this analysis has been frequently overlooked or underestimated. In the next chapter it will be shown how this analysis can help in the understanding of the combination of cues by the visual system.

Chapter 3

Cue Combination

> “I had made the mistake of thinking that the experience of the layout of the environment could be compounded of all the optical slants of each piece of surface. . . Convexities and concavities are not made up of elementary impressions of slant but are instead unitary features of the layout” (Gibson, 1979, p166)

From the pioneering work of Marr (1982) most research has considered cues as separate sources of information that are processed separately in the visual system (see Landy, Maloney, Johnston, & Young, 1995). It is assumed that different modules compute depth independently from each cue and the result of the computation is a depth map of the surface at each point of the image. Once the depth map is computed for one cue it is merged with the ones obtained from different cues to obtain a common final estimate. The process of joining information obtained from different cues is called cue combination.

There are many ways to combine information. Clark and Yuille (1990) defined a general dichotomy to classify the different possibility. Weakly coupled cue combination is where the estimate of one module does not affect the computation, nor the output, of any other module. Each module must be able to provide a unique veridical solution from the cue to which it is devoted. The purpose of cue combination is to reduce the noise in the final estimate by averaging the redundant estimates of the depth map. Each estimate is weighted depending on the relative reliability of the cue.

“Strongly coupled” cue combination means that the results of the computation of one module interacts with other modules, usually by altering the influence of constraints or assumptions necessary for the computation of the estimate. The outputs of the modules are interdependent, because the computation that a one module performs based on one cue can be affected by the result of others. Strong coupling not only reduces the noise in the estimate can be computed even in the case that information is underconstrained. In this case the assumptions needed for the computation are derived from the interaction of the modules.

This framework has been extended to visual perception. The two categories, weak and strong, define a continuum on which each proposed approach falls. At one end the models emphasize modularity, the estimates are recovered independently and the estimates are linearly combined. These models are called "weak fusion" (Clark & Yuille, 1990), "weak observer" (Landy et al., 1995) or "additive models of perception" (Massaro & Cohen, 1993). Every module is dedicated to the analysis of only one cue and produces an estimate that can be compared to the others (Maloney & Landy, 1989). That is each module computes an estimate of the same property. The rule of combination used is linear, like an algebraic sum (Maloney & Landy, 1989) or an average (Taylor, 1962). The weights given to each estimate change relatively slowly, and depend on the relative reliability of the cues (Jacobs & Fine, 1999; Backus & Banks, 1999; Ernst & Banks, 2002; J. M. Hillis, Banks, & Landy, 2002).

At the other end of the continuum is the "strong observer", where the computation is not divided in separate modules. Cues can influence each other's interpretation and the emphases of the model is posed on the holistic processing of information rather than on its separability. The rule of combination is completely unrestricted; estimates are not divided corresponding to different cues. Such an approach is consistent with the model proposed by (Nakayama & Shimojo, 1992) where the observer simply chooses a scene interpretation that maximizes the probability of the image.

Landy et al. (1995) emphasized that the weak and the strong observer are only theoretical formulations that represent the two extremes of a continuum. Real models occupy a position that is determined by the amount of freedom in the interaction between the modules, e.g. the quantitative influence a module has in changing qualitatively the output of other modules. The most accepted model of cue combination, the modified weak fusion (MWF) model proposed by Landy et al. (1995), lies somewhere in the middle because the structure is modular, but there are limited interactions between cues.

Domini, Caudek, and Tassinari (2006) recently proposed a different approach called the Intrinsic Constraint (IC) Model because it stresses the importance of constraints between image signals for the estimate of shape. The fundamental observation of the approach is that image measurement are interdependent, and therefore the computation is not actually separated in modules devoted to a cue, but the solution is computed inter-independently.

We will cover the relevant detail of these approaches before considering other problem of cue combination. More importance in the analysis will be given to the issues related to the topic of this dissertation, rather than to give an exhaustive description of the theories.

3.1 Modified Weak Fusion

Landy et al. (1995) proposed a different model of cue combination based on a linear combination rule, but with many difference from strictly weakly coupled fusion approaches (see also Kersten & Yuille, 2003; Young, Landy, & Maloney, 1993). The model they proposed received great empirical support and now it is considered to be the most comprehensive model of cue combination.

The authors believed that the modularity of the system is very important and that the final linear interaction among estimates is essentially correct. They wanted, however, to place the accent on the different kinds of information provided by various cues and how one can consider such differences while integrating the estimates from different modules. The authors (see also Maloney & Landy, 1989; Jacobs, 2002) state that cues in the retinal image are differently "meaningful" (see S. S. Stevens, 1959). Each source provides information measured in a particular scale type that is qualitatively different from other types. The integration process has to somehow account for these differences.

In order to account for the different type of information that the cues provide, the authors proposed that the visual system makes each estimate to be “commensurable” with the others by transforming it to a common scale (scale convergence Birnbaum, 1983). Cutting and Vishton (1995), in an early account of this problem, assumed that all sources were reduced to a lower-order description, namely to an ordinal representation of the world (see also Cutting & Bruno, 1988). However, Landy et al. (1995) stated that instead of reducing the solution, the visual system "promotes" the status of the cues by making them all sources of absolute depth information once a number of unknown parameters have been specified Maloney and Landy (1989). The output of the module could, therefore, be thought of as a "depth-map-with-parameters" [1].

Promotion is a type of interaction between cues where each module provides the others with its current incomplete depth map. The information coming from other modules is used to compute the missing parameters needed for the interpretation. Each module could, therefore, fill in the missing parameters using incomplete output from the other modules. Using the promotion mechanism, two cues that are present in the same location of the retinal image can be made commensurable if the one with lower status is promoted to the value of the higher one. The two cues can interact in different manners (for an exhaustive description see Landy & Brenner, 2001):

- First, an absolute cue can provide the missing parameter for a lower order cue that must be promoted if they are both present in some areas. For example, stereo information which provide absolute surface distances, can be used to promote occlusion to a higher order.

- A second possibility for promotion is when cues have different scaling behavior.

[1] as advanced by Shopenhaurer (1847) and present in Gogel's formulation as well (Gogel & Tietz, 1977).

Cues of this type provide a metric description only once the viewing distance has been determined. The visual system has to use information from a qualitatively different cue to define the missing parameters (E. B. Johnston, Cumming, & Landy, 1994) usually it does so by deriving the viewing distance, or the motion of the object. Richards (1985), for example, developed a method to obtain the correct viewing distance parameter from motion and stereo by equating the shape specified by the two cues. Since the cues scale differently with distance (depth from motion scales linearly whereas disparity scales quadratically) the correct viewing distance is the one and only value at which the estimates of the modules are in agreement.

- A third possibility is if two sets of scene estimates are combined to mutually constrain the assumptions needed for their interpretation. A possible example is the interaction in the interpretation of shading and texture. While shading provides curvature estimates once the light source parameters have been individuated, texture needs constraints on the distribution and composition of the texels on the surface. The two cues could solve the problem of interpretation by combining the information they provide.

- A fourth and final possibility is when a cue promotes itself if two sets of data are gathered. For example the interpretation can be mutually constrained from different parts of the scene or from multiple views of the same object.

All these instances of promotion are examples of the same computation: the determination of the scaling parameter for the interpretation of depth. Once this missing value is specified the cues can be combined using simple rule of combination. It is important to spell out that for the MWF the information sharing can be used only for purposes of promotion, so to determine the depth scaling. The output of each module once these parameters have been obtained is a depth-map and a map of estimated reliability. These two maps for each of the cues are feeded to the actual combinatory stage.

Many results are consistent with the modified weak fusion scheme. Non-linear interactions between cues as the ones incorporated in this scheme have often been reported.

There are results consistent with the presence of a hierarchy of cues and a selection of the best suited for interpreting the observed scene. Schwartz and Sperling (1983)found that linear perspective does not provide useful information for such disambiguation when coupled with proximity-luminance covariance (PLC) which in turn is found to override geometric cues of form and motion. Prazdny (1986) suggested that kinetic information vetoes stereoscopic disparity information. He showed subjects a sequence in which the motion was consistent with a rotating three-dimensional shape but the disparity information was consistent with a flat disk in front of a background.

Many results are consistent with the view that one cue disambiguates the depth derived by another. Braunstein, Anderson, and Riefer (1982), for example, found that occlusion information in a motion sequence disambiguates the sign of depth derived from kinetic information. Proffitt, Bertenthal, and Roberts (1984) obtained a similar result investigating multistability in point-light walkers stimuli, indicating that occlusion reduces it. Braunstein, Andersen, Rouse, and Tittle (1986) found that stereo disparity disambiguates the sign of perceived depth from motion information. B. J. Rogers and Collett (1989)reported a similar finding for stimuli containing motion parallax and stereo disparity by asking the subjects to estimate depth of corrugated surfaces. They evaluated the combination rule using a matching technique with different magnitudes of the two cues and found that observers minimize the discrepancies between depth signaled by stereo disparity and parallax transformation by reducing the amount of rotation required by the interpretation of the scene. This result is inconsistent with a linear combination strategy but is plausible under promotion. Blake and Bülthoff (1990)showed that the disparity information provided by the presence of a highlight on a shiny curved surface can disambiguate the sign of perceived depth from shading information across the surface.

Another point that this model does, is the definition of cues to flatness. Bülthoff (1991) used a shape matching task with texture and shading information. He found that their combination could be modeled by a ”strong coupling” citeB1991because perceived depth increases with the availability of more cues. Landy et al. (1995)revises this interpretation by saying that the absence of a cue is not the same as the presence of a cue that indicates a flat object. Many other cues could be present signaling a flat display, so by increasing the number of cues available, the cues to flatness receive less weight and the apparent additivity is predictable.

A different set of findings is consistent with the presence of a promotion mechanism for cues that have a reliability that depends on distance (distance scaling). E. B. Johnston et al. (1994) suggested that motion and stereoscopic disparity interact to improve judged distance and solve the scaling problem when subjects were asked to choose which of a set of cylinders appeared circular (apparently circular cylinder). The results indicate that perceived shape was determined by both motion and disparity, whereas distance and size where affected only by the disparity cue. Tittle, Todd, Perotti, and Norman (1995) found that binocular disparities and motion didn't always result in veridical estimation of shape. The results with the cylindrical task indicate that the shape of an object is dependent on both its orientation and viewing distance. The authors interpret this result as providing support for the idea that the perceptual solution is achieved through an Affine representation.

Brenner and Damme (1999)used an adjustment task for the size and depth of an ellipsoid (which had to be matched to a tennis ball) and a reaching task for its perceived distance. The inclusion of rotation information for the stimulus improves perceived shape, but did not influence the apparent distance or its size. On the other hand, enhancing distance cues improves the three judgments altogether. The results indicate that the three tasks could depend on different representations or on different weighting of the cues. The authors concluded that with the rotation the improvement to the measure of distance in the shape module does not transfer to other judgments.

The MWF model of cue combination is a way of overcoming the limitation of a linear combination strategy wile preserving its inherent simplicity. Interactions between cues are allowed, but fundamentally limited to promotion. This approach acknowledges that cues do not carry information about Euclidean depth. The information that cues carry can be promoted to depth by specifying some missing parameters. All the combinations between cues happen in a [2] depth map.

3.2 Intrinsic Constraint between Image Signals

The weak observer does not make use of any knowledge about the image formation in the combination of cues. Domini et al. (2006) proposed an alternative hypothesis that constraints implicit in the image signals, are used in the interpretation of cues. The various image signals are not processed independently, but they are combined *at the signal level* before a 3D interpretation is provided. Their proposal states that the magnitude of the different image signals is first scaled by the amount of noise present in the measurement. This assures a common metric between signals. In this way different signals can be represented along orthogonal axes in a multidimensional space, called the "signal space". The visual system reduces the dimensionality of the signal space to a one-dimensional manifold in a manner comparable to a statistical methods for dimensionality reduction. The authors proposed the use of Principal Component Analysis as a possible method of reducing the dimensionality of the image

[2] unique representation of

signals. This has the relevant advantage of obtaining a new signal that has a higher correlation with the depth than each of the signals in isolation. It is, in other words, the best estimate of the Affine structure of the distal object.

This model allows a more accurate analysis of the problem of cue combination. In the MWF approach, all the image signals are assumed to be related to the relative depth of the points in the optic array. Instead, the IC Model stresses the importance of a correlation between image signals in the image formation process.

Let us consider, as an example, the case of a monocularly viewed textured patch curved in depth along the horizontal dimension and illuminated by a light source. The optic array contains two types of information, the texture gradient and the shading pattern. If the object is big enough and there is a perspective effect, only one of the two cues would carry information about depth, namely the size of the texels (as stated in the analysis in Chapter 2). But let's also assume that the object is small enough that this information is not present in the image. Then none of the two cues carry information about depth unless some other knowledge is available to the viewer. Recall from the last chapter that the by measuring $\arccos(m/M)$ it is possible to obtain an estimate of slant from texture and by measuring κI_s it is possible to estimate curvature from shading. For the MWF the visual system should promote the two cues to depth maps in order to combine them. None of the cues can be expressed as a depth map without other information. The only possibility to create a depth map is to estimate the slant at every point across the surface and then reconstruct depth from these values. For shading the estimate of slant is also not available from the image signals. For both cues these processes are noisy and therefore not very reliable. Indeed in this framework, shading is considered to be not very informative.

The ICM says that the image signals of shading and texture are related, but the relation is not in terms of depth as the MWF assumes. For example, if two adjacent points in the image have known values m/M (see Section 2.2), then it is possible to predict what the value of I_s should be (Section 2.3). In fact, the different properties of shape that generate these signals are related because the orientation and curvature of a patch geometrically determines the slant of adjacent patches.

This example illustrates how image signals are mutually constrained even if they are not indicative of the same property. Similarly to what the MWF hypothesizes, cues do not specify shape until some parameters have been solved for. These parameters are necessary for estimating shape properties and not for promotion of cues as in the MWF.

These parameters might be assumed or estimated by a global process. The Bayesian approach to perception has a similar distinction, priors (assumptions) are discarded if sufficient evidence (estimate) is present in the image. In a recent analysis of structure from motion I showed (Di Luca et al., 2004) that isolated patches are perceived with a slant and angular velocity that is dependent on deformation according to the default value $\omega^* = \sigma^* = k\sqrt{def}$ (Domini & Caudek, 1999). Estimates of both slant and rotation are made from the same image signal and they cannot be disentangled. When the patch is embedded in a larger optic flow, a process of spatial integration between different areas in the retinal image changes the interpretation, so that the whole surface is perceived as moving rigidly (and therefore with the same angular rotation ω. In this case, a global process can be used to estimate the necessary parameters and disentangle slant and rotation.

3.3 The representation problem

> "The eye sees only what the mind is prepared to comprehend" Henri Bergson

One of the most common views of perception is that the visual system uses the information available in the retinal image to reconstruct the Euclidean shape of the environment. It has often been assumed that this shape is expressed in the form of a feature map (Barrow & Tenenbaum, 1978). This idea was first proposed by Craik (1943) and Gibson (1950) stating that knowledge about the 3D shape of objects can be described as a point-by-point mapping of depth and orientation for each surface within the field of view see also Gibson (1979). Feature maps are retinotopic[3] and contain measurements of a local property computed from the retinal signals (Gibson, 1979; Barrow & Tenenbaum, 1978; Todd, 2004). In these maps some type of information has to be made explicit (Marr, 1982, p10).

[3] each point in the optic array is associated with an estimated (i.e. Landy et al., 1995)

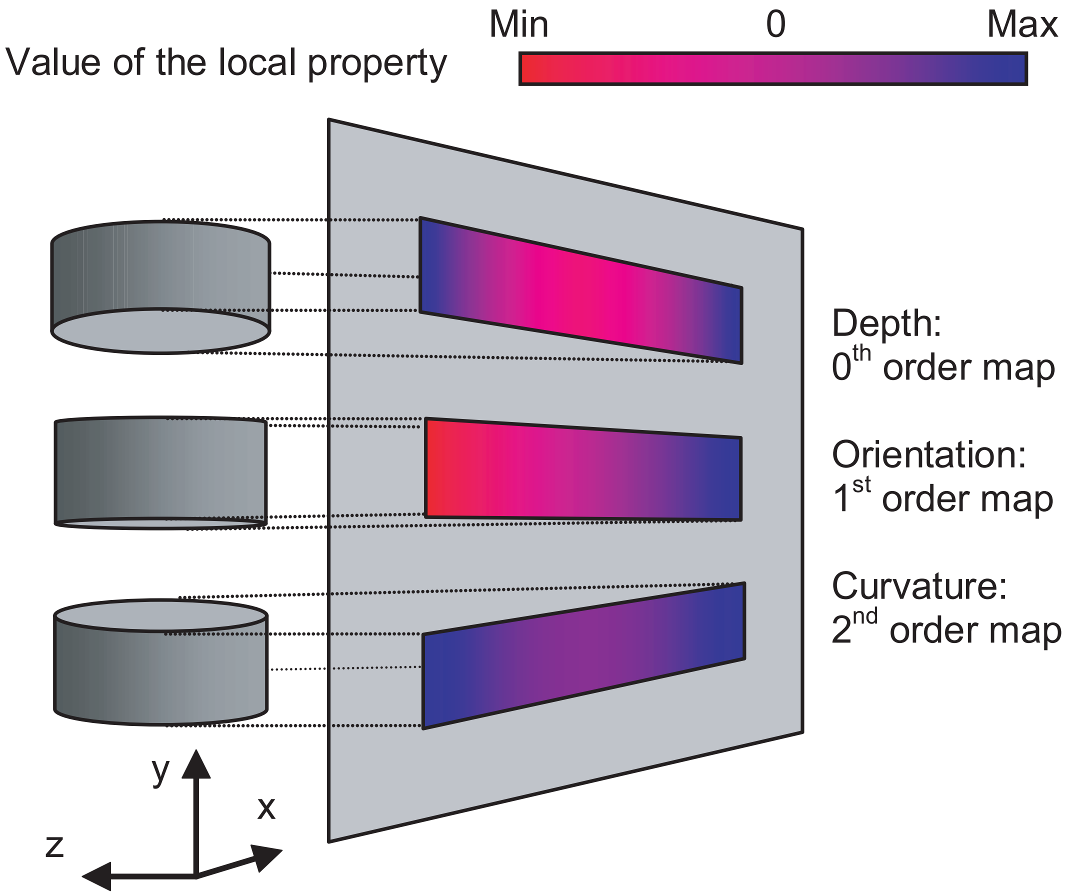


Figure 3.1: Representation of different feature maps for the same object.

Many models of perception assume the use of a feature map, although its existence is still challenged. Moreover, even if this map exists, there is still debate about which feature it is based on. It is also assumed that once one feature has been computed from the image signals it should be easy to derive other features.

According to differential geometry, feature maps based on Euclidean properties are formally equivalent. The only difference between possible maps is the property explicitly represented. If depth is used as the represented property in a feature map, it is trivial to compute the values of orientation and curvature by taking the derivative along the depth dimension. Depending on the order of the spatial derivative taken along the depth dimension (see Koenderink et al., 1992) these maps take different names: Depth-map, Slant-map, Curvature-map. Depth-maps associate relative depth, the 0th order structure, to each point on the object projection. Orientation-maps associate local orientation, the 1st order structure, to each point on the object projection. Curvature-maps associate curvature magnitude, the 2nd order structure, to each point on the object projection.

Many of the early accounts of vision were based on depth-maps, probably because they are descendants of computer vision approaches where the representation of depth is undoubtedly the easiest property to use (Barrow & Tenenbaum, 1978). There are more recent studies that favor the use of shape descriptors such as surface orientation ("2-half D sketch" Marr, 1980; Reichel, Todd, & Yilmaz, 1995; Koenderink et al., 1996), invariant landmarks (Phillips, Todd, Koenderink, & Kappers, 2003), or directional curvature (Cutting & Millard, 1984; A. Johnston & Passmore, 1994b; Curran & Johnston, 1994; Bülthoff & Mallot, 1988).

Several studies have shown that the equivalence between Euclidean representation may not be relevant to the functioning of the visual system (Koenderink et al., 1996; Todd & Bressan, 1990; Tittle et al., 1995; Fermuller, Cheong, & Aloimonos, 1997; Domini & Braunstein, 1998) While physical space is Euclidean, perceived space could be expressed in a different geometry (for a discussion see Cutting, 1986; Todd &

Norman, 2003), where geometry is a system of definitions and theorems - called invariants - which do not change under a specified group of transformations (Cutting, 1986, p65).

Visual space could be, for example hyperbolic but the amount of curvature is so small that it goes unnoticed in most circumstances (see i.e. Helmholtz, 1910). According to Todd and Bressan (1990)perceived space is Affine, where angles and lengths in different directions are not preserved. The Affine account of perceived space has received much empirical support, but there is also some evidence that it does not entirely describe visual perception (e.g. Domini & Braunstein, 1998). One of the findings indicating that perception is not based on Euclidean properties is inconsistency of results obtained with different judgments of shape properties. Koenderink et al. (1996)were among the first to notice the confusion between geometry and perception. Participants compared relative depth and estimated slant on a mannequin torso. Subsequently, the authors compared the reconstructed shapes derived from the two judgments. If the two judgments were equivalent, then the two shapes should have been identical. While the reconstruction did show inherent topological similarities a significant difference in global orientation was evident, a "change of pose [. . .] – a torsion in the waist that twists the thorax" of the mannequin used as a stimulus (Koenderink et al., 1996, p170). The reconstruction was so different that the authors concluded that perceived orientation is not based on a depth map. Using synthetic shading information A. Johnston and Passmore (1994a) obtained similar results. The measured sensitivity for curvature and orientation indicates that they are directly estimated from the retinal information rather than being derived from each other.

This empirical evidence indicates that geometrical and perceived shape are fundamentally different and that perceptual space has non-Euclidean characteristics. As a consequence, feature maps characterized by different properties are not equivalent. Tasks based on the represented feature become easier but tasks based on other features are then more difficult. Thus, the choice of a feature on which the map is based has profound consequences for the perception of the environment. The selection is limited because while some features are easy to derive from the image signals, others can not be directly derived. It would then seem logical to base the map on easily extracted features. To appreciate this point, let's consider the work of Lappin and Craft (2000). The authors stressed the importance of shape index (Koenderink, 1990, p319-324) as a representational unit because of its relation with retinal signals. There are a range of retinal signals that allow us to make a precise discrimination of the local shape of a surface without any further processing or assumption. Conversely, no similar retinal information exists for other descriptors. As we discussed in chapter 2 there is no single one-to-one mapping between image signal and curvedness (Tittle, Norman, Perotti, & Phillips, 1998). Yet our perception of the environment is not limited to shape index, we can also perceive curvedness.

Let's summarize what has been said so far to understand why this is possible.

If perceived space is Euclidean, feature maps are interchangeable because they are mathematically equivalent. If perceived space is non-Euclidean, feature maps are not equivalent and the choice of a one will affect perception. Either the perceived properties of shape are all derived from the same 'primary' description, or there exist multiple descriptions of the same object derived from the retinal signals. With a primary description, all perceived properties are consistent, but the estimate is more precise for the 'primary' property. If there are multiple descriptions, the perceived properties do not need to be consistent.

We will now look into this latter possibility, that there exist multiple descriptions of the same object based on only one geometric property of the three-dimensional shape (Tittle, Perotti, & Norman, 1997). K. A. Stevens (1995) was one of the first proponents of this theory and called these descriptions “representations”. His theory defined a task as a matter of dimensionality reduction of parallel representations. In this view, since perceived space is non-Euclidean, the representations can coexist while being mutually inconsistent (see Koenderink et al., 1996; Mausfeld, 2003). “It is not strictly necessary for any global ‘internal representation’ to exist in the first place, and if there is *one* such entity there seems to be no reason why there couldn't be *several*, perhaps unrelated ones” (Koenderink et al., 1996, p169). The existence of different representations might be due to the very nature of cues. Cues are in fact generated by some aspect of the environment and while they carry useful information about some of these properties, they do not specify completely every aspect of the environment.

The representation of space is still an open question and needs to be addressed to fully understand cue combination. The representation then serves as a medium for solving conflicting cues (e.g. Attneave, 1972). Thanks to the explicit representation of one characteristic of the world, cues that specify that characteristic can be combined directly. Cues that offer information about a property of the world are more likely to be represented in an explicit manner for that property. Interaction among competitive information can be achieved in this format without re-mapping (which in turn would decrease the accuracy of the solution). Direct mapping of cues into representation has some shortcomings because the combination depends on the task the organism has to solve (see i.e. Schrater & Kersten, 2000). We will explore this point in the next section.

There is not a clear answer to the problem of representation that we have explored in this section. The debate is particularly open on three points: the geometry of perceived space, the features explicitly represented in it, and the presence of multiple representations (or the problem of independence of perception and response K. A. Stevens, 1995).

3.4 The influence of the task in cue combination

In the classical formulation, the goal of the visual system is to create a full 3D representation of the scene with Euclidean properties that can be accessed to generate a large number of behaviors. This same type of representation was adopted by researchers in machine vision; B. K. P. Horn (1975, 1977), for example, defined shape as a local orientation map (see also Pentland, 1984; K. A. Stevens, 1995). A similar approach was adopted by Marr (1982) and Marr and Nishihara (1978). There are many findings indicating that the visual system may generate an internal representation of space whose geometry is reflected in the performance on different tasks (for example J. Foley, 1977; Loomis, Da Silva, Fujita, & Fukusima, 1992; Philbeck & Loomis, 1997; Brenner & Damme, 1999). This implies that the different tasks require different computations but it is unclear how to select the computation given a task (Todd & Bressan, 1990; Tittle et al., 1995). Langer and Bülthoff (1994) summarize different types of cues underscoring their importance in shape perception. The opposite view, that a common internal representation is not used at all and different tasks are solved in different ways by the visual system, has also been confirmed by different findings (Bradshaw, Parton, & Glennerster, 2000; Knill, 2005).

Rogers showed that different task demands involving the manipulation of the same information leads to inconsistent results (B. Rogers & Bradshaw, 1993). When subjects were asked to adjust the pattern of disparity until it appeared to be fronto-parallel, the results at different distances of observation indicate complete consistency. When the same experimental setting was used to ask the subject to judge the amplitude of a corrugation, the constancy for this task was small (Bradshaw, Glennerster,

& Rogers, 1996). Because a unique representation of the scene cannot lead to a difference in performance for the tasks, the data suggest a direct strategy. The adjustment task can be solved without an estimate of the viewing distance, however the judgment itself requires such a evaluation (see Bradshaw et al., 2000). The visual system may use the simplest possible strategy to solve any task with which is faced. (Glennerster, Rogers, & Bradshaw, 1996) proposes a hierarchy of mechanisms to perform tasks that require increasingly precise information, where the visual system would choose the lowest order one.

Bülthoff and Mallot (1990) showed how the data collected using a depth probe lead to a reconstructed shape with significant errors. The errors were reduced when the task was changed to a global depth comparison between shapes. The authors analyzed different pairs of cues using the two types of judgments and in summarizing their results they propose that the integration of cues leads to the perception of different descriptors of shape (range, shape, orientation). They state that the global orientation of an object can be recovered more easily from texture cues (Bülthoff & Mallot, 1988)while the curvature (shape) is easier with shading. Highlights have an effect only on the reported shape and not on measures made with the depth probe. The authors proposed a strong fusion scheme for the integration of the modules.

Dennett (1991) described the difference in task performance while experimenting with inverting goggles (Stratton, 1897). After adaptation to the inversion, some aspects of the world appeared to be normal, while others did not adjust and so were inconsistent. A single representation of space could not lead to this inconsistency. The best way to account for the difference in performance across tasks is to assume that the visual system uses different mechanisms in each case.

Another situation where performance on tasks is inconsistent arises for reaching responses, which only seem accurate in the presence of binocular information (Pagano & Bingham, 1998; Bingham & Pagano, 1998). In this case the task performance depends on the type of information provided. This also indicates that the visual system can obtain a veridical estimate of a property of the world if and only if it has the appropriate information for the evaluation. However, Milner and Goodale (1995) suggest a different interpretation: two separate visual pathways could lead to different solutions for purposes of perception and action. There are other findings that support this view. For example, the same visual stimulus produces accurate walking responses (Loomis et al., 1992)as opposed to inaccurate verbal judgments (Pagano & Bingham, 1998) and matching (J. F. Norman, Todd, Perotti, & Tittle, 1996). In this case, depth perception probed by different tasks (for a review see Landy et al., 1995, p402-404) produces different response patterns (Koenderink et al., 1996) indicating that many representations are used and they don't need to be consistent (Graziano, Yap, & Gross, 1994).

The results discussed in this section indicate that the analysis of a cue is mapped onto the representation that is more appropriate given the type of information that the cue carries. Other computational processes allow the conversion of one type of information in other ones. In the next section I will describe how this way of analyzing information can be applied to the cue combination problem.

3.5 Theoretic proposal

As described in Chapter 2, shape properties can be estimated directly from a local measurement of the image signals or they can be obtained indirectly by integrating many measurements across the image.

The computation of an estimate from image signals is possible only if the image signal carries information about the shape property, so that the magnitude of the signal and the magnitude of the property covary. Similar to the concept of cue validity proposed by Brunswik (1956), here the cue is lawfully connected to the shape properties. If such an image signal is present, the estimate can be obtained by a "local analysis" of image signals.

A local analysis is composed of a measurement made in a small region of the retinal image, and a computation on this measurement. The estimate of a local analysis is a single value associated with the area of the image. The measurement of image signals and the computation of an estimate are contained in a "spatial operator" a module that encapsulates the information in spatial terms (i.e. see Koenderink et al., 1992). The estimate at one point is independent from other estimates and from the image signals in neighboring regions. The processing of the image as a whole happens in parallel between these encapsulated modules. If there is no image signal that relates to the shape property, the visual system has to first measure some other image signal that is locally unrelated to the property, but that integrated across the image allows the computation of the property.

As described in Chapter 2, the estimate of the shape property is limited in precision by two factors: noise in the measurement, and noise in the computation. When the estimate is not computed from local information, the amount of noise in the computation is higher than in the local case. In fact, more than one measurement is needed for the estimate, each of which is affected by noise. These measurements are subsequently combined so that the noise term has an addictive effect. This happens because the variance of the sum of any number of mutually independent random variables is the sum of the individual variances. The expected value is instead the average of the expected values of the individual random variables.

To simply show this point, let $X_1, X_2, ...X_n$ be an independent-trials process with expected value $E(X_j) = \mu$ and variance $V(X_j) = \sigma^2 = E((X_j - \mu)^2)$. Let the sum of these random variable be $S_n = \Sigma X_i$, and their average be $A_n = \frac{S_n}{n}$. The expected values of these quantities are $E(S_n) = n\mu$ and $E(A_n) = \mu$. The variances of these quantities are $V(S_n) = n\sigma^2$ and $V(A_n) = \frac{\sigma^2}{n}$. The average of independent random variable has a lower variability than the single variables in isolation. However, the variability of the sum of the same variables has more variability than the single variables.

The estimate of a surface property with non-local analysis is affected by additive noise, so it is less reliable than local ones. The goal of perception is to obtain a precise estimate of the shape properties, so local analysis should be preferred over non-local ones. However, the non-local estimates should not be discarded completely. In order to decrease the uncertainty in the final estimate, the visual system combines the estimates from multiple sources of information when available simultaneously. In this way it takes advantage of the reduction of variability in the average of random variables described above.

According to the MWF the visual system relies more on cues that provide a precise estimate (Yuille & Bülthoff, 1994; Landy et al., 1995; Clark & Yuille, 1990). Most of the studies that adopted this view concluded that the visual system estimates the precision of the cue by measuring the variability of the estimate in different trials. But here there is an interesting conundrum, as we showed in Chapter 2 the reliability of the cues depends on the property to estimate. Thus the weight assigned to each cue should depend on the property judged in these trials.

Consider the situation where two image signals A and B provide local information about a shape property X with similar reliability. In this case, the visual system combines the estimates from the two signals weighting them equally. Now assume that only the image signals A provides information about a second property Y, while the other signal B is not related to this property. The estimate of this property Y will be more precise if based more on the image signal A rather than on B. Therefore, in the same viewing situation the weight given to a cue depends on the property judged. For example, we have seen that whereas texture cues are informative about orientation and somewhat about curvature, shading is limited to curvature. In this case, tasks in which performance is based on curvature estimates are more precise if based on both the cues. Tasks depending on orientation will be more accurate if based mostly texture, which provides information about this property and therefore is more precise. The variability on the estimate of orientation from shading information, on the other hand, will be higher and the visual system should not rely heavily on it. The variability of the combined estimate should be therefore weighted differently in the two cases in order to be optimally reduced.

The expected value of the estimate in these cases follows a similar destiny. There are many indications that cues do not provide information to make a veridical (unbiased) estimate of the shape properties. The result is that the value of the combined estimate will be closer to the more reliable cue for that particular property.

Without considering the relationship between cues and properties and without the presence of multiple representations of the object, a weighted average based on depth information as proposed by the MWF is not an appropriate strategy for integrating cues. In the next chapter I will show numerically that the estimates of shape properties are related to the type of information available to the viewer and that optimal combination of information is influenced by the property considered. The difference in the estimates of the geometric properties indicates that the visual system does not recover the Euclidean structure of the environment. Instead we conclude that there must be multiple representations of shape properties.

Chapter 4

Simulation

In the previous chapters, I described the information provided by each cue as well as my hypothesis about how the visual system might use the image signals to estimate the shape properties. To prove that in some cases the measurement of an image signal provides sufficient information to make a local estimate of a certain surface property and that the noise involved affects the estimate I ran a simulation of an ideal observer that uses the simple equations of table 2.1 to estimate geometric properties. These equations are based either on local computations or on integrals and derivatives of these signals. I did not consider any global nor symbolic processing of information.

The relationship between properties and information is extremely important in shape perception, and its analysis allows us to predict how the information provided by each cue is related to the perceptual estimate of different shape properties. This theoretical analysis does not consider the effects of noise in a quantitative manner. The noise in the measurement is different for each pair of signals and properties depending on the type of computation involved and the viewing situation.

An informative method to evaluate the effect of noise in a particular situation is to create a simulation and compute the estimates of each property by using the same conditions and strategies as the visual system. The goal of this simulation is to quantitatively verify that in viewing situations similar to the experiments, the performance of an optimal observer can be explained by the local specification of properties by the image signals. It is important to make the image signals available to the ideal observer as similar as possible to what was used in the experimental

	Motion	Texture	Shading
Depth	L+I+II	I+II	II
Slant	L+D+I	L+I	I
Curvature	L+D+DD	L+D	L

Table 4.1: The computation of shape properties can be made in different ways: (L) locally by considering only the image signal, (D) by derivation from lower order properties or (I) by integration of higher order ones.

condition and estimate the same properties as the subject are required to report. [1]

This simulation is comprised of 'information' and a 'computation' stages. The information stage is designed to create the same type of information available in the single-cue experimental viewing conditions (described in Chapter 5). Three cues are been considered here: motion, texture and shading. The computation stage simulates the computations of the ideal observer: measurement of the image signals and computation of the shape properties.

The 'information' stage involves the creation of a surface viewed at a distance of $zf = 2000mm$ with size $r = 50mm$ and stretch factor $S = 0.05$ that defines the shape $z = Sx^2$, similarly to the experimental conditions described in Chapter 5. The three types of cues used by the participants were simulated as described below. The image signals contain noise similar to the one a human observer would encounter in the same viewing condition. The simulation consisted of 50 repetitions of the same estimate using the same information perturbed by different samplings of the noise.

In the 'computation' stage, the ideal observer uses the information in the display to estimate the shape of the surface in terms of three properties: relative depth,

[1] In Chapter 5 the simulation will replicate the experimental conditions.

orientation and curvature for each point in the image. The estimates of the properties can be computed in three different ways:

- By measuring the image signal that provides local information
- By deriving a higher-order property from the estimate of a lower-order one made locally
- By numerical integration of higher-level properties along a straight line starting from the center of the surface.

Each of the cues has different combination of the three ways of computing properties. Table 4.1 summarizes how in the simulation each of the properties has been computed. Each of the table's entries is one possible way of estimating a shape property and will be considered in the following simulation. The different estimates will be kept separate and analyzed for magnitude and variability.

4.1 Motion simulation

A series of identifiable features was simulated to be on the surface, aligned with the y axis. As in the experiments, the features were points but here they were equally spaced in the vertical direction, $5mm$ from each other on the image plane. The surface was simulated as rotating around a vertical axis passing through the middle of the surface with a period of $1sec$ and an oscillation amplitude that changed according to the equalization procedure described in Section 5.2. The value of velocity for

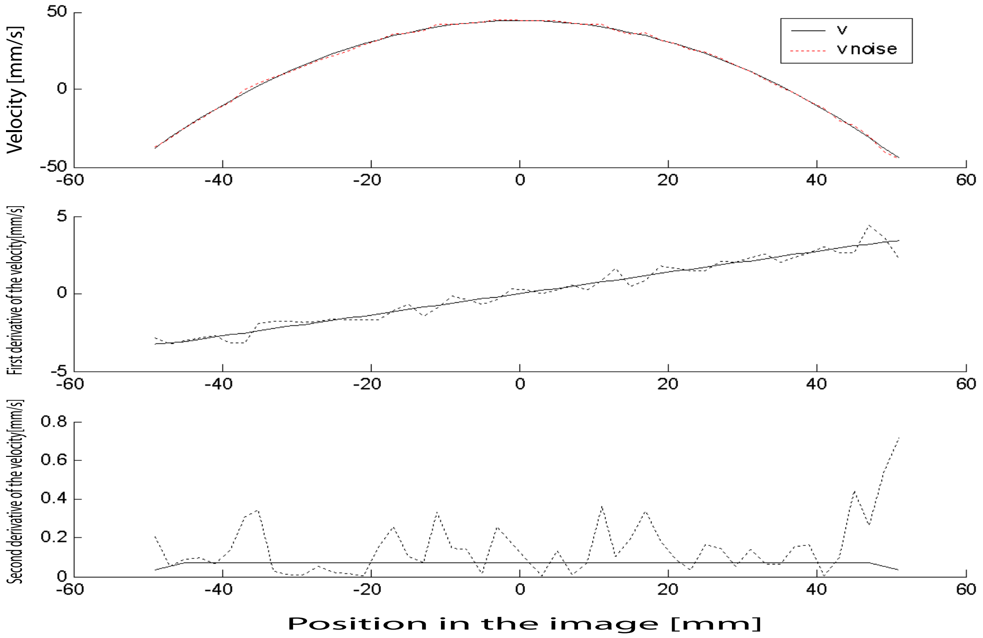


Figure 4.1: Image signals related to the motion cue (with and without noise). Top: Velocity of the surface features on the image. Middle: First derivative of the velocity. Bottom: Second derivative of the velocity.

each feature i was simulated as such: $v = (z_i - z_0) * sin(\omega)$ where ω is the angle of rotation. After some experimentation, the angle of rotation was chosen to be $\omega = 20$. The simulated observer made use of three image measurements: velocity v measured on each point, velocity gradient ∇v and second order velocity gradient $\nabla^2 v$. The simulated noise affected each signal in the same manner. For the velocity, the error in the estimation had a standard deviation $= 60 arcsec/sec$. For the other two signals, the standard deviation was calculated accordingly.

The simulated observer estimated the three properties both directly and indirectly for each point of the stimulus. Indirect properties were obtained by differentiation or integration of the direct ones. From the velocity signals, it is possible to estimate

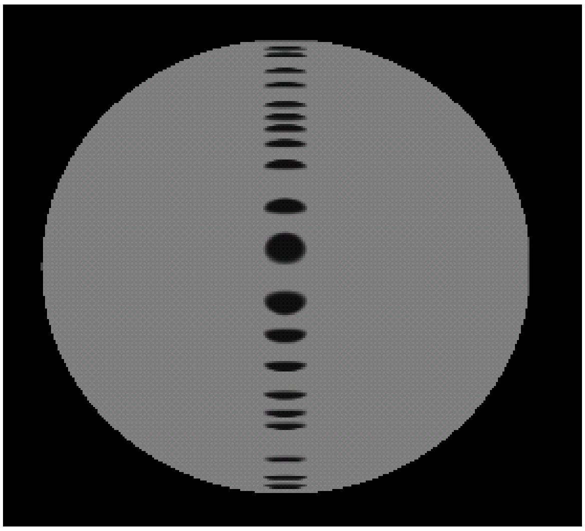


Figure 4.2: The rendering algorithm used in the experiments has been here utilized to depict how the surface would look with the texture.

depth directly, since speed correlates with depth, and then to calculate the first and second order gradients to obtain slant and curvature. For the direct estimation of slant it is possible to use the velocity gradient and then calculate the gradient to obtain curvature and the integral to calculate depth. The estimation of curvature is extracted from the second order gradient of velocity and then the two integrals are computed to obtain the estimate of slant and depth.

4.2 Texture simulation

A series of features was simulated on the surface with a spacing of $10mm$ in the image plane. The features can be thought as circular texture elements with ray $r = 5mm$ that are projected on the image plane in orthographic projection. This is slightly

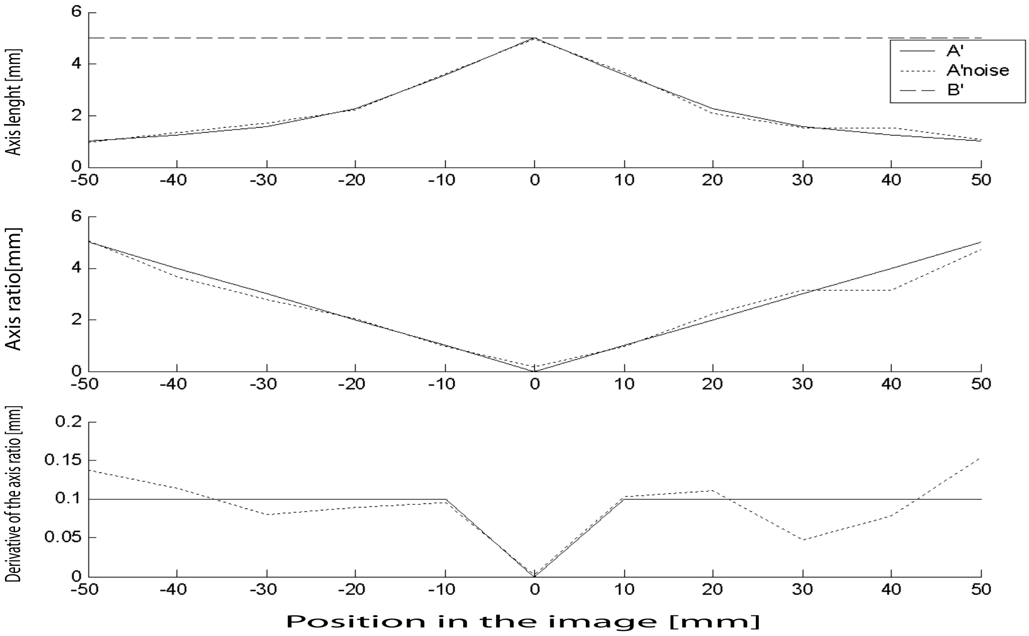


Figure 4.3: Image signals related to the texture cue with and without noise. Top: Length of the projection of the two axis. Middle: Ratio of the two axis. Bottom: Derivative of the ratio.

different from the type of stimuli used in the experiment where a volumetric texturing method was used and the features were positioned randomly across the surface. In this case, the texture information is generated only from the size of the major axes of the ellipses generated by the projection of the circles. This method was employed because of the simpler way it can be simulated.

The simulated observer measures the elongation of the axis A and B for each of the features. The minor axis A' is foreshortened according to the formula $A = rA{*}cos(\sigma^o)$, where rA is the size of the simulated feature in the direction corresponding to the minor axis that for a circular feature corresponds to r, and σ^o is the slant of the surface in the area of the feature. The major axis is simply $B = rB$, which for a circular feature corresponds to r. The measurement of these quantities was simulated to contain an error with a normal distribution and standard deviation of $sd[^o] = 5arcsec$ for each measurement so that on the image, this error resulted on a random distribution with $sd[mm] = zftan(2/3600) = 0.0485mm$. The simulated observer can adopt two strategies to reconstruct the surface using the information from the texture cue: 1) measure the values A and B for each of the features, estimate the slant, and subsequently obtain curvature estimates by derivation of the local values of slant and depth estimates by integration. 2) measure the *change* in the size of the features across the image, estimate the local curvature, and subsequently obtain slant and depth by integration. To obtain the estimate of slant, it is possible to use the length of the minor axis A that has been foreshortened if one knows what is the original dimension A_r on the surface. It is possible to make an adequate guess about this dimension by assuming that B_r is equal to A_r. This assumption is equivalent to an *isotropy assumption.* In this way, by measuring B in the image and using it as an estimate of A_r it is possible to have an estimate of σ since $A = A_r * cos\sigma[^o]$ and $tan\sigma[^o] = \frac{dz}{A} = \sigma$. In fact, since

$$\sigma' = \frac{A_r sin(\sigma[^o])}{A} = \frac{A_r\sqrt{1-(cos\sigma[^o])^2}}{A} = \frac{A_r\sqrt{1-(\frac{A}{A_r})^2}}{A} = \frac{\sqrt{A_r^2 - A^2}}{A} \tag{4.1}$$

that becomes $\sigma = \sqrt{A_r^2/A^2 - 1}$, to estimate the slant from the axis of the ellipse it is possible to simply apply the formula $\sigma = \sqrt{B^2/A^2 - 1}$ to the ratio of the two

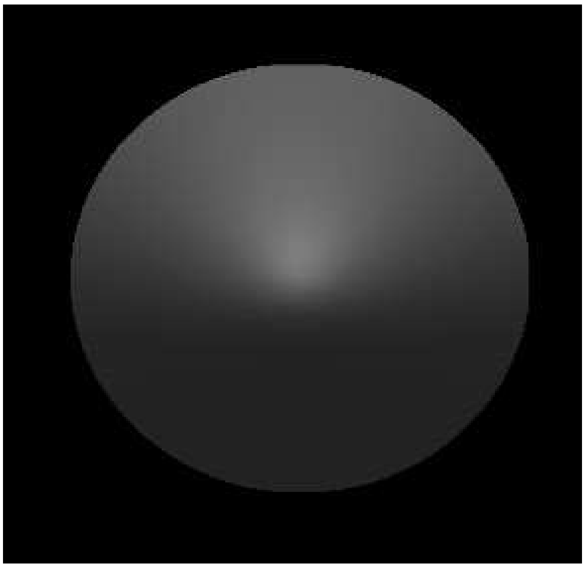


Figure 4.4: Rendering of the surface with shading cue.

orthogonal axis of the ellipse, which is defined as the *eccentricity* of the ellipse. To estimate the curvature from the change in the horizontal dimension, instead, one has to use the formula $C = \nabla\sqrt{A_r^2/A^2 - 1}$. This formula can be applied to the signals coded as the *difference* in size between features (see Section 2.2).

4.3 Shading Simulation

To create a shading cue the surface described above was simulated to have Lambertian reflectance properties and it was illuminated by a distant light source positioned 20^o above the observer. The value of the incidence angle was manipulated until the value of curvature between shading and texture cue were properly equalized (see equalization Section 5.2). From this image, the luminance values

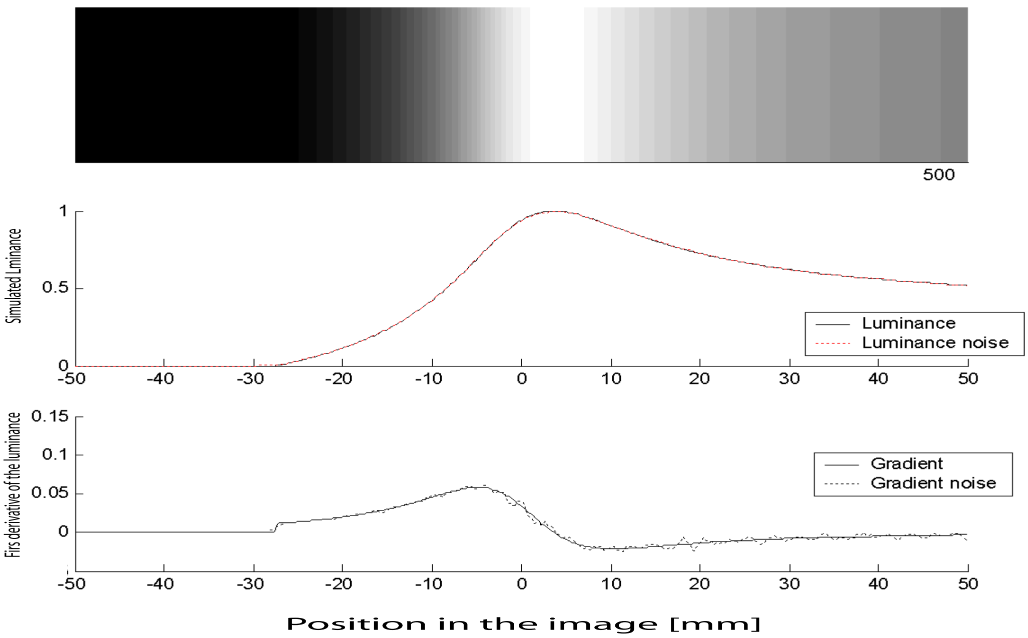


Figure 4.5: Image signals related to the shading cue with and without noise. Top: Shaded image used as a stimulus. Middle: luminance profile. Bottom: Gradient of the luminance profile.

on the vertical line and passing trough the center of the surface were analyzed. These values were quantized with a resolution of $20arcsec$. The value of luminance at every quantized point in the image can be described by the formula $I = \max\left((\tan(\sigma) * \cos(alpha) + \sin(alpha))\sqrt{(1+\tan(\sigma)^2)}, 0\right)$.

The obtained values of luminance were perturbed by a random value with a standard deviation equal to a fixed ratio of the luminance value $s.d._x = 0.01$ of I_x. To simulate the effect of retinal diffraction (see Howard, 2002), a low pass filter was applied to the perturbed signal using a zero-phase filter whose cutoff was set to be $60arcsec$. The curvature of the surface was recovered indirectly trough an iterative process. Starting from the center, it is possible to integrate the value of curvature

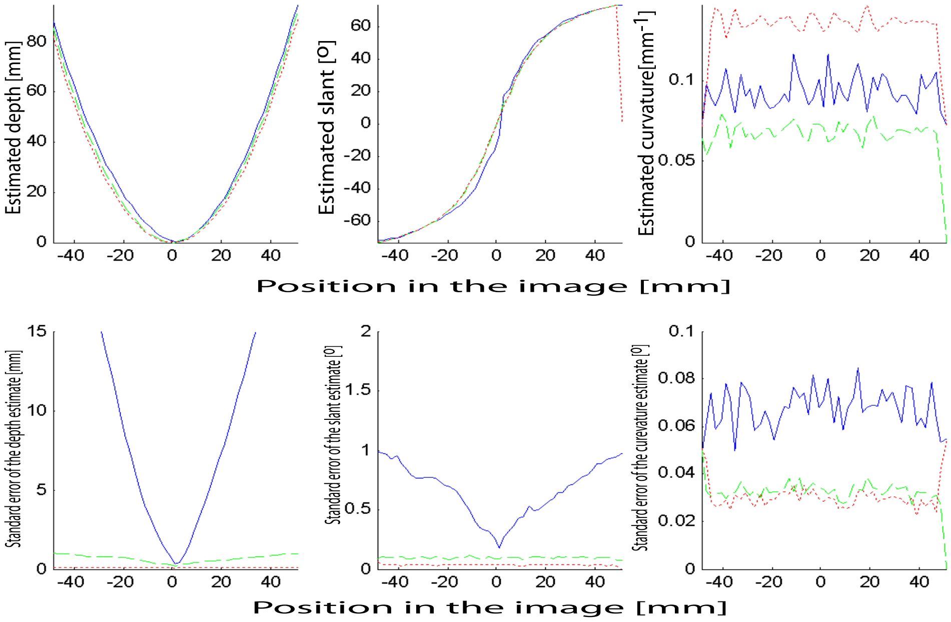


Figure 4.6: Results of the motion simulation. The different columns depict the types of information estimated. Left column: estimated depth. Middle column: estimated slant. Right column: estimated curvature. Top row: average value of the properties estimated in the different repetitions. The lines specify which information was picked up in the stimulus to estimate each property. The pattern represents the type of direct information used for the estimation, red for depth, green for slant, blue for curvature. Bottom row: error in the estimation across trials computed as the standard error of the estimate across repetitions.

to obtain slant starting with the assumption that at the center the slant is $s(0) = 0$ and curvature is $C(0) = I_x \cos(a_h)$ where a_h is the hypothesized position of the light source. The iteration proceeds by calculating the slant at each point using the formula $s(x + 1) = s(x) + C(x) * dx$ where $dx = 20arcsec$. The curvature at $x + 1$ can be subsequently calculated from the gradient of luminance Ix.

4.4 Results

The results of the simulation using the motion cue are shown in figure4.6; they indicate that curvature estimated directly is most reliable and produces overestimation for both slant and curvature. The first graph shows that the estimates of depth reproduce the shape of the simulated surface. The second graph shows that the value of slant in degrees is similar to the values simulated. The third graph shows that the estimated curvature is underestimated when computed indirectly. Estimates obtained using depth and slant information have a smaller amount of noise. Depth and slant can be estimated with small amount of error while curvature cannot. In general the estimate of curvature is more variable and does not allow for accurate estimates either directly or indirectly. The simulated observer is better off calculating curvature starting from the estimates of depth and slant.

The results of the simulation using the texture cue are shown in figure 4.7. The average values of estimated depth and slant are similar to the simulated one. Curvature, on the other hand, has a bias toward underestimation when calculated directly. This means that the shape would appear less curved in the center compared with the real shape. With the level of error chosen in the simulation, the estimates made starting from the direct estimates of slant are more reliable than the ones made using curvature. The error in the estimation of depth and curvature increases moving from the center of the object to the periphery. This pattern does not appear for slant when estimated directly. This difference in the pattern of noise is due to the fact that slant

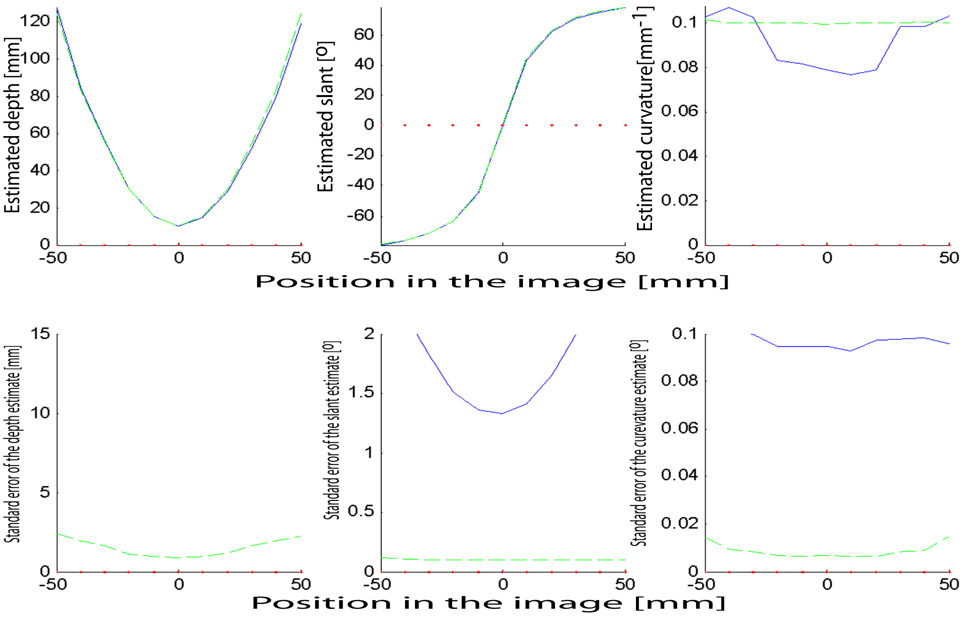


Figure 4.7: Results of the texture simulation. See caption of figure 4.6.

is estimated locally, however to estimate indirect properties the simulated observer computes an integral or a derivative. These computations increase the error when more steps are needed because the errors add up.

The results obtained with the shading cue are generally more variable and subject to more errors than the one obtained with motion and texture, as figure 4.8 shows. The only property that can be estimated directly is curvature, so the graph contains only one plot. The graphs depicting the average estimate indicates that depth is underestimated on the illuminated part of the surface when compared with the side in shadow. The estimated slant has a discontinuity in the middle of the figure. Curvature changes gradually across the surface, with a peak near the center on the shadow side, where the penumbra begins. In the most illuminated part of the surface,

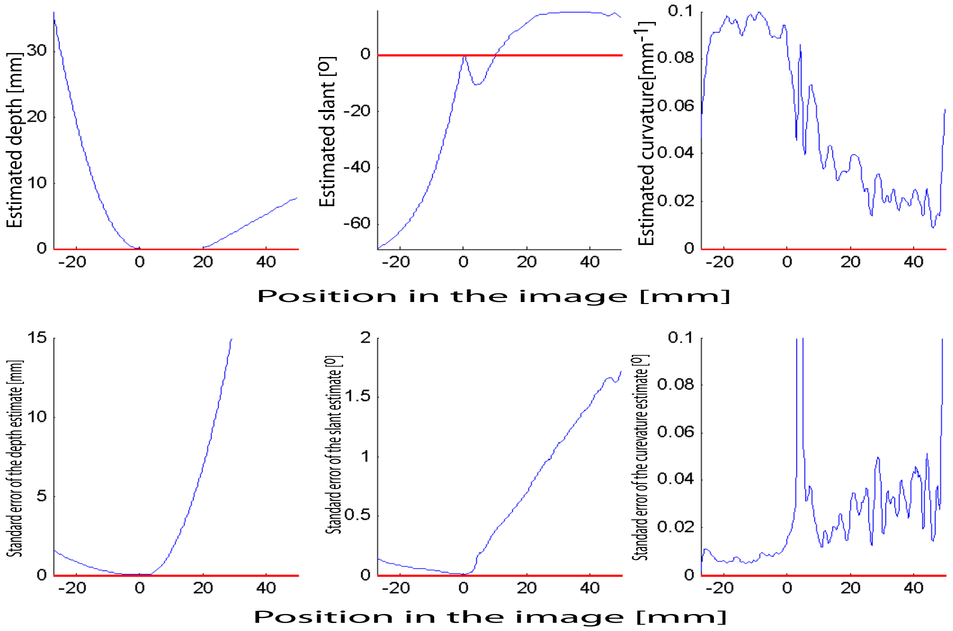


Figure 4.8: Results of the shading simulation. See caption of figure 4.6.

there is an evident decrease of perceived slant and depth. At this level of noise, the amount of noise in the estimate of curvature is comparable to what was obtained directly from the other two cues. At the same time, the error in the estimate of slant and depth rapidly increase when moving from the center of the surface toward the illuminant.

This simulation introduces noise in the measurement of the image signals, but does not consider noise in the computation of the property. In fact, if the visual system is able to estimate one property directly, the level of noise for the property is less than the one computed using its value because of neural noise. Even without this factor, the simulation already demonstrates that an optimal visual system produces direct estimates of properties that are more accurate than indirect ones. Depth and slant from the velocity signals have lower noise than curvature and depth is more reliable at the center of the shape. Slant from the texture signal is more reliable than curvature and depth especially when computed directly. Finally, depth and slant have low variability at the center, but the error increases for the parts in light whereas for the curvature the increase in noise is less significant.

Chapter 5

Experiments

“Experience is the name everyone gives to their mistakes.” Oscar Wilde, Lady Windermere’s Fan, 1892, Act III

I showed that cues carry different information regarding shape properties. For certain properties there exist an image signal that can be detected and used to make an estimate. For other properties, there is no signal that can serve this purpose and the visual system can only rely on an indirect computation of the property. Here I want to demonstrate that the visual system does not consider cues to be equivalent as often it is assumed.

If cues are not equivalent in specifying different properties, a task that based on a specified property should provide different performance than the one based on other properties. A single shape will be judged differently depending on the task. Two shapes perceptually equal for one property are not necessarily equal for other properties. In the first experiment I will verify this prediction. I will first equalize two cues in one property and then compare the shapes for other properties.

If cues are perceived differently depending on the property judged, what happen when cues are contemporary present on the same object? It is possible that the information provided by the cue is joined to create a representation that exploits all properties specified by the cues, or perhaps the properties of shape are analyzed in isolation and the cues are combined differently for each of the properties. The second experiment will test these possibilities. I will measure the informativeness of cues for different properties and explore the behavior of the visual system when multiple cues are present.

The third experiment is designed to test whether the perception of shape properties can be accounted for by a single shape. Judgment of properties at different points on the surface can be used to estimate the profile of the shape most coherent with these judgments. The reconstructed shape is an estimate of the property from which the judgments have been made. Differences in the shape reconstructed are an indication of the processing of information by the visual system. Similar shapes obtained with different task would indicate that properties are evaluated from a common medium. Shapes that significantly differ from each other would instead indicate that the properties of shape are analyzed independently and accessed when the task requires so. In this case, the perceived shape would be multiply analyzed and represented in the visual system.

5.1 General methods

Observers

Participants were undergraduate and graduate students from Brown University. They had normal or corrected-to-normal vision, they were naïve to the purpose of the study but some of them were familiar with psychophysical experiments. Participation was voluntary, all participants provided written consent and the undergraduate students were paid for their time. The only exception is when I participated and this is specified in the description of the experiment.

Apparatus

The stimulus displays were presented on a ViewSonic P70f color monitor controlled via a Dell Dimension 8100 with a Nvidia FX9600 graphic card. The resolution of the monitor was $1280x1024$ and the refresh rate was $60Hz$. Brightness and contrast of the monitor were reduced at 15% of the maximum. The luminance of the monitor was linearized using photodiodes and a custom procedure. The experiment and the stimuli were created and displayed using custom software that makes use of OpenGL libraries. Stimuli were rendered at $60fps$. A custom procedure for spatial calibration of the stimuli was employed.

The monitor was viewed monocularly using an eyepatch and a chin rest. The viewing distance was kept at $250cm$ to prevent the use of accommodation cues (e.g. Mather, 1997)(Watt, Akeley, Ernst, & Banks, 2005). An occlusion screen limited the portion of the monitor visible from the chin rest to a $20x20cm$ square region ($4^o35'$ of viewing angle). A series of screens was positioned in front and around the monitor to prevent the participant from viewing any internal part of the apparatus. The screen was occluded to the view until the screen was darkened to reinforce the impression that the stimuli were tangible objects inside a box. A dim light source ($5W$) positioned on the floor under the table supporting the apparatus and away from the line of sight maintained a small amount of light in the room to prevent dark adaptation.

Stimuli

The stimulus depicted a smooth convex surface of revolution with quadratic profile displayed in parallel projection. The surface was defined by the formula

$$z = S(x^2 + y^2) + z_0 \tag{5.1}$$

where S is called the "stretch factor" that determines the profile of the surface along the depth dimension, z is the simulated depth in cm behind the monitor's surface and x and y are the position on the monitor in cm and z_0 is the distance of the monitor from the participant. I will use this convention to indicate the stretch factor: S_x indicates a surface of stretch factor S containing only one cue (x), S_2 or S_{ab} indicates a surface with two cues (a and b), S_3 indicates a surface with 3 cues, $S_{3.x}^P$ indicates a surface with three cues that has the same magnitude of the property P as the surface containing the cue x. The surface contour was the same in all trials; it was a circle with radius $r_b = 5cm$ ($1^o09'$).

The surface was simulated using three independent sources of information: motion, texture and shading. In the different shapes, these sources of information were created by changing the value of the stretch factor S_x in the formula. This value changes the depth profile of the surface: low values simulate a flatter ellipsoid, and high values create an oblong shape. The cues could be showed independently, so that they could be either present or absent from the stimulus. For example, one shape might have contained shading and motion while another could have contained only texture. The details of how the stimuli were generated when every cue was presented in isolation are described below and in figure 5.1.

Motion: The surface had a simulated low albedo and there were a number of high albedo dots ($diameter = 0.1cm$). The dots were randomly distributed on the image plane and the minimum distance between the center of the dots was $0.25cm$. The motion of the contour and of the dots simulated a 1 Hz sinusoidal oscillation around a vertical axis passing through the surface at half its depth. The projected shape of the dots on the image did not change at any point of the rotation and the surface contour moved rigidly with the surface.

Texture: A volumetric texture technique was employed to generate a sculptured representation of the surface. The surface intersected a number of spheres with $0.5cm$ diameter, which center coincided with the surface. The minimum distance in 3D between the center of the spheres was $0.75cm$. Points of the surface contained in the spheres had a different simulated albedo than points outside. An antialiasing technique was used to create a smooth border between the two regions. In the experiment described in Chapter 7 the spheres did not coincide with the surface, instead their position was randomized in the volume to be carved out.

Shading: A standard Phong model with Lambertian reflectance function determined the amount of screen luminance at every point (J. D. Foley & Dam, 1983). Since the object was convex, no shadow was needed in the model. The simulated light source was positioned on a plane tilted 15^o on the right from the vertical. The angle subtended between its rays and the line of sight was varied in different conditions. The light source consisted of an array of nine parallel-rays lights spaced 5^o apart in the two directions and disposed in a square array. This extended illumination source resembles the type of lights we encounter everyday and creates a realistic appearance of the shape. Small dots with low albedo were present on the surface to make this stimulus consistent to the other two cues. These dots are also necessary to create a velocity signal with stimuli containing motion and shading that is consistent with the one obtained by motion alone.

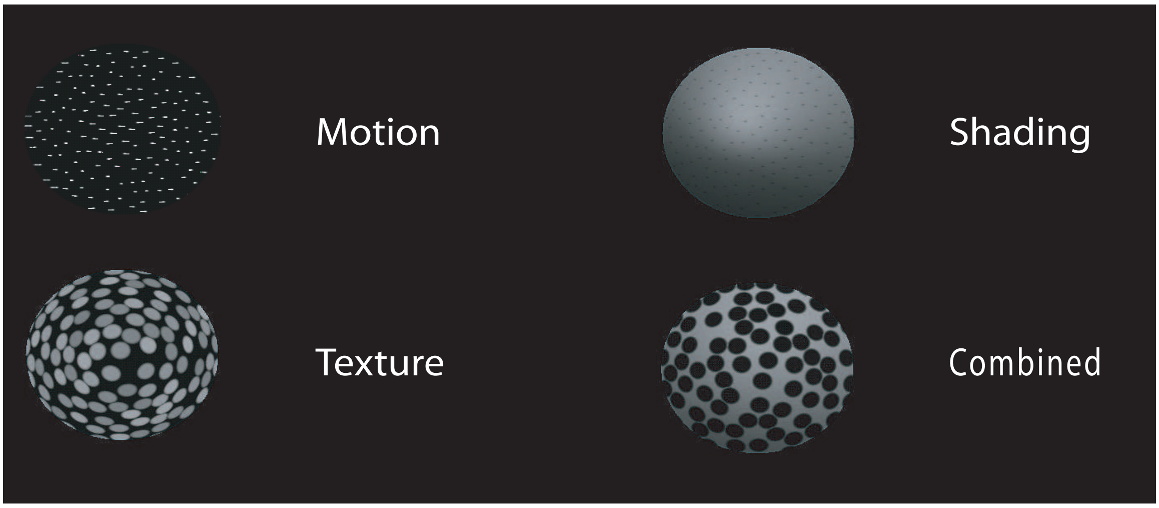


Figure 5.1: The stimulus used in the experiments with different cues.

For stimuli that contained two or three cues this scheme was slightly modified. For multi-cue stimuli that included texture, the albedo of the surface was determined only by the volumetric procedure and the small dots described above were not used. For stimuli that contained shading, the albedo of the surface was high and the albedo of the spheres and dots was low. When the shading cue was absent, the contrast was reversed so that the surface was brighter than the circle or the dots.

Procedure

All the participants were run individually. After the subject had read the instruction sheet, the experiment started with a signal-equalization session described in Section 5.2 and continued with a number of sessions timed to end after two 25-minutes blocks that contained a variable number of trials.

In each of the trials, the participant was presented with an icon indicating the type of judgment they had to make about the shape. After 300ms the icon disappeared and two stimuli appeared sequentially for 1000ms each on the two sides of the screen(except for the experiment in Chapter 7 where the stimuli were displayed continuously). The participant was asked to compare the two stimuli according to the type of judgment required by the icon. When a response was given, the next icon appeared. The three judgments required to the participant are (see figure 5.2):

- Depth, where the participant judged the elongation of the surface in depth from the tip to its base (from the center of the projected image to the bounding contour);
- Orientation, where the participant judged the slant of the surface at the near

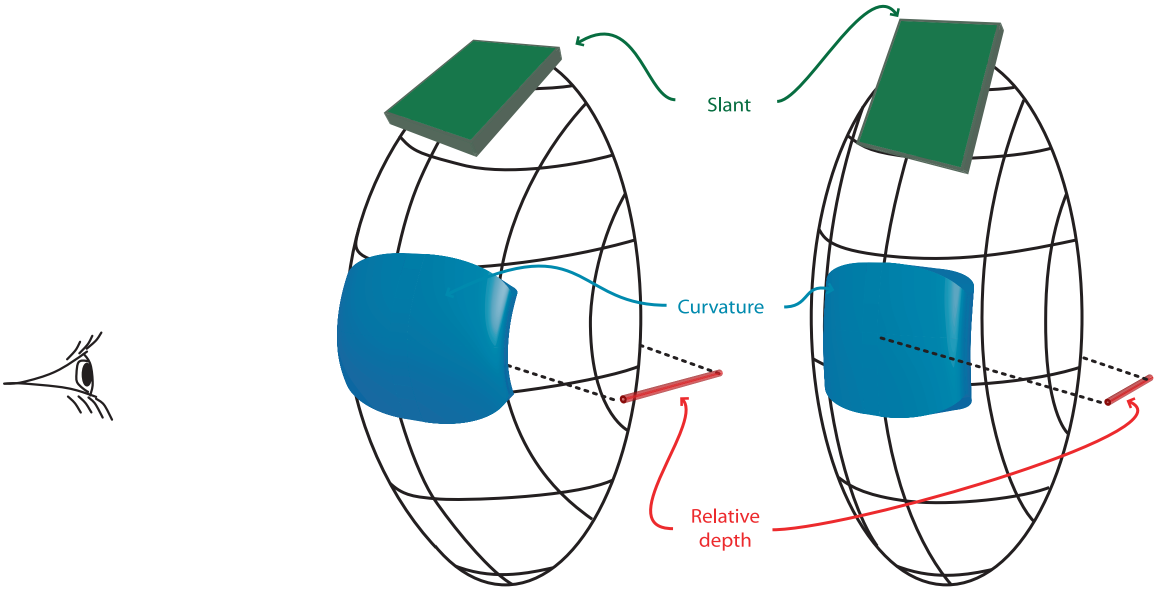


Figure 5.2: Geometric properties of the stimuli that participant judged during the experiments.

Geometric Property	Formula
Depth	$S'_D \;=\; D'[cm]/r_b^2$
Orientation	$S'_O \;=\; \tan(90 - O'[^o])/r_b$
Curvature	$S'_C \;=\; C'$

Table 5.1: Formula that related the geometric property participants judge and the stretch factor S

the bounding contour at the top of the stimulus;

- Curvature, where the participant judges the curvature of the surface at the tip (the closest point to the subject and the center of the projected image).

These judgment are based on three geometric properties of shape $\{D', O', C'\}$ whose magnitude is related to the stretch factor S in formula 5.1, as described in table 5.1.

All variables were studied within participants. Psychometric functions were fitted using psignifit version 2.5.6 (see http://bootstrap-software.org/psignifit/), a software package which implements the maximum-likelihood method described by Wichmann and Hill (2001). Confidence intervals were found by the BCa bootstrap method implemented by psignifit, based on 500 simulations.

5.2 Equalization

The goal of this procedure was to normalize the perceived magnitude of curvature of two surfaces. The participant saw two stimuli on the screen: one was defined by texture, and the other by either motion or shading. The subject was then asked to choose the stimulus that appeared to have a greater curvature at the center. The angle of rotation (for the moving object) and direction of illumination (for the shaded object) were varied as independent variables as depicted in figure 5.3. These two parameters influence significantly shape perception (Curran & Johnston, 1994, 1996; Domini & Caudek, 1999; Perotti et al., 1998) Four interleaved staircases (3-1 2-1 1-2 2-3) were employed to find the angle of illumination ϕ and rotation ω which would equate the perceived curvature for the three cues. The staircases started at 60, 50, 20, 10 degrees for the rotation and 30, 25, 10, 5 degrees for the motion. The staircase step was 5 degrees before the first reversal and 1 degree afterwards. The session stopped after 3 reversals for each staircase. If the staircases were not completed within the 25 minutes session or if the fit did not converge, the participant's data was not analyzed.

Participants compared the perceived curvature of the texture stimulus with the

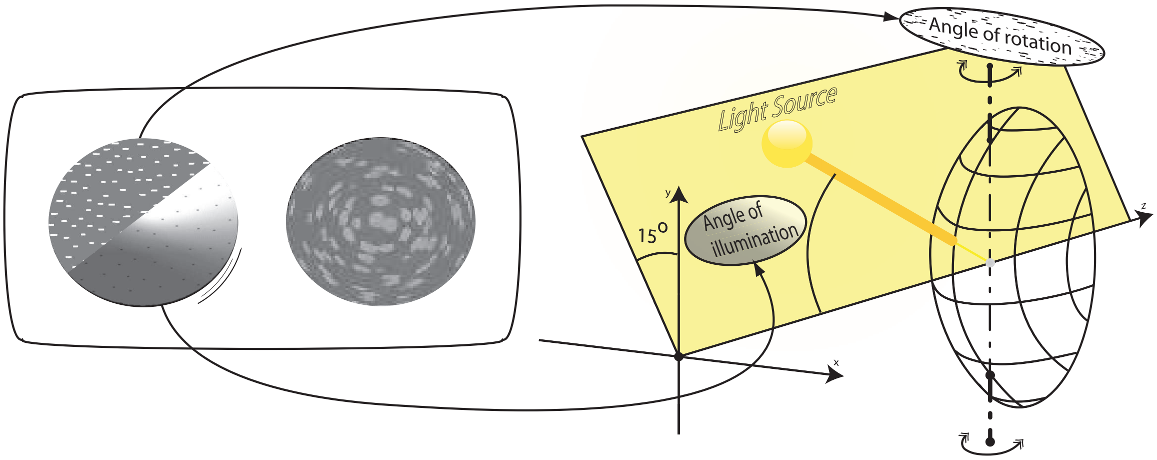


Figure 5.3: Angles used as independent variable of the staircases in the Equalization session.

perceived curvature from the motion or shading stimulus. The three simulated surfaces had the same stretch factor S ($S = 0.03$ in the first experiment, $S = 0.045$ in the second experiment, and $S = 0.2$ in the third experiment). I found the angle of rotation for the motion stimulus and the angle of illumination for the shading stimulus so that perceived curvature from motion and shading matched perceived curvature from texture. These values were then used in the rest of the experiment

Chapter 6

Experiment 1: Single-cue comparisons

In the classic accounts of shape perception like MWF, it is assumed that all cues provide the same type of information about shape. Initially, each cue is analyzed in isolation by a module that computes an estimate of shape in the form of a depth map. Subsequently, all depth maps are averaged according to the reliability of cues in order to achieve a unique representation of shape. Other geometric properties can then be computed from this single representation.

In this experiment I demonstrate that cues are differentially informative about geometric properties of shape. I manipulated the viewing condition in order to make different surfaces perceptually equivalent in one property and then compared them for other properties. If cues were equally informative about different geometric properties, the two surfaces would be perceived as equal also for other properties. If they are perceived as being different, the perception of geometric properties cannot be consistent and each aspect of shape must be extracted directly from the image signals. This result would explain why tasks that require estimation of different geometric properties yield inconsistent response patterns.

6.1 Method

The participants were ten undergraduate students naïve to the purpose of the experiment. The experiment began with the equalization session described in 5.2. For each participant the angles computed were used in the rest of the experiment. All participants completed the experiment in 2 sessions plus the equalization.

In the experimental sessions, participants compared the texture stimulus with the motion and the shading stimulus for each of the three tasks in figure 5.2. They pressed a mouse button to select the stimuli that appeared to have a larger magnitude for the property specified for the task. The two stimuli appeared sequentially for two seconds each. The shape of the texture stimuli was defined by a stretch factor $S = 0.03$. The shape of the motion and shading stimuli was changed by applying a different stretch factor S. The value of S was determined using four interleaved staircases as described for the equalization in 5.2 starting at 2.0, 1.5, 1.0, and 0.5 times the stretch of the texture stimulus. The staircase's step was 0.125 times the stretch of the texture stimulus. A 2x3 design consisted of two cue conditions (either motion or shading was compared to texture) and the three task conditions that defined the property judged by the participant (depth, slant, curvature).

6.2 Results

Participants described the stimuli as being egg-like objects or cones. Some observers reported that some shapes appeared initially to be concave but they were able to flip them without effort. Participants found the comparison of the shape properties to be easy to perform. The psychometric function obtained in the Equalization session for a sample subject is shown in figure 6.1.

The PSE for curvature between the motion and the texture stimuli was achieved

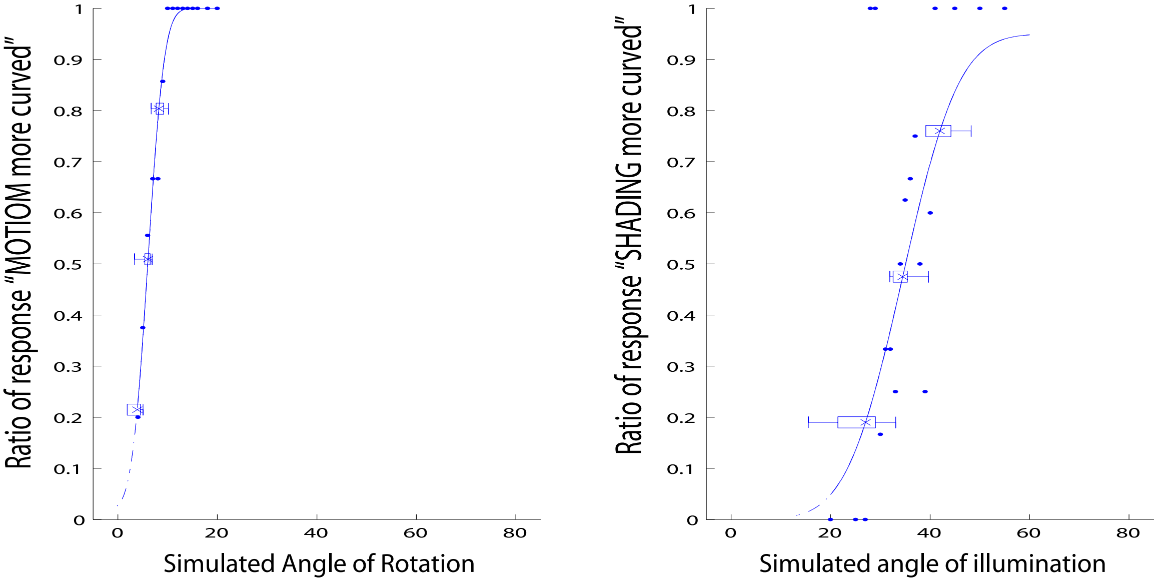


Figure 6.1: Psychometric curves in the equalization session for one participant.

with an angle of rotation that across observers was $14.5^o \pm 4.3^o (s.e.)$. For the comparison of shading and texture, the angle of illumination was $24.9318^o \pm 4.1^o (s.e.)$.

Figure 6.2 shows the result of the fit in the experimental conditions for one subject. The left graph indicates that the PSE between motion and texture is obtained with a smaller stretch of the motion stimulus than the one for judgments of slant and curvature. The higher slope of the curve for the comparison of depth indicates higher reliability for this judgment. The right graph indicates that the PSE between shading and texture is obtained with a stretch factor of the shading stimulus that is smaller for judgments of curvature than the one required for judgments of depth and slant. Slant judgments are also somewhat less accurate. The value of the PSE in the six experimental conditions and the standard deviations estimated from the psychometric curves are summarized in figure 6.3.

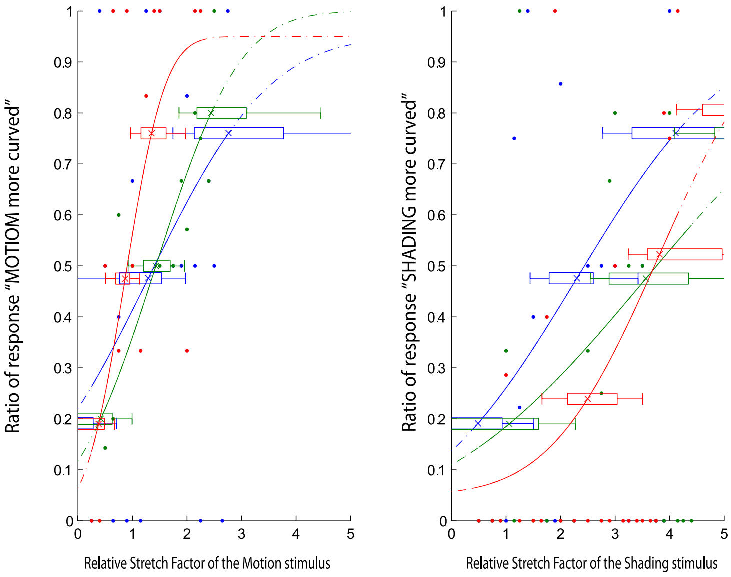


Figure 6.2: Subject 06. Left: Responses in the condition motion-texture as a function of the Stretch factor of the motion stimulus and fit with psychometric function for the three tasks. Right: same graph for the shading-texture condition.

A 2(cue: motion-texture, shading-texture) x 3(task: depth, slant, curvature) repeated-measure analysis of variance (ANOVA) on the stretch factor required to obtain the PSE revealed a main effect of cue-pair ($F(1,9) = 68.634, p < 0.001$) and a main effect of the task ($F(1,9) = 27.700, p < 0.001$). The interaction between the two independent variables was also significant ($F(2,18) =, p < 0.001$).

One tail t-test reveal that for the motion-texture pair, the stretch factor for depth was smaller than the one for curvature ($t(18) = 2.0688, p = 0.026$) and the one for slant was larger than the one for curvature ($t(18) = 1.6071, p = 0.062$). This indicates that the to perceive the same property magnitude the simulated curvature needed to be larger than curvature for slant and smaller for depth. For the shading-texture pair, the same t-test revealed that both depth and slant were larger than the curvature

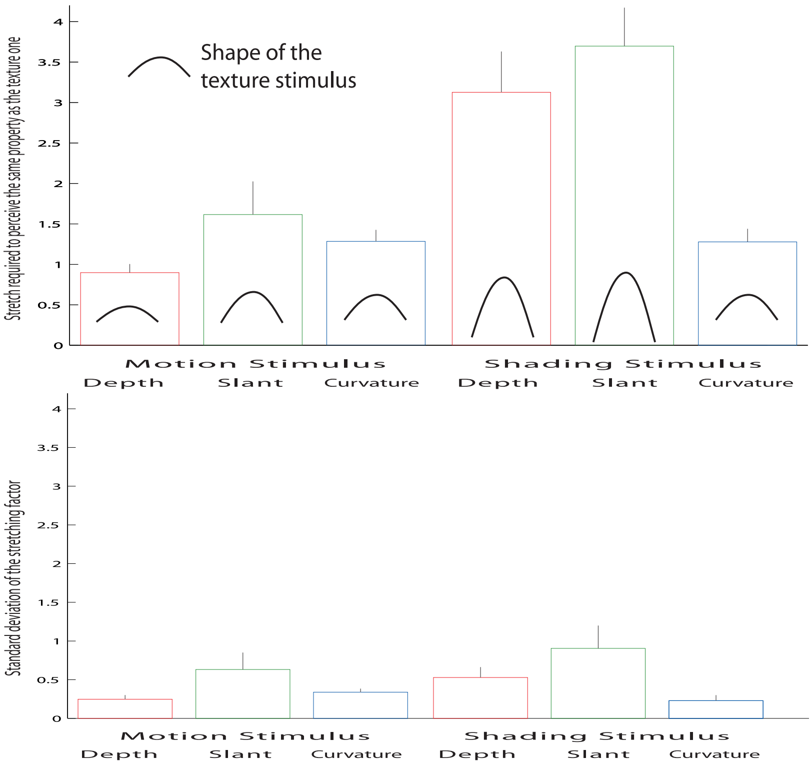


Figure 6.3: Average values obtained from the psychometric fit Top: PSE Stretch Factor. Bottom: Mean of subjects Standard Deviation estimated from the psychometric function. The error bars are the standard error of the mean across observers.

$(t(18) = -4.5712, p < 0.001; t(18) = -3.3138, p < 0.01)$. For both depth and slant comparisons, the PSE was obtained with a larger simulated shape of the shading stimulus.

The high values of the standard deviations registered in this experiment (figure 6.3) indicate a low reliability of the judgments. The ANOVA computed on the standard deviation of the stretch factor revealed a main effect of the task $(F(1,9) = 3.876, p < 0.05)$ while the cue-pair was not significant $(F(1,9) = 1.443, p = 0.260)$ and neither was the interaction $(F(2,18) = 2.052, p = 0.158)$. The single tail t-tests revealed no significant difference in standard deviation for the motion-texture conditions $(t(18) = -0.1906, p = 0.57452; t(18) = 1.0644, p = 0.1506)$ while for shading-texture the standard deviations differed significantly $(t(18) = -4.0787, p < 0.001; t(18) = -2.6201, p < 0.01)$.

The variability of participants' judgments computed as Weber fractions by dividing the standard deviations by the mean stretch factor for each stimulus and each observer is reported in figure 6.4. These values indicate a variability of the estimate that spans 25-100% of the PSE. The inter-subject variability is also conspicuous for the motion stimulus in the slant and depth judgments. The inter-subject variability is most likely due to a limited number of repetition of the measurement. The ANOVA on the Weber fraction revealed that neither the main effects $(F(1,9) = 2.983, p = 0.118; F(1,9) = 0.623, p = 0.547)$ nor the interaction $(F(2,18) = 0.464, p = 0.636)$ were significant.

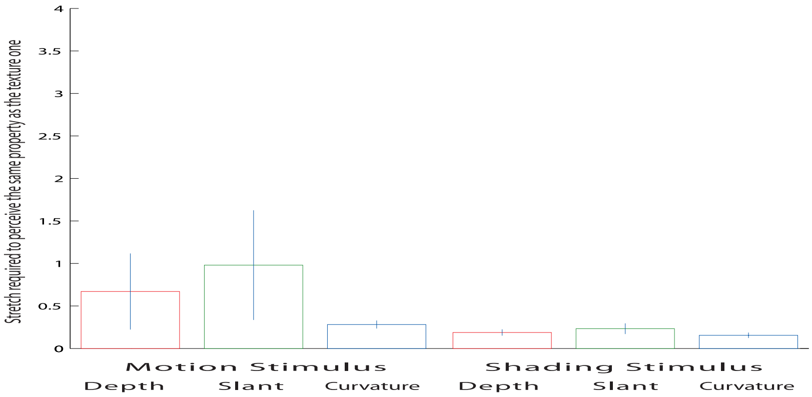


Figure 6.4: Average Weber fraction

6.3 Discussion

Classic models of perception based on a unique representation of shape predict that by equalizing curvature the perceived shape should be equal also for other properties. The representation of the shape should be either unique or equivalent with different properties. Shape comparison of single-cue stimuli should not be affected by which property is being judged. Contrarily, the results indicate that when two shapes are perceived as having the same curvature, shape comparisons for other geometric properties are still affected by the task performed. When cues differ in the type of information they provide as discussed in Chapter 2, the perceived property reflects this difference. Motion information is reliable for depth but less for slant judgments whereas texture information is reliable for slant and curvature but not for depth. Shading information is reliable for curvature but not for slant and depth, whereas texture information is reliable for slant and not reliable for curvature as shading is. When cues are not equivalent in the type of information they provide, there are two possibilities. The first possibility is that the stimulus that varies in shape contains a cue that specifies a property with a greater magnitude than texture, as in the case of motion for depth. In this case, the relative stretch factor required for the varied stimuli is less than the one required for the curvature (that has been equalized across stimuli). The motion signal specifies a larger perceived depth than the texture does and a smaller stretch factor is needed when comparing depth. The second possibility is that the varied stimulus specifies the property with lower magnitude than texture, as in the case of shading for slant. In this case, the varied stimulus needs a higher value of the property to be perceived as equal to texture. The shading signal specifies a smaller perceived slant than the texture does and a larger stretch factor is needed when comparing slant.

It is interesting to notice that when comparing two stimuli generated with different cues the property judged is extremely important. In the case of a static textured stimulus and one defined only by the motion of punctiform features, the perceived magnitude of different properties that can be used to describe 3D shape changes significantly. In this experiment it is shown that although three shapes are equalized so to have the same perceived curvature, texture is still more slanted toward the contour but less deep, and motion is more elongated in depth but less slanted.

There are three possible explanations that can be offered to explain this phenomenon. The first possibility is that judging different properties changes the way cues are interpreted in the construction of a unique representation of shape. For example, in the Bayesian account of human perception it is argued that different priors are used in the interpretation of the information when a property is judged. Although this is a possibility that cannot be ruled out by these findings, it is still not clear how the different judgments are integrated in a unique representation. In this experiment the judgments are not made on the same position so it is possible that the integration of the different properties creates a representation of shape that is not veridical and does not conform to the simulated quadratic shape. This possibility is addressed in the experiment in Chapter 7.

The second possibility is that, notwithstanding the precautions taken in the experiment, the stimuli contained cues to flatness that are not accounted for in the analysis of the signal (see Section 3.1). These cues are combined as every other type of cues with the information provided by the simulated stimuli. The weights assigned to the cues depends on the reliability of the estimate which in turn varies with the tasks (this possibility is consistent with the PSE registered). The Weber fractions, however, indicate that there is no change in performance for the two stimulus pairs in judging different properties. Although this result may be due to very large variability in the motion condition, no change in performance across the six condition indicates that there is no change in the reliability of the information provided in the display. According to the MWF and any other account based cue weighting, the weight assigned to cues must be the same because the reliability is the equivalent. If the cue to flatness do not change when changing type of stimulus, and because the weight sum to one, the weight assigned to the cue to flatness is also the same across conditions. Their influence does not change, it only decreases the magnitude of the perceived property. The classical formulation of the cues to flatness cannot account for these results. However, the possibility that cues to flatness change depending on the stimulus used cannot be ruled out. This possibility will be considered in more detail in the next experiment. Moreover, if task determines a different set of weights, perceived shape should change when switching tasks. If the judged property changes the *whole* representation of shape, there would not be a shape constancy as the task changes. Our experience of the world does not conform to a world that changes continuously.

A third explanation involves the tendency that certain shapes have to appear concave rather than convex. This possibility could explain why certain properties were be judged to have a greater magnitude. For example, depth judgments would be greatly affected by a depth reversal whereas slant judgments near the occluding contour would be more consistent. This explanation does not apply to the difference that cues have in the specification of a property. The effect should be the same for the three cues and should not invert as it does comparing motion and shading. If cues have a different tendency to appear concave *in different parts of the image* the perceived shape should appear distorted. This possibility is investigated with the experiment described in Chapter 8.

The explanation I propose is that cues provide different information for each of the properties analyzed in this experiment. The property to be judged is based on only one source of information and the information is differently biased. In other words, different image signals independently specify properties of shape and the perception of one property is independent from that of other properties. The representation of shape according to one property might be different from that based on another property, even though the signals are related by a geometric constraint. A slanted surface, for example, is necessarily related to a gradient of depth values. The two signals created by the difference in depth of points on the surface and by the orientation are therefore necessarily related by the laws of differential geometry. The measurement of these signals, however, is affected by noise and the noise has a different effect on them. According to the ICModel described in Section 3.2, for example, the perceived magnitude of a property is scaled by the noise level. Thus the difference in noise between signals generates a difference in estimation of the two properties when represented independently. The IC model's original formulation does comprises only one representation of shape. However, the heuristical processing of information that characterize its functioning allows for modifications like the simultaneous representation of different properties.

It is important to underscore that the MWF account of perception cannot be applied to this case. This model would require cues to be promoted to depth-maps using the information available from other cues. However, in this experiment the cues are presented in isolation so they cannot be be possibly promoted. This model may be reformulated to conceive cues as being informative also about other geometric properties, but this would still not explain the data collected for two reasons. First, the MWF states that variation in the estimate of a property is only due to the effect of noise and there are no biases in the modules. The data collected exhibit a constant deviation from the veridical estimates hypothesized by the model. Second, the MWF states that the goal of the visual system is to reconstruct a unique representation of shape (even if we suppose that it is not in the form of a depth map). The three properties required by the tasks are derived from this representation and therefore they are necessarily consistent. Third, the possibility that a task would weight a particular cue more than another is ruled out because the stimuli have only one cue each unless one assumes that there are cue to flatness in the display.

The results of this experiment allow us to conclude that cues are differentially informative about geometric properties of shape. Judging different properties of the same shape induces a pattern of inconsistent responses according to the relationship existing between image signal, property, and the amount of noise in the signal.

6.4 Simulation of experimental conditions

To test the proposed explanation, that the image signals provide different information for each of the properties judged by the participant, I used the method described in Chapter 4 to simulate the same experimental conditions and information. The ideal observer described uses the same type of information as the subjects would, if they

	Motion	Texture	Shading
Curvature [mm^-1]	0.086	0.084	0.085
Slant [^{o}]	74.44	66.82	34.71
Depth [mm]	77.87	39.83	15.15

Table 6.1: Averages of the estimated properties for single cue stimuli with the same viewing conditions as in the single-cue experiment.

analyzed the information independently for each property. If this simulation produces a pattern of results that is coherent with the responses given by the subjects, it is likely that the type of strategy used to compute an estimate is the same.

The simulation was conceived to compute a perceptual estimate for each task and each cue presented to the participants. A window of $2cm$ was defined on the stimulus at the center for the curvature judgment, at the top for the slant, and in both positions for the depth estimate. The average values of the estimated property were computed from the image signals created by the stimuli of the experiment.

The averaged estimated properties for 50 trials are shown in table 6.1. These are the estimated values obtained by averaging the different results for the points that fell within the window. If the cue allowed for multiple estimates, the values were averaged. For the slant, the computation was done in terms of the tangent of the angle. The value in degrees in the table was computed only afterwards.

Table 6.1 shows that once the cues are equalized in terms of curvature, the slant obtained by shading was smaller than the one obtained by motion and texture. The pattern of responses predicted from these values and shown in figure 6.5 support the interpretation hypothesized in the section above regarding the independence of the estimated properties. This pattern is very similar to the one registered experimentally ($r^2 = 0.92$). If the image signals are treated as independent sources of information for one geometric property, the projection of a single-cue shape specifies inconsistent values.

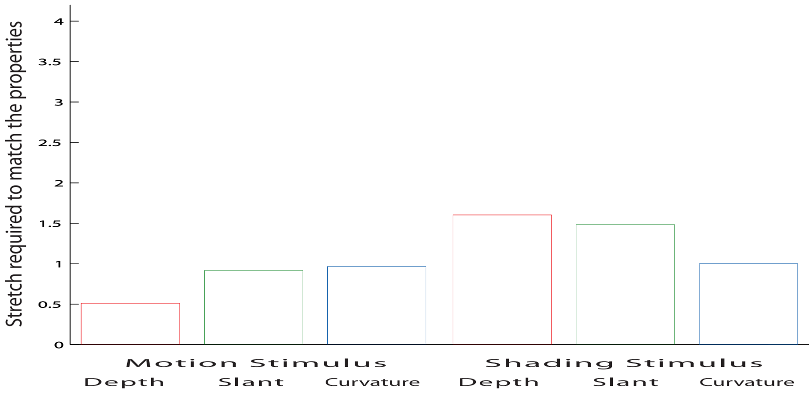


Figure 6.5: Results of the simulation of the estimated properties for single cues.

The estimated depth has a significant value for motion, an intermediate one for texture and a small value for shading. Therefore, the shaded stimulus perceived to have the same magnitude as texture needs to be equal for curvature, but larger for slant and depth as shown in figure 6.5. For motion, the stimulus needs to be flatter for depth and equal for curvature. Contrary to what has been found in the empirical data, if the simulated motion and texture shape are equal, the perceived slant is also almost equal. It is unclear why the perceived slant from motion information is larger than the prediction. It is possible that such a difference is only apparent and due to the choice of a Euclidean solution. The relative magnitude of slant and depth estimates for the motion and texture cues are opposite. A different type of noise model for one of the signals (i.e. a multiplicative noise for the motion as described in Section 2.1) with a larger amount of noise might have led to a result compatible with the human data. This hypothesis needs to be tested with a new set of simulated data.

Chapter 7

Experiment 2: Multiple cue comparisons

The visual system measures image signals and uses these values to compute the geometric properties of shape in the scene. Each of the two processes required for the estimation of a property (measurement and computation) are sources of noise that decreased precision. Both steps, measurement and computation, are imprecise because of biological limitations.

Direct estimation of a property from an image signal leads to lower noise because of the small number of steps required in the computation. On the other hand, estimates that require complex computations or that involve more than one measurement bear lower magnitude of the perceived property, even if they are derived from the very same signals used for the direct estimate of a different property as the one above.

When shape is defined by a single cue, the estimate of a property will be affected by smaller amounts of noise if that cue allows an estimation of a property with fewer steps. According to the IC Model, the signal is scaled by the amount of noise in its measurement. Therefore cues that lead to lower noise in the estimate will also produce a larger estimate of the property.

Here I want to measure this effect and be able to differentiate cues by the type of information they provide. This measurement is a direct estimate of the influence of a cue on the perception of different shape properties. Moreover, I want to test the hypothesis that cue combination is done independently for each property of shape. The information provided by a cue can be combined with what is provided by other cues, but independently for each property. If this is the case, the information carried by each cue in isolation for a property can be used to determine the perceptual solution when more than one cue is present, but only for that property.

7.1 Method

Four undergraduate students, one graduate student and I participated in the experiment. Except for myself, the other participants were naïve to the purpose of the study but all were familiar with psychophysical experiments.

All of the participants completed the experiment in five sessions plus the equalization. Sessions were composed of two 25-minutes blocks and contained a variable number of trials. The experiment started with the equalization session with a shape defined by a stretch factor $S = 0.045$.

Participants compared one target property (depth, slant, or curvature) of two stimuli, the test and the probe. The test stimulus was defined by various combinations of cues. The probe was simulated using all three cues at the same time (motion, texture and shading) and was changed in shape across trials. For each condition, a 1-up-1-down staircase procedure was used to find the amount of stretch of the probe stimulus composed by three cues S_3 that was necessary to perceive the same magnitude of the target property as in the test stimulus. The initial value of stretch S_3 used in the staircase was alternated to be higher and lower than the value of the test in the different sessions (0.5 and 1.8 times respectively).

There were 36 different conditions: six cue conditions crossed with two stretch factor conditions and the three task conditions. The cue conditions included three conditions with single-cue test stimuli and three conditions with paired-cue stimuli. In the single-cue conditions, the test stimulus was defined by only one cue (motion, texture, shading). The shape of the test stimulus was defined by a stretch ratio that was 1.00 time or 1.66 times the shape used in the equalization (stretch factor conditions). In the paired-cue conditions, the test stimulus was defined by using the three possible combinations of cue pairs (motion-texture, motion-shading, texture-shading). The two cues were generated by the same surface, with the same stretch factor, either 1.00 or 1.66 times larger than the equalization.

7.2 Results

Single cue conditions: Figure 7.1 shows the values of the stretch factor $S_{3.x}$ for which the probe stimulus and the single cue test stimulus appeared to be perceptually equal, when judging the three target properties. Lower bars indicate that the PSE between the probe and the test stimulus was obtained with smaller values of S_3, the stretching of the probe. This value is measurement indicates the magnitude of the perceived property with relation to the probe. Values closer to the relative stretch factor (1.00 and 1.66) indicate that the property of the test stimulus has been perceived veridically. Lower values indicate underestimation of the property with respect to the simulated value. The magnitude of the PSE is also an index of the information conveyed by the cue about a property.

The comparison of PSE across tasks indicates that different properties are not perceived consistently for the same stimulus. They follow the prediction I expected from the signal analysis. Motion is informative about depth, texture is informative about slant and curvature, and shading is informative about curvature. This pattern is more definite in surfaces created using higher stretch factor. The equalization worked for curvature judgments with a stretch factor of 1.00. In fact, for the curvature task and simulated stretch equal to the equalization, there is no statistical effect of cue (1 way ANOVA (cue) repeated measures, $F(2, 10) = 0.056, p = 0.946$). The equalization, however, was effective only at this level. A 2 way ANOVA (cue and stretch) has a significant main effect of stretch ratio ($F(1, 5) = 35.584, p < 0.01$) and a significant interaction cue by stretch ($F(2, 10) = 6.547, p < 0.01$).

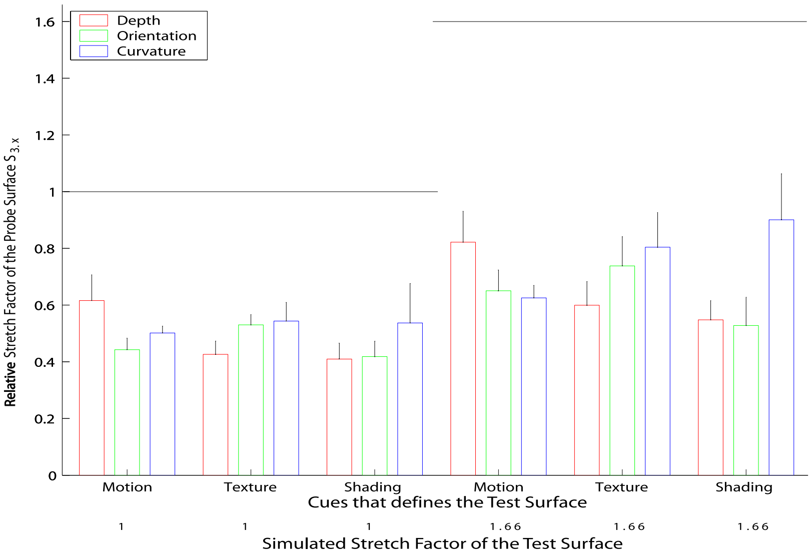


Figure 7.1: Average values obtained from the psychometric fit of the stretch factor of the probe stimulus when compared with single cues test stimuli.

A 3 way ANOVA (task, cue, simulated stretch) shows a main effect of stretch

$(F(1,5) = 26.878, p < 0.01)$ a task-stretch interaction $(t(2,10) = 8.373, p < 0.01)$ and a task-cue-stretch $(t(4,20) = 5.265, p < 0.005)$ interaction. The effect of stretch and task-stretch simply indicates that the simulated stretch influences the perceived properties. Higher values of stretch factor bear greater perceived magnitude of the property especially for one of the tasks. The three way interaction indicates that the increase in perceived magnitude property dependent also on the cue. If the cue is informative for a property according to the analysis proposed, the magnitude perceived will increase more than the non specified properties.

The stretching factor of the probe $S_{3.1}$ in the single cue conditions is significantly smaller $(t(5) = -8.3893, p < 0.001; t(5) = -14.8246, p < 0.001)$ than the one simulated for the test stimulus $S_1 = 1$ and $S_2 = 1.66$ respectively. The average values are in fact $S1' = 0.72 \pm 0.07(s.e.)$, $S2' = 1.07 \pm 0.12(s.e.)$.

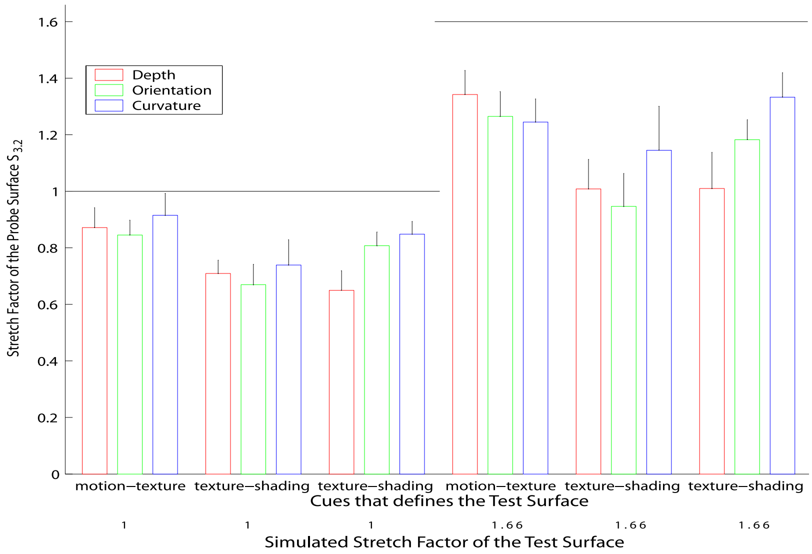


Figure 7.2: Stretch factor of the probe stimulus required for the perception of the same geometric property as the paired-cue stimulus.

Paired-cue conditions: Figure 7.2 shows the values of the stretch factor $S_{3.2}$ of the that created the same perceived property as the test. Perceived properties are not completely coherent even with stimuli composed of two cues. The perceived properties stimulus are coherent for the motion-texture stimulus, but they are not for the other two pairs. The texture-shading stimulus induces higher values of curvature and the texture-shading stimulus induces lower magnitudes of perceived depth and higher magnitudes of perceived curvature. Perceived depth for this stimulus is less than what would be needed to simulate the perceived curvature and orientation.

A three way ANOVA (task, cue-pair, simulated stretch) indicates that the only significant factors are the main effect of stretch ratio ($F(1, 5) = 242.367, p < 0.001$) and the 3-way interaction ($F(4, 20) = 2.921, p < 0.05$). The same analysis on the standard deviation obtained from the estimated psychometric function, indicates that the only significant factor is the simulated stretch ($F(1, 5) = 19.449, p < 0.01$). The ANOVA on the Weber fractions, instead showed a significant interaction between the three factors ($F(4, 20) = 3.994, p < 0.05$).

7.3 Discussion

In this experiment I measured the relative contribution cues have for the perception of different geometric properties of shape. Using a probe composed of all cues used in the stimuli, I could measure the relative contribution of each cue relatively to this maximal standard. This value constitutes a standard to which all information in the experiment can be compared. Underestimations in the perception of shape properties from cues are indicated by a lower value of magnitude of the probe.

Cues provide information only for some properties: depth for motion, slant and curvature for texture, and curvature for slant. The difference in the information provided can be quantized in figure 7.1. Judgments about shape depend on both the cues and the judged property.

When cues are combined, the information about shape is processed independently for each property. The judgment depends on the task also when more than one cue is present. It appears that the properties judged on the paired-cue stimuli might be related to the information available from the cue in isolation. For example, in the texture-shading stimuli in figure 7.2 depth is smaller than slant that in turn is smaller than curvature. Both texture and shading provide information about curvature and only texture can provide information for a reliable perception of slant. Neither cue allow to estimate depth (see 7.1). It appears that the magnitude of the perceived property of the two cues in isolation is related to the magnitude when combined. I will now use the IC model to relate the judgments in single and paired stimuli. In the final part of this discussion I will define the formulas to predict the participant's results in paired-cue stimuli from the data obtained in the single-cue stimuli for each of the three properties. This prediction can be made without free parameters or weights.

Other possible explanations of this data may be made using different theoretical frameworks. For example, it is possible to use a Bayesian approach to describe the noise in the measurements for each of the cues (for example Kersten, Mamassian, & Yuille, 2004). The probability distribution of the cue should change when a different task is given to the subject because it specifies a different "cost function". For example, this approach has been modeled by Schrater and Kersten (2000) who analyze the influence of the same information on different representations. Bayesian inference requires choosing a common depth representation to combine the cues, but there are different options regarding which type of depth comparisons are made explicit. As the authors did for size and shadow position, an explanation of these results can not be achieved by analyzing only the informational contribution of each cue. The authors admit that for the Bayesian framework to work, the whole posterior probability distribution must be computed (which is not always straightforward). The approach that I propose, as it will be seen below, does not require any more knowledge nor parameters in the prediction of the combined results.

This experiment requires the subject to compare two stimuli, a probe created using 3 cues (motion, texture and shading) with a stretch defined by S_3 and a test created either by using 2 cues (a and b) or 1 cue (x) with a stretch defined by S_a, S_b or S_x respectively. A staircase procedure was used to modify the probe stimulus until the perceived shape property (depth, orientation and curvature) was matched to the test stimulus. According to the IC Model, when the test and probe stimuli are perceived as having the same property P, the magnitude of the perceived property

π_3^P of the test is equal to the one of the single cue test π_x^P or the cue pair test π_2^P. I will use this convention: x is one of the three cues used in isolation (motion, texture and shading) and the numbers 2 and 3 indicate how many cues are present in the stimulus. So $S_{3.2}$ indicates the stretch factor of the probe when matched with a cue pair and $S_{3.a}$ is the value when matched with cue a. At the PSE, $\pi_{3.2}^P = \pi_2^P$ for one comparison and $\pi_{3.1}^P = \pi_x^P$ for the other, where the indexes 3.1 and 3.2 indicate the different comparisons.

The IC model discussed in Section 3.2 hypothesizes that perception of a property π_x^P is related to ρ (that is the score obtained from the PCA of the standardized image signals described in Section 3.2) by a monotonically increasing function $\pi_x^P = f(\rho_x^P)$. At the PSE, by combining this and the above formulas, we have the equalities $f(\rho_{3.1}^P) = g(\rho_x^P)$ and $f(\rho_{3.2}^P) = h(\rho_2^P)$. If we make the assumption that the function relating the PCA score ρ_x^P to the perceived quantity π_x^P does not change for a particular property in different conditions of stimulation, we can describe the goal of the experiment as finding the stimuli that satisfies $\rho_{3.1}^P = \rho_x^P$ and $\rho_{3.2}^P = \rho_2^P$ for the three properties P, the two simulated stretches S and the six cue conditions.

Equalization

Since the rotation and direction of illumination calculated in the equalization session are used in every stimulus, the test stimulus defined by single cues should appear to have the same amount of curvature. This minimizes fluctuations in the perceived shape that are not due to the factor of interest. In fact, the equalization should insure that for the same simulated stretch $S_m = S_t = S_s$ there is the same perceived π_x^C and consequently the same ρ_x^C for curvature for the three cues.

In the experimental session, if no other factors influence the perception of shape, the pattern of responses should also be equal for the other two properties. If there are no differences in the perceived shape with different tasks in single-cue conditions we should find that the matching shapes are equal ($S_x^D = S_x^O = S_x^C$). Modifications from this scheme would indicate that there is some interaction between the cue contained in the stimulus and the task. I expect that the conditions were the property can be estimated in a smaller number of steps by a cue, and indirectly by another cue, should produce a PSE where $S_d > S_i'$ as quantized below.

Standardization

In the IC model, the quantity ρ_x^P for a stimulus defined by a single cue x is equal to the standardized signal $\bar{S_x}$,

$$\rho_x^P = \bar{S_x}. \tag{7.1}$$

The standardized signal $\bar{S_x}$ is equivalent to:

$$\bar{S_x} = \bar{k}_x^P P_x + \bar{\varepsilon}_x \tag{7.2}$$

where $\bar{\varepsilon}_x$ is the standardized error term in the measurement $\bar{P}_x$ with mean 0 and variability 1 and $\bar{k}_x^P$ is the proportionality constant for the cue x and the distal property P. The standardization is obtained by the formula specifying the non-standardized signal:

$$S_x = k_x^P P_x + \varepsilon_x \tag{7.3}$$

By dividing the non-standardized signal P_x by the standard deviation σ_x of the noise term ε_x, the formula becomes:

$$\frac{S_x}{\sigma_x} = \frac{k_x^P}{\sigma_x} P_x + \frac{\varepsilon_x}{\sigma_x} \tag{7.4}$$

$$\bar{S_x} = \bar{k}_x^P P_x + \bar{\varepsilon}_x = \rho_x^P. \tag{7.5}$$

Because $\pi_x^P = f(\rho_x^P)$, the equalization session that precedes the experiment was used to equalize the perceived curvature π_x^C of the single cue condition ($\pi_m^C = \pi_t^C = \pi_s^C$) and therefore to equalize also ρ_x^P. The simulated property P in all the conditions of the equalization session was the curvature $C_e = 1.5$, so we can substitute in $\rho_m^C = \rho_t^C = \rho_s^C$ and from these values and we obtain:

$$\bar{k}_m^C 1.5 + \bar{\varepsilon}_m = \bar{k}_t^C 1.5 + \bar{\varepsilon}_t = \bar{k}_s^C 1.5 + \bar{\varepsilon}_s. \tag{7.6}$$

$\bar{\varepsilon}_x$ has equal variance and mean 0 because it was standardized of the standardization, so it becomes evident that the equalization produced $\bar{k}_m^C = \bar{k}_t^C = \bar{k}_s^C$ that we can summarize as $\bar{k}_e^C$.

Underestimation

The apparent underestimation of depth with single cue stimuli can be explained by the IC Model and the formulas provided above. In fact, if the single cue test stimulus provides a value of ro $\rho = \sigma_x = k_x S$ that has to be equal to the one provided by the probe stimulus $\rho = \sigma_{probe} = S'\sqrt{k_m^2 + k_t^2 + k_s^2}$. Assuming k is equal for the cues composing the probe (because of the equalization), $\rho = S' k_{cue}\sqrt{3} = k_{cue} S = k_{cue}\{1.0, 1.6\}$. Therefore it is possible to calculate the expected value of the stretch of the probe that matches the single cue stimuli to be $S' = \frac{k_{cue}\{1.0,1.6\}}{k_{cue}\sqrt{3}} = \{0.57, 0.96\}$. The values obtained are not statistically different from the ones predicted by the IC Model (2 tails, $t(5) = 1.5292, p = 0.19; t(5) = -0.7023, p = 0.51$).

Multiple cues

When stimuli contain n cues the IC Model indicates that the perceived property ρ_n^P can be determined by analyzing the conditions where the cues composing the stimuli are presented in isolation. For the IC Model, the ρ_n^P value can be simply estimated from the values of ρ_x^P by a simple Pythagorean equation $\rho_n^P = \sqrt{\sum (\rho_i^P)^2}$. With two cues present in the display (test stimulus) the ρ_n^P for two combined cues is:

$$\rho_2^P = \sqrt{{\rho_a^P}^2 + {\rho_b^P}^2}, \tag{7.7}$$

whereas for a probe stimulus that is composed of the three cues ρ is:

$$\rho_3^P = \sqrt{\rho_m^{P\,2} + \rho_t^{P2} + \rho_s^{P2}}. \tag{7.8}$$

For every cue in isolation, the IC model defines ρ_x^P to be equal to the scaled retinal signal which in turn is equal to $\rho_x^P = \bar{k}_x^P P_x$. Substituting this quantity for each cue in the formulas above, it is possible to calculate: the value of ρ_n^P for a property P and n cues from k_x^P, the values of the singles standardized proportionality constant of the cue x, and P_n, the value of the simulated property of the stimulus. For displays that contain two cues as test cue pair test stimulus ρ_2^P is:

$$\rho_2^P = P_2\sqrt{\bar{k}_a^{P2} + \bar{k}_b^{P2}}. \tag{7.9}$$

Similarly, for the probe stimulus ρ is:

$$\rho_3^P = \sqrt{(\bar{k}_m^P P_3)^2 + (\bar{k}_t^P P_3)^2 + (\bar{k}_s^P P_3)^2} = P_3\sqrt{\bar{k}_m^{P2} + \bar{k}_t^{P2} + \bar{k}_s^{P2}}. \tag{7.10}$$

For curvature, the equalization session has the effect of equating $\bar{k}_x^P$ for different cues, so that

$$\rho_x^C = \bar{k}_e^C C_x. \tag{7.11}$$

So, for curvature we should find that

$$\rho_3^C = \sqrt{\rho_a^{C2} + \rho_a^{C2} + \rho_a^{C2}} = \sqrt{\bar{k}_e^C {C_e}^2 + \bar{k}_e^C {C_e}^2 + \bar{k}_e^C {C_e}^2} = \bar{k}_e^C C_e \sqrt{3}. \qquad (7.12)$$

Task

These formulas predict the behavior of the visual system with two cues from the data obtained in the single cue. In the experimental sessions, we ask participants to compare the probe stimulus to the test stimulus to find the PSE for a property P. In this case in the two types of comparison at the PSE the equalities $\pi_{3.2}^P = \pi_2^P$ and $\pi_{3.1}^P = \pi_x^P$ are satisfied. According to the formulas above for the comparison with the single cue stimulus, we can write

$$P_{3.1}\sqrt{\bar{k}_m^{P2} + \bar{k}_t^{P2} + \bar{k}_s^{P2}} = \bar{k}_x^P P_x \qquad (7.13)$$

where $P_{3.1}$ is value of the simulated property P for the probe stimulus when compared with the single cue test. From this equality, we derive $\bar{k}_x^P$:

$$\bar{k}_x^P = \frac{P_{3.1}}{P_x}\sqrt{\bar{k}_m^{P2} + \bar{k}_t^{P2} + \bar{k}_s^{P2}} \qquad (7.14)$$

that can be used to predict what value of P_3 we should obtain when the probe is compared with a stimulus defined by two cues. In fact, if for simplicity we define

$\sum = \sqrt{{\bar{k}_m^P}^2 + {\bar{k}_t^P}^2 + {\bar{k}_s^P}^2}$, $\rho_3^P = \rho_2^P$ can be written as:

$$P_{3.2} \sum = P_2 \sqrt{{\bar{k}_a^P}^2 + {\bar{k}_b^P}^2} \tag{7.15}$$

where it is possible to substitute $\bar{k}_x^P$ to obtain:

$$P_{3.2} = \frac{P_2 \sqrt{\left(\frac{P_{3.a}}{P_x} \sum\right)^2 + \left(\frac{P_{3.b}}{P_x} \sum\right)^2}}{\sum} \tag{7.16}$$

$$P_{3.2} = \frac{P_2}{P_x} \sqrt{{P_{3.a}}^2 + {S_{3.b}}^2} \tag{7.17}$$

Equation 7.17 allows me to make predictions about cue combination from single cue measurements without any free parameter or weight. For tasks involving 2 cues, the perceived amount π_2^P of the shape property P measured in the experiment as the PSE value $P_{3.2}$ can be calculated from the PSE values $P_{3.x}$ obtained for stimuli defined by the cue composing the pair. If the simulated properties in the single cue and paired-cue condition are equal, the formula reduces to the sum of the squares of the values.

This result can be explained by the fact that neither of the cues present in the display, shading and texture, provides direct information for depth. Depth can be computed using an indirect method that results in more error in the estimation and a consequent lower quantity of depth in the display.

Figure 7.3 indicates the predicted values of $S_{3.2}^P$ made using the corresponding

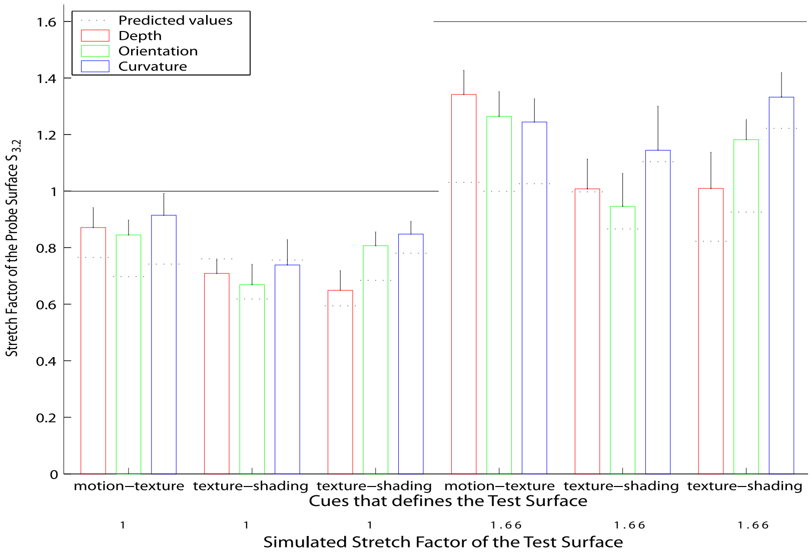


Figure 7.3: Predictions of the values of the stretch factor in the cue pair condition made using the values obtained experimentally in the single-cue conditions.

values of $S^P_{3.1}$ superimposed on the data from Experiment 2. The pattern across surface properties matches the data for all cues but one: motion-texture for curvature ($S = 1.66$). Most of the values fall within one standard error of the observer's means. The only quantitative difference is in the velocity-texture stimulus. This stimulus bears values of the perceived that are higher than expected. This effect is probably due to the interaction of motion and texture cues, a factor that was not considered in the predictions because it is created by the interaction of the cues. The deformation of the texture elements in time that happens when both cues are present creates an additional image signal that can be used to estimate information about shape. This information has an additive effect to the value of S because it can be considered as a third cue in the computation of ρ in the paired-cue stimulus. To verify this possibility, I modified the formula 7.17 for this condition so that

$$S_{3.2}^{P} = \sqrt{{P_{3.a}}^{2} + {S_{3.b}}^{2} + (\frac{P_{3.a}}{2} + \frac{P_{3.b}}{2})^{2}} \tag{7.18}$$

Notice that the same cue interaction is present on the probe stimulus. However, here we should expect no modification in the expected value, because the influence of such information was present in both single and paired cue conditions, therefore canceling out in the term Σ in formula 7.17. This modification provides a new estimate that is not different from the one registered experimentally as depicted in figure 7.3. These findings accounts for the two known cues, motion and texture, and a third cue 'D', with a value in isolation approximated as the mean of the S in the other conditions.

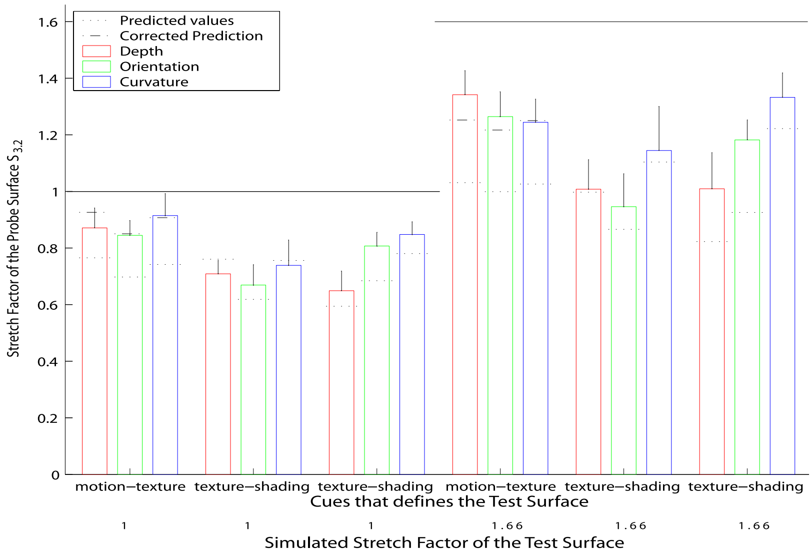


Figure 7.4: Predictions of the values in the cue pair condition modified according to the equation 7.18 and superimposed to the experimental results.

The pattern of responses and the predictions of a cue-integration scheme that keeps the computation separate for each property allows me to conclude that cues have a different influences in the specification of shape properties and, that the visual system keeps separate the estimates of the shape properties also when combining cues. Perceived properties of a shape containing multiple cues are mutually inconsistent, but can be predicted from the perceived properties from single-cue stimuli.

As anticipated in the discussion of the previous experiment, this result may be consistent with the presence of cues to flatness in the display. When a cue is present in isolation it might be claimed that all other cues specify a flat surface. The absence of a cue specifies a flat surface, so all absent cues should specify the same surface and can be accounted for by a decrease in the perceived property that is inversely proportional to the number of cues present in the display. If more cues are present in the display, the perceived properties increase because there are less cues specifying a flat surface. It must be notice that cues to flatness do not alone explain this data. If cues specify the properties of shape in an equal manner, the results should be consistent across properties. Perceived properties should be equal because derived from the same representation and the same image signals. The data I collected can be obtained only if cues are differently informative for the different tasks. This mechanism implies the existence of different representation of the shape. So whether the IC model or cues to flatness offer an adequate explanation of these results, the two are identical with respect to the explanation I propose: cue properties are perceived independently.

Chapter 8

Experiment 3: Surface reconstruction

In the experiments described in chapters 6 and 7 I tested the hypothesis that cues convey different information. The methodology used in the first experiment allowed me to compare stimuli containing different cues and estimate their differential contribution to the estimation of properties. The probe used in the second experiment allowed me to estimate directly the influence of each cue and cue pair for each property. However, the judgment performed by the subjects was made on different parts of the stimulus.

In this experiment I had participants estimate different properties at the same position of the stimuli to rule out the possibility that perceived shape is just deformed instead of being multiply represented. To show what the perceived shape looks like I combined different judgments across the surface and reconstruct it with the average result obtained for each subject.

To reduce the number of trials needed for such a reconstruction I substituted an method of adjustment for the method of constant stimuli used in the previous experiment. Participants modified the shape of the 3-cue probe stimulus by continuously changing the stretch factor. The task was to make the perceived property at a determined point match the magnitude of the property on a test surface on the corresponding point.

8.1 Method

Six undergraduate students naïve to the purpose of the study participated in the experiment, which began with an equalization session. The method of equalization was conceptually similar to that described in Section 5.2, but some modifications were made in order to use the method of adjustment. The participants were presented with the probe stimulus containing three cues and the test stimulus. The test stimulus had a stretch factor $S = 0.2$ and in different trials contained each of the three cues in isolation. The shading and motion cues had different values of rotation or illumination direction in different trials. Participants modified the shape of the left surface by increasing and decreasing the stretch factor using the keyboard in order to match the perceived curvature at the center. The shape of the surface was recomputed on-line as described below. A regression on the values of adjusted stretch was used to find the values of rotation and illumination that yield a perceived curvature equal to the average value adjusted with the texture stimulus.

The probe stimulus differed slightly from that described in the general methods in Section 5.1. The position of the spheres used to create the texture was randomized in the volume, instead of having the center on the surface. The minimum distance between the spheres was equal to 2 times the radius of the spheres $(2 * 0.5cm)$. When adjusted, the surface was calculated using different stretch factors, so that it intercepted different spheres and the intersected shape and size changed. The shading and motion cues were not modified from the original scheme, but they were recalculated at every modification of the surface. The two stimuli, probe and test, were presented simultaneously to the subject and the viewing time was not restricted. The difference in phase of the rotation was randomized at every trial.

Depending on the required judgement one or two yellow antialiased dots were superimposed on the stimulus along the vertical axis. For depth judgments one dot was positioned on the center of the surface and the other dot identified the position where the judgment should be made. For slant and curvature judgment only the dot identifying the location was present. Six positions were evaluated in different trials at a distance of 0.40, 1.25, 2.10, 2.90, 3.75, 4.60 cm from the center of the stimuli along the vertical. The motion of the dots was consistent with the simulated 3D rotation of the stimulus, so the dots appeared as being as small light sources like LEDs positioned on the surface.

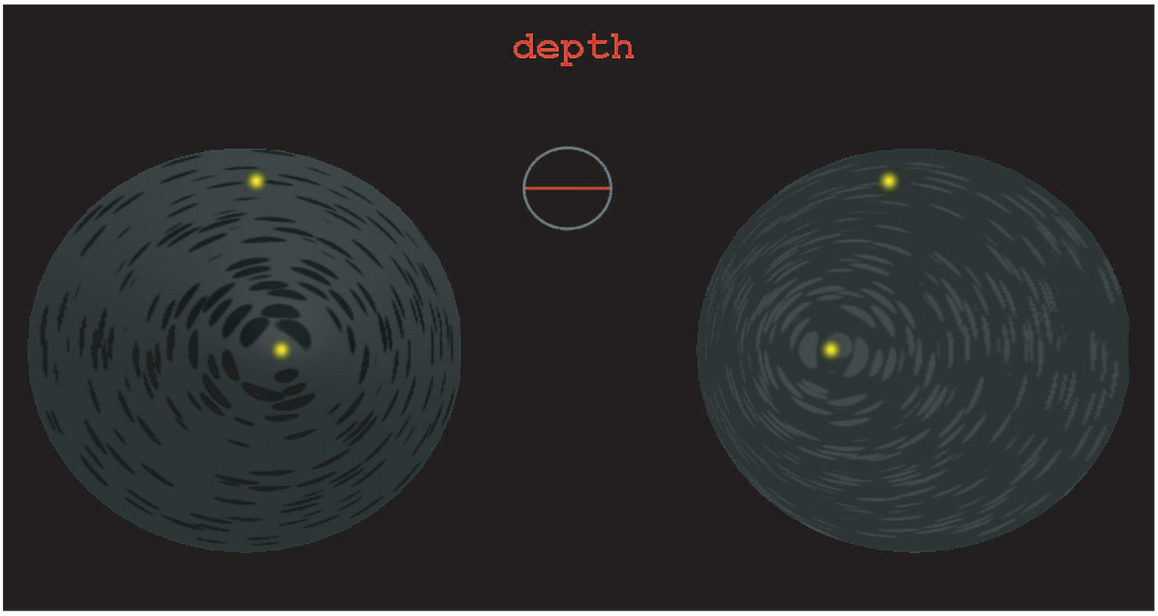


Figure 8.1: Appearance of the probe (left) and test (right) stimuli in this experiment.

There were 108 different conditions: six cue conditions, six position conditions and the three task conditions. The cue conditions were divided into two groups as in the previous experiment, three single-cue and three paired-cue. Participants that did not complete at least two adjustments per condition in two 25-minutes sessions were discarded from the analysis.

8.2 Results

The value of the stretch factor for each position, cue, and task averaged across observers is depicted in figure 8.2. The adjusted shape for the motion cue was systematically higher when judging depth. For texture, the higher value was slant. For shading the higher value was curvature, but instead of a constant stretching factor across the image, subjects reported higher values of curvature at the center. For the motion-texture stimulus, the values differentiate only in the middle of the radial extent, where depth is higher than slant and curvature is lower. For the motion-shading stimulus depth and slant are higher on the middle of the radial extent as well. For texture and shading, curvature has a peak at the center and slant decreases as with the radial distance.

To compare these results with the ones obtained in the second experiment described in Chapter 6, I computed the average of the adjusted values of the stretch factor across the position conditions. These values are also shown in figure 8.3. Motion has higher values for depth, texture for slant and shading for curvature. the motion-texture and the motion-shading stimulus have higher values of depth and slant and lower value of curvature. The texture-shading stimulus has a lower value of depth.

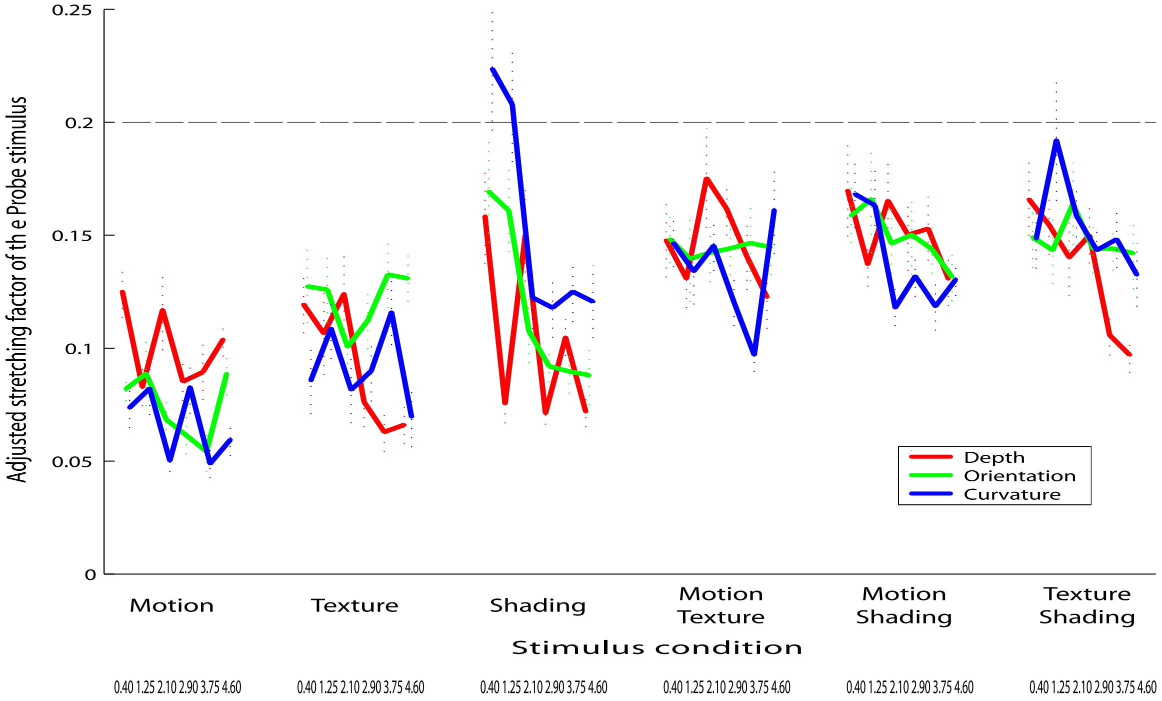


Figure 8.2: Adjusted stretch factor for the different conditions.

A 3 (task: depth, orientation, curvature) x 6 (stimuli) x 6 (height) repeated measures ANOVA on the adjusted stretch factor values revealed a main effect of stimuli ($F(5, 25) = 13.494, p < 0.001$) and height ($F(5, 25) = 2.710, p < 0.05$). The interactions between task and stimulus ($F(10, 50) = 5.442, p < 0.001$) and between stimulus and height ($F(50, 250) = 1.611, p < 0.05$) were also significant.

Using the data collected in the different conditions, I reconstructed the surfaces perceived by the subjects in each condition. From the value of the stretch factor, it is possible to use table 5.1 to find the value of the geometric property at the different

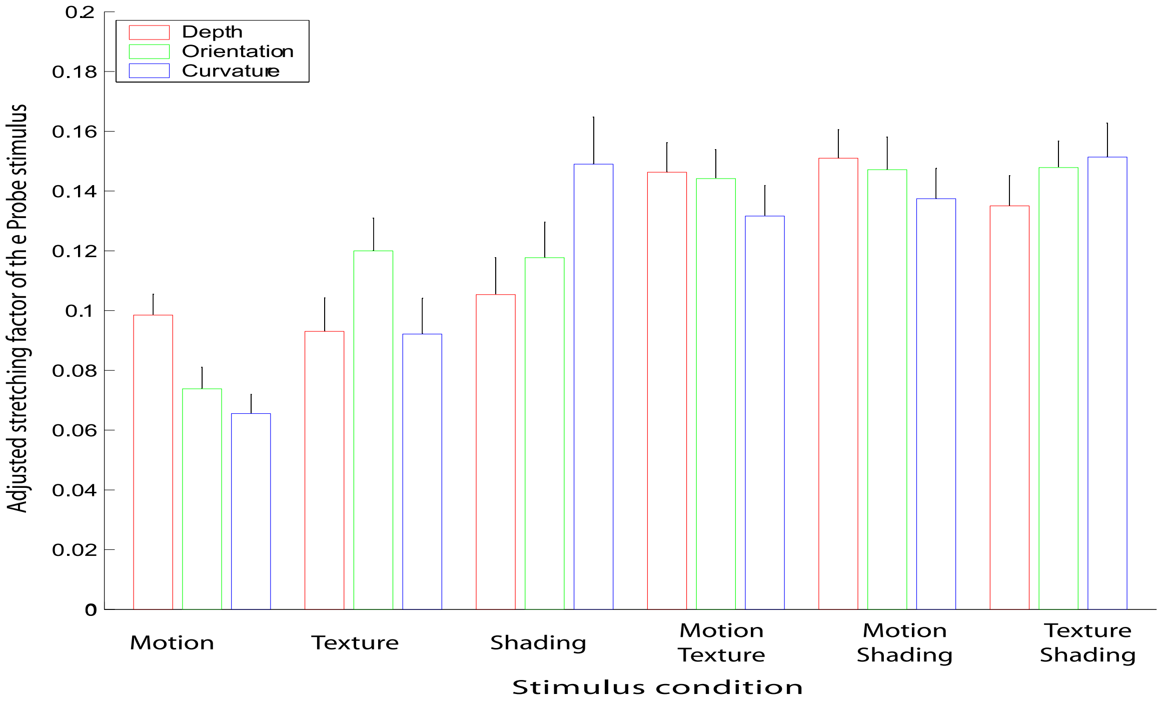


Figure 8.3: Adjusted stretch factor averaged across position on the surface.

points. Depth values are equivalent to the depth profile of the perceived surface. Slant values can be used to build a profile by integrating from the center of the stimulus. Curvature values can also be used for a numerical integration by assuming that at the center the perceived slant is 0. The surfaces obtained are shown in figure 8.4.

8.3 Discussion

The values of the stretch factor obtained in this experiment confirm that the data obtained in the second experiment, discussed in Chapter 6, cannot be explained by a unique representation of shape. In this experiment the judgments were performed on the same positions across the surface, but judgments of different properties produced

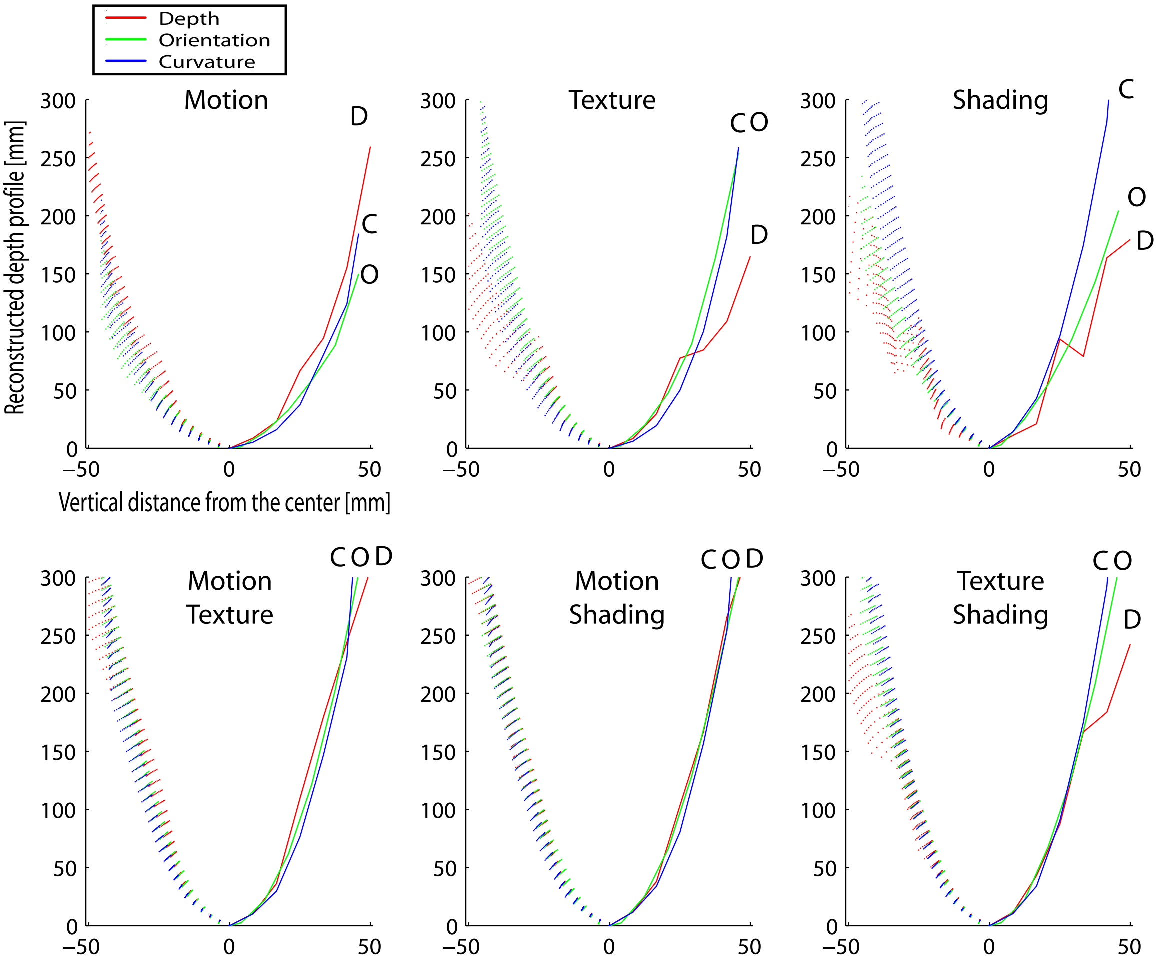


Figure 8.4: Surfaces reconstructed from the value of the adjusted stretch factor. The right side is obtained from the average values of the stretch factor across observers. The left side represents the different surfaces obtainable with with values within one standard error from the mean.
Top row: motion, texture, shading. Bottom row: motion-texture, motion-shading, texture-shading.

inconsistent results for the same shapes. Pooling results across positions on the surface (figure 8.3) creates a similar pattern of responses obtained previously for the single cue stimuli (Figure 7.1) and somewhat resembling for the combined cue (Figure 7.2). The correlation between the average data collected in this experiment and Experiment 2 is $r^2 = 0.6974$ for $S = 1.00$ and $r^2 = 0.7535$ for $S = 1.66$. Actually, the pattern for the single cue stimuli appears even more extreme than the one obtained above. The cues appear to have inconsistent properties when analyzed in isolation. For the combined condition, instead, the difference between tasks is less salient. While testing the experiment myself (the data is not reported) I realized that it is very easy to adjust the overall shape of the surface, rather than the local shape. This tendency is enhanced with multiple cues; having the test surface a greater stretch factor, the two shapes look more alike. The more information there is in the stimulus, the least the "beholder share". This pattern of responses cannot be produced by a single representation of a surface with Euclidean properties.

The data collected allows reconstruction of a shape that is most consistent with the perceived properties at each point. Figure 8.4 shows that the reconstructed shape for single cue stimuli is consistent with the data of the previous experiments, confirming that cues carry different types of information about shape. The shape reconstructed with evaluation of motion information shows that perceived depth yields a more elongated profile. Texture results in judgments of more curvature and slant while shading produces higher magnitude of curvature only. The above results refer to overestimation with respect to the other geometric properties of the same stimuli.

Using the adjusted values obtained in the single cue conditions I estimated the values of stretch factor in the combined cue conditions according to the Formula 7.17 and then reconstruct the shape of paired-cue stimuli. This procedure creates surfaces very similar to the ones obtained using the adjustments in the cue pair conditions (Figure 8.5). Although it is still not possible to exclude that these results have been produced by the effect of other processes, they are consistent with the proposed explanation of a combination of cues that happens in isolation for each of the properties. Moreover, the values of the adjusted properties for each spatial position with single cues are predictive of the values obtained for the same locations in multi cue stimuli. This indicates that the values of the perceived properties at one location are combined across cues to determine the perceived property at locations on the cue combined stimulus. This result shows that the information from each cue is combined only in isolation in the spatial domain. Information from neighborhood regions has only limited influence in the perception of local shape properties.

It is important to underline that, although I do not exclude global processing of image signals such as what is necessary to compute the parameters discussed in Section 3.5, the reconstruction performed here in both cases does not rely on any other information (global processing, priors, or assumptions) and it is obtained without free parameters. On the other hand, it is possible to conceive other types of explanation for these results. For example, it is possible that a Bayesian description of the perceptual process would predict the same type of results. To attempt this explanation, it is necessary to describe correctly three probability distributions: the likelihood function, that relates the image signals and the value of the perceived properties; the prior distribution, the description of the natural frequency of occurrence of the property in the environment; and the decision rule, the description of cost and gains of the response as evaluated by the subjects. The data collected can either be due to the use of a different decision rule when judging properties or to the different geometrical relation of the properties to the signal. The latter explanation requires much more analysis than the one provided in this work because each of the functions needs to be characterized experimentally before being able to make a prediction. I would like also to underline that the functions relative to different properties are related by the same geometrical relations analyzed in Chapter 2. The same geometrical relation relates both the likelihood and the prior of two properties. Therefore if no other factors are involved, cues may produce estimates that are inconsistent across properties. Nevertheless, different cues should produce the same pattern of responses because the signals are affected by the same geometric relation of the distributions. Therefore, the likelihood and prior distributions alone cannot predict the different pattern of responses created by each of the cues used in this experiment. On the other hand, the possibility that the decision rule could account for these results cannot be discarded from the data collected in this experiment. Although no feedback was used in any of the conditions, the subjects could have adopted a different decision rule for each of the properties. It is not possible to characterize the decision rule in any manner, given the experimental conditions used here. In principle, a decision rule that explain these results exist and cannot be ruled out.

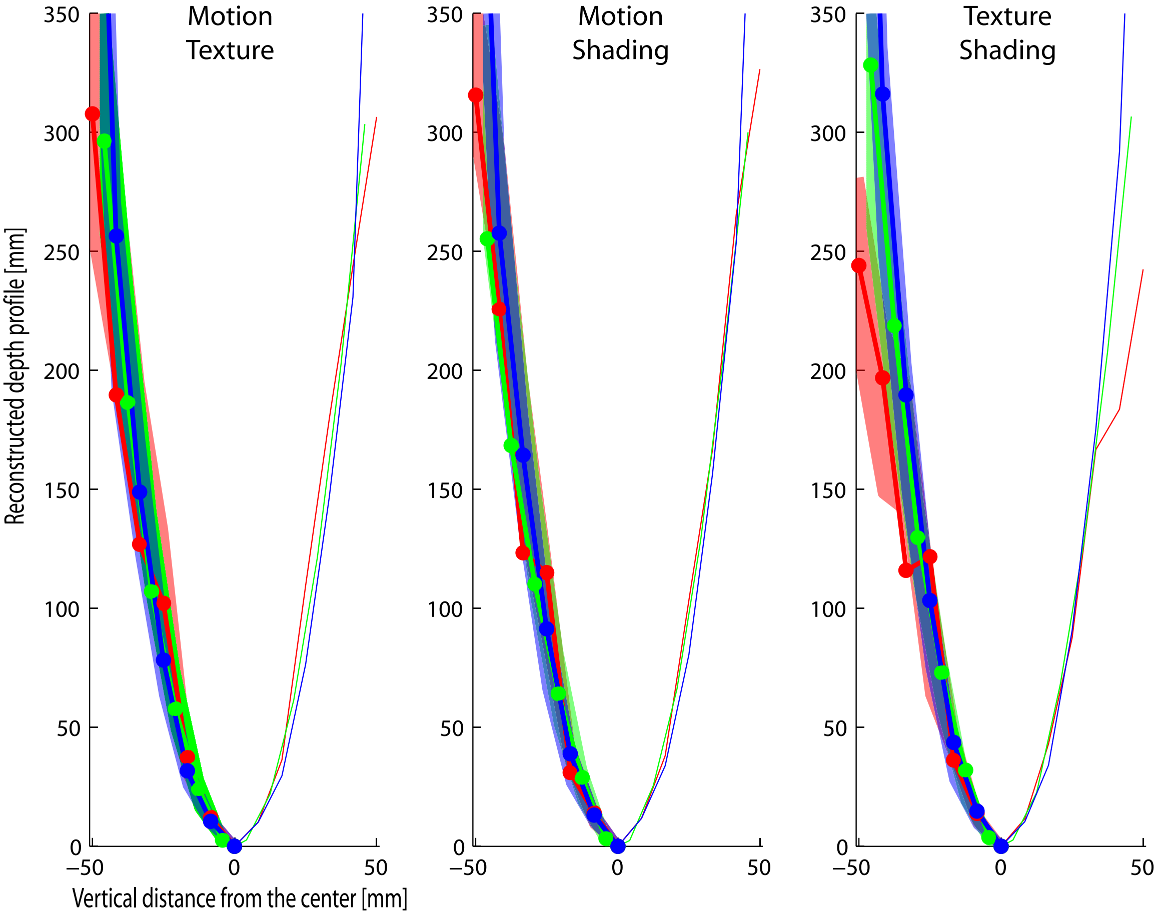


Figure 8.5: Right branch: reconstructed profile created using the responses in the paired-cue conditions. Left branch: predicted profile in the paired-cue condition created using responses in the single-cue conditions. Left: motion-texture. Center: motion-shading. Right: texture-shading.

Chapter 9

Conclusions

> “There is a road from the eye to heart that does not go through the intellect.” G. K. Chesterton

Shape perception is dependent on the analysis of information conveyed by cues created by the projection of a visible surfaces onto the retina. It is possible to identify the information used to estimate shape and to relate this information to different geometrical aspects of the environment. This information can be thought of as image signals measured by the visual system in order to estimate shape. The measurement and subsequent estimate are affected by noise, but the effect of this noise depends on the computations performed on the image signals. Estimates differ in precision because of the effect of noise in the measurements, complexity of computations, and spatial encapsulation of the processing. In certain cases, the measurements are simple to make, the estimates are linearly related to the signals, and the computation can be performed on local measurements. In other cases, the measurements are affected by large amounts of noise, the computations are performed on a large area of the image, many parameters are required, and the solution is not linearly related to the image signals.

This classification resembles a classical dichotomy in the study of perception that revolves around the question of direct vs. indirect perception[1]. According to the indirect approach to perception, shape perception is conceived as a problem of inverse geometry, where perception is the recovery of the Euclidean structure of objects. A

[1] Similarly to Haeckel’s “biogenic law”, there is a resemblance between the ontology (the study of the basic categories of being) of shape perception and its philology (the study of ancient texts).

controversial argument for direct perception was proposed by (Gibson, 1979), when he described visual perception in terms of "direct pickup" of information contained in the optic array. According to this view, inference, computational mechanisms and representation are superfluous concepts. Information about the estimates are simply detected in the optic array. The detection mechanism depends on the nature of the pattern but the interpretation process depends only the properties being computed.

The data reported in this work supports a characterization of cues in terms of the noise in the estimate. In contrast to the assumptions from the classical approach to shape perception, the data indicate that this estimate depends also on the geometric property considered. Three-dimensional shape can be described by different geometric properties like relative depth, local orientation and curvatures (e.g. Koenderink et al., 1996). Most of the psychophysical research on 3D shape perception asserts that cues are equally informative for every 3D property(Bülthoff & Mallot, 1990). However, I provide empirical evidence that cues are differently informative about geometric properties of shape.

Perceived properties are mutually inconsistent and their magnitudes depends on the mathematical relation between image signal and geometry of the property. Using a simulated observer, I showed that this relation can be quantified to predict the perceived magnitude of the geometric properties. The results obtained in the first experiment (Chapter 6) with perceptual comparisons of single cue stimuli show a similar pattern.

The inconsistencies between the perceived magnitude of different properties implies that they are perceived independently. Perceived properties of shape are computed from the image signals, and there is no single representation that resolves these inconsistencies. The shapes reconstructed from local judgments of geometric properties in different positions are qualitatively different when the properties convey different information about shape. My results cannot be accounted for with any single Affine or Euclidean representation of shape.

Despite the possible advantages resulting from combining information characterized by different noise patterns, the data show that when cues are combined by the visual system the combination happens independently for each property. I have demonstrated that the IC model, a non linear model of cue integration, can be modified to account for my results. The judgments of properties in multi-cue stimuli can be predicted from the same judgments on in single-cue stimuli. The modification to the IC model requires each image signal to be represented only in the appropriate "signal space" (see Domini et al., 2006) according the information that it carries. The combination of the signals then happens only within each signal space. The MWF instead cannot be modified in such a way because one of the required mechanism is the promotion of cues. The computations involved change the information carried by the cues and all cues become informative about depth. This requirement of the MWF precludes any possible modification that could account for the data collected in this dissertation.

In the first pages of this work I described the qualitative shape of an apple on my desk in my own words. This description gave you the reader a mental image of the shape of the apple, but this description is not sufficient to convey my perceptual experience. Because I am not describing my experience using the language of the visual system I cannot describe my percept. It is still not clear what the basic properties estimated by the visual system are. We recognize that the shape of the apple is perceived through its geometric properties. My contribution has been to prove that the computations of these properties from the image signals are kept separate by the visual system. This is an important step in furthering our understanding of perception but it still leaves many questions unanswered. Although this might seem like a complication, it demonstrates that the representation of shape by the visual system is of the simplest form possible.

References

Aloimonos, J., & Swain, M. (1988). Shape from patterns: Regularization. *International Journal of Computer Vision*, *2*, 171 - 187.

Arnheim. (1954). *Art and visual perception. a psychology of the creative eye.* Berkley: University of California Press.

Attneave, F. (1972). Representation of physical space. In A. Melton & E. Martin (Eds.), *Coding processes in human memory.* Washington, D. C.: V. W. Winstone.

Backus, B. T., & Banks, M. S. (1999). Estimator reliability and distance scaling in stereoscopic slant perception. *Perception*, *28*(2), 217-242.

Barrow, H. G., & Tenenbaum, J. M. (1978). Recovering intrinsic scene characteristics from images. In *Computer vision systems* (p. 326). San Diego, CA: Academic Press.

Berends, E. M., Liu, B., & Schor, C. M. (2005). Stereo-slant adaptation is high level and does not involve disparity coding. *Journal of Vision*, *5*, 71-80.

Berkeley, G. (1709). *An essay towards a new theory of vision* (4th ed.). Dublin:

Jeremy Pepyat.

Bingham, G. P., & Pagano, C. C. (1998). The necessity of a perception-action approach to definite distance perception: monocular distance perception to guide reaching. *J Exp Psychol Hum Percept Perform*, *24*(1), 145-68. (0096-1523 Journal Article)

Birnbaum, M. H. (1983). Scale convergence as a principle for the study of perception. In H. G. Geissler & V. Sarris (Eds.), *Modern issues in perceptual psychology.* North Holland.

Blake, A., & Bülthoff, H. H. (1990). Does the brain know the physic of specular reflection. *Nature*, *343*, 165-168.

Blake, A., Bülthoff, H. H., & Sheinberg, D. (1993). Shape from texture: ideal observers and human psychophysics. *Vision Res*, *33*(12), 1723-37.

Bradshaw, M. F., Glennerster, A., & Rogers, B. J. (1996). The effect of display size on disparity scaling from differential perspective and vergence cues. *Vision Res*, *36*(9), 1255-64.

Bradshaw, M. F., Parton, A. D., & Glennerster, A. (2000). The task-dependent use of binocular disparity and motion parallax information. *Vision Res*, *40*(27), 3725-34.

Braunstein, M. L. (1994). Decoding principles, heuristics and inference in visual perception. Erlbaum.

Braunstein, M. L., Andersen, G. J., Rouse, M. W., & Tittle, J. S. (1986). Recovering viewer-centered depth from disparity, occlusion, and velocity gradients. *Percept Psychophys*, *40*(4), 216-24.

Braunstein, M. L., Anderson, G. J., & Riefer, D. M. (1982). The use of occlusion to resolve ambiguity in parallel projections. *Percept Psychophys*, *31*(3), 261-7.

Brenner, E., & Damme, W. van. (1999). Perceived distance, shape and size. *Vision Research*, *39*(5), 975-986.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Bruss, A. (1982). The eikonal equation: Some results applicable to computer vision. *J. Math. Physics*, *23*, 890-896.

Buckley, D., & Frisby, J. P. (1993). Interaction of stereo, texture and outline cues in the shape perception of three-dimensional ridges. *Vision Research*, *33*(7), 919-33.

Bülthoff, H. H. (1991). Shape from x: Psychophysics and computation. In M. S. Landy & J. Movshon (Eds.), *Computational models of visual processing* (p. 295-305). Cambridge, MA: MIT Press.

Bülthoff, H. H., & Mallot, H. A. (1988). Integration of depth modules: Stereo and shading. *Journal of the Optical Society of America A: Optics and Image Science*, *5*(10), 1749-1758. (Journal Article Oct)

Bülthoff, H. H., & Mallot, H. A. (1990). Integration of stereo, shading and texture. In A. Blake & T. Troscianko (Eds.), *Ai and the eye* (p. 295-305). Chichester, England: Whiley.

Carman, G. J., & Welch, L. (1992). Three-dimensional illusory contours and surfaces. *Nature (London)*, *360*, 585-587.

Caudek, C., & Proffitt, D. R. (1993). Depth perception in motion parallax and stereokinesis. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 32-47.

Christou, C., Koenderink, J. J., & Doorn, A. J. van. (1996). Surface gradients, contours and the perception of surface attitude in images of complex scenes. *Perception*, *25*, 701–713.

Clark, J., & Yuille, A. (1990). *Data fusion for sensory information processing systems.* Norwell, MA: Kluwer academic publishers.

Craik, K. (1943). *The nature of exploration.* Cambridge, England: Cambridge University Press.

Curran, W., & Johnston, A. (1994). The effect of light source position on perceived curvature. *Investigative Ophthalmology and Visual Science*, *35*, 1741-.

Curran, W., & Johnston, A. (1996). The effect of illuminant position on perceived curvature. *Vision Res.*, *36*, 1399-1410.

Cutting, J. E. (1986). *Perception with an eye for motion.* Cambridge, MA, USA: MIT Press.

Cutting, J. E., & Bruno, N. (1988). Additivity, subadditivity, and the use of visual information: a reply to massaro (1988). *Journal of Experimental Psychology:*

General, *117*(4), 422-4.

Cutting, J. E., & Millard, R. T. (1984). Three gradients and the perception of flat and curved surfaces. *Journal of Experimental Psychology: General*, *113*(2), 198-216.

Cutting, J. E., & Vishton, P. M. (1994). Perceiving layout: The integration, relative dominance, and contextual use of different information about depth. In *Conference of the psychonomic society.* St. Louis, Missouri.

Cutting, J. E., & Vishton, P. M. (1995). Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In W. Epstein & S. Rogers (Eds.), *Perception of space and motion. handbook of perception and cognition (2nd ed.)* (Vol. xix, p. 69 - 117). San Diego, CA, US: Academic Press, Inc.

Daniilidis, & Spetsakis. (1996). Visual navigation. Hillsdale, NJ: Lawrence Erlbaum Associates.

De Bruyn, B., & Orban, G. A. (1988). Human velocity and direc- tion discrimination measured with random-dot patterns. *Vision Research*, *28*, 1323-1335.

Dennett, D. C. (1991). *Consciousness explained.* Boston, MA: Little, Brown.

Dijkstra, T. M. H., Snoeren, P. R., & Gielen, C. C. A. M. (1994). Extraction of three-dimensional shape from optic flow: a geometric approach. *Journal of the Optical Society of America A*, *8*, pp. 2184-.

Di Luca, M., Domini, F., & Caudek, C. (2004). Spatial integration in structure from motion. *Vision Research*, *44*, 3001-3013.

Di Luca, M., Domini, F., & Caudek, C. (submitted). Effects of contextual information on stereo-motion processing. *Vision Research*.

Domini, F., & Braunstein, M. L. (1998). Recovery of 3d structure from motion is neither euclidean nor affine. *Journal of experimental psychology: human perception and performance*, *24*, 1273-1295.

Domini, F., & Caudek, C. (1999). Perceiving surface slant from deformation of optic flow. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 426-444.

Domini, F., Caudek, C., & Tassinari, H. (2006). Stereo and motion information are not independently processed by the visual system. *Vision research*, *46*, 1707-1723.

Erens, R. G., Kappers, A. M., & Koenderink, J. J. (1993). Perception of local shape from shading. *Percept Psychophys*, *54*(2), 145-56.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429-33.

Fermuller, C., Cheong, L., & Aloimonos, Y. (1997). Visual space distortion. *Biol Cybern*, *77*(5), 323-37.

Foley, J. (1977). Effect of distance information and range on two indices of visually perceived distance. *Perception*, *6*, 449-60.

Foley, J. D., & Dam, A. V. (1983). *Fundamentals of interactive computer graphics.*

Reading, MA: Addison-Wesley.

Gärding, J. (1992). Shape from texture for smooth curved surface in perspective projection. *Journal of Mathematical Imaging and Vision*, *2*, 329-352.

Gibson, J. J. (1950). *The perception of the visual world.* Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston, MA: Houghton-Mifflin.

Glennerster, A., Rogers, B., & Bradshaw, M. (1996). Stereoscopic depth constancy depends on the subject's task. *Vision Research*, *36*, 3441-3456.

Gogel, W. C., & Tietz, J. D. (1977). Eye fixation and attention as modifiers of perceived distance. *Percept Mot Skills*, *45*(2), 343-62.

Gombrich, E. H. (1974). *Art and illusion part iii the beholder's share.* London, UK: Phaidon Press).

Graziano, M. S., Yap, G. S., & Gross, C. G. (1994). Coding of visual space by premotor neurons. *Science*, *266*(5187), 1054-7.

Gregory, R. L. (1968). Perceptual illusions and brain models. *Proc R Soc Lond B Biol Sci*, *171*, 179-296.

Helmholtz, H. v. (1910). *Physiological optics* (1962 ed.). New York: Dover.

Hildreth, E. (1984). *The measurement of visual motion.* Cambridge, Mass: MIT Press.

Hillis, J., Watt, S., Landy, M., & Banks, M. (2004). Slant from texture and disparity cues: optional cue combination. *Vision Research*, *4*, 1-3.

Hillis, J. M., Banks, M. S., & Landy, M. S. (2002). How are texture and stereo used in slant discrimination? *Journal of Vision*, *2*(7), Abstract 325.

Hoffman. (1985). Inferring the relative three-dimensional positions of two moving points. *J. Opt. Soc. Am. A*, *2*, 350-.

Hoffman. (1986). The computation of structure from fixed-axis motion: rigid structures. *Biological Cybernetics*, *54*, 71 - 83.

Hogervorst, M. A., & Eagle, R. A. (1998). Biases in three-dimensional structure-from-motion arise from noise in the early visual system. *Proc R Soc Lond B Biol Sci*, *265*(1406), 1587-93.

Horn, B. (1986). *Robot vision.* Cambridge, Massachusetts: MIT Press.

Horn, B., & Brooks, M. (1989). *Shape from shading.* Cambridge, Massachusetts: MIT Press.

Horn, B. K. P. (1975). Obtaining shape from shading information. New York: McGraw Hill.

Horn, B. K. P. (1977). Understanding image intensities. *Artificial Intelligence*, *8*, 201-231.

Howard, I. P. (2002). *Seeing in depth.* Toronto: I Porteous.

Ikeuchi, K. (1984). Shape from regular patterns. *J. of Artificial Intelligence*, *22*, 4975.

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences*, *6*(8), 345-350.

Jacobs, R. A., & Fine, I. (1999). Experience-dependent integration of texture and motion cues to depth. *Vision Research*, *39*(24), 4062-4075. (Journal Article Dec)

James, W. (1890). *The principles of psychology.*

Jau, J., & Chin, R. (1990). Shape from texture using the wigner distribution. *CVGIP*, *52*, 248-263.

Johnston, A., & Passmore, P. J. (1994a). Independent encoding of surface orientation and surface curvature. *Vision Research*, *34*(22), 3005-12.

Johnston, A., & Passmore, P. J. (1994b). Shape from shading. ii: Geodesic bisection and alignment. *Perception*, *23*(2), 191-200.

Johnston, A., & Passmore, P. J. (1994c). Shape from shading. i: Surface curvature and orientation. *Perception*, *23*(2), 169-189.

Johnston, E. B., Cumming, B. G., & Landy, M. S. (1994). Integration of stereopsis and motion shape cues. *Vision Research*, *34*(17), 2259-2275. (Journal Article Sep)

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *55*, 271-304.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Curr Opin Neurobiol*, *13*(2), 150-8.

Knill, D. C. (n.d.). Discrimination of planar surface slant from texture: Human and ideal observers compared. journal = Vision Res, volume = 38, pages =

1683-1711, year = 1998.

Knill, D. C. (1998). Ideal observer perturbation analysis reveals human strategies for inferring surface orientation from texture. *Vision Res*, *38*(17), 2635-56.

Knill, D. C. (2005). Reaching for visual cues to depth: The brain combines depth cues differently for motor control and perception. *Journal of Vision*, *5*, 103-115.

Koenderink, J. J. (1986). Optic flow. *Vision Res*, *26*(1), 161-79.

Koenderink, J. J. (1990). *Solid shape.* Cambridge, MA: MIT Press.

Koenderink, J. J. (2001). Multiple visual worlds. *Perception*, *30*, 1-7.

Koenderink, J. J., Doorn, A. J. van, & Kappers, A. M. (1996). Pictorial surface attitude and local depth comparison. *Perception and Psychophysics*, *58*(2), 163-173.

Koenderink, J. J., Doorn, A. J. van, & Kappers, A. M. L. (1992). Surface perception in pictures. *Perception and Psychophysics*, *52*, 487-496.

Landy, M. S., & Brenner, E. (2001). Motion-disparity interaction and the scaling of stereoscopic disparity. In L. R. Harris & M. R. M. Jenkin (Eds.), *Vision and attention* (p. 129-151). New York: Springer Ve.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. J. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, *35*(3), 389-412. (Journal Article Feb)

Langer, M. S., & Bülthoff, H. H. (1994). Measuring visual shape using computer graphics psychophysics. In *Proceedings of the eurographics workshop on rendering techniques.* London, UK: Springer-Verlag.

Lappin, J. S., & Craft, W. D. (2000). Foundations of spatial vision: From retinal images to perceived shapes. *Psychological Review*, *107*(1), 6-38.

Leonardo, d. V. (1888). *The notebooks of leonardo da vinci from the english translation by jean paul richter.*

Liter, J. C., & Braunstein, M. L. (1998). The relationship of vertical and horizontal velocity gradients in the perception of shape, rotation, and rigidity. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 1257-72.

Liter, J. C., Braunstein, M. L., & Hoffman, D. D. (1993). Inferring structure from motion in two-view and multiview displays. *Perception*, *22*, 1441-1465.

Loomis, J. M., Da Silva, J. A., Fujita, N., & Fukusima, S. S. (1992). Visual space perception and visually directed action. *J Exp Psychol Hum Percept Perform*, *18*(4), 906-21.

Malik, J., & Rosenholtz, R. (1997). Computing local surface orientation and shape from texture for curved surfaces. *International Journal of Computer Vision*, *23*, 149-168.

Maloney, L. T., & Landy, M. S. (1989). A statistical framework for robust fusion of depth information. In W. A. Pearlman (Ed.), *Proceedings of spie: Visual communications and image processing iv.* (Vol. 1199, p. 1154-1163).

Mamassian, P., Kersten, D., & Knill, D. C. (1996). Categorical local-shape perception.

Perception, *25*, 95-107.

Marr, D. (1980). Visual information processing: the structure and creation of visual representations. *Philos Trans R Soc Lond B Biol Sci*, *290*(1038), 199-218.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information.* San Francisco, CA, USA: Freeman and Co. ("Seeing" is to know what is where by looking)

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three dimensional structure. *Proceedings of the Royal Society of London B*, *200*, 269-294.

Massaro, D. W., & Cohen, M. M. (1993). The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General*, *122*(1), 115-124.

Mather, G. (1997). The use of image blur as a depth cue. *Perception*, *26*(9), 1147-58.

Mausfeld, R. (2003). Conjoint representations and the mental capacity for multiple simultaneous perspectives. Cambridge, Mass.: MIT Press.

Milner, A. D., & Goodale, M. A. (1995). *The visual brain in action.* London: Oxford University Press.

Mingolla, E., & Todd, J. T. (1986). Perception of solid shape from shading. *Bio. Cyber.*, *53*, 137-151.

Nakayama, K., & Shimojo, S. (1992). Esperiencing and perceiving visual surfaces. *Science*, *257*.

Norman. (1993). The perceptual analysis ofstructure from motion for rotating objects undergoing affine stretching transformations. *Perception and Psychophysics*, *53*, 279-91.

Norman, J. F., Todd, J. T., Perotti, V. J., & Tittle, J. S. (1996). The visual perception of three-dimensional length. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(1), 173-186.

Pagano, C. C., & Bingham, G. P. (1998). Comparing measures of monocular distance perception: verbal and reaching errors are not correlated. *J Exp Psychol Hum Percept Perform*, *24*(4), 1037-51. (0096-1523 Journal Article)

Pentland, A. P. (1982). inding the illuminant direction. *Journal of the Optical Society of America*, *72*, 448-455.

Pentland, A. P. (1984). Local shading analysis. *IEEE Transactions on Analysis and Machine Intelligence*, *6*, 170-187.

Perotti, V. J., Todd, J. T., Lappin, J. S., & Phillips, F. (1998). The perception of surface curvature from optical motion. *Perception and Psychophysics*, *60*, 377388.

Philbeck, J. W., & Loomis, J. M. (1997). Comparison of two indicators of perceived egocentric distance under full-cue and reduced-cue conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(1), 72-85.

Phillips, F., Todd, J. T., Koenderink, J. J., & Kappers, A. M. (2003). Perceptual representation of visible surfaces. *Percept Psychophys*, *65*(5), 747-62.

Prazdny, K. (1986). Three-dimensional structure from long-range apparent motion. *Perception*, *15*, 619-625.

Proffitt, D. R., Bertenthal, B. I., & Roberts, R. J. (1984). The role of occlusion in reducing multistability in moving point-light displays. *Perception and Psychophysics*, *36*(4), 315-323.

Purdy, W. C. (1960). The hypothesis of psychophysical correspondence in space perception. *General Electric Technical Information Series*, *R60ELC56.*

Ramachandran, V. S. (1988). Perception of shape from shading. *Nature*, *331(6152)*, 163-166.

Reichel, F. D., Todd, J. T., & Yilmaz, E. (1995). Visual discrimination of local surface depth and orientation. *Perception and Psychophysics*, *57*, 1233-1240.

Richards, W. (1985). Structure from stereo and motion. *Journal of the Optical Society of America A: Optics and Image Science*, *2*(2), 343-9.

Rock, I. (1984). *Perception* (Vol. New York). Scientifica American Library.

Rogers, B., & Bradshaw, M. F. (1993). Vertical disparities, differential perspective and binocular stereopsis. *Nature*, *361*, 253-255.

Rogers, B. J., & Collett, T. S. (1989). The appearance of surfaces specified by motion parallax and binocular disparity. *Q J Exp Psychol A*, *41*(4), 697-717.

Rosenholtz, R., & Malik, J. (1997). Surface orientation from texture: isotropy or homogeneity (or both)? *Vision Research*, *37*, 2283-93.

Schrater, P. R., & Kersten, D. (2000). How optimal depth cue integration depends on the task. *International Journal of Computer Vision*, *40*, 73-91.

Schwartz, B. J., & Sperling, G. (1983). Luminance controls the perceived 3-d structure of dynamic 2-d displays. *Bulletin of the Psychonomic Society*, *21*(6), 456-458.

Shopenhaurer, A. (1847). *On the fourfold root of the principle of sufficient reason* (2nd ed.). LaSalle, Ill.: Open Court.

Stevens, K. A. (1979). *The resolution of conflict between disparity and overlap cues to depth.* Unpublished doctoral dissertation.

Stevens, K. A. (1981). Constructing the perception of surfaces from multiple cues. In G. W. Humphrey (Ed.), *Understanding vision.* Oxford, UK: Blackwell.

Stevens, K. A. (1984). On gradients and texture "gradients". *Journal of Experimental Psychology: General*, *113*, 217-220.

Stevens, K. A. (1995). Integration by association: Combining three-dimensional cues to extrinsic surface shape. *Perception*, *24*(2), 199-214.

Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theories.* New York: John Wiley.

Stratton, G. M. (1897). Vision without inversion of the retinal image. *Psychological Review*, *4*, 441-481.

Super, B., & Bovic, A. (1995). Shape from texture using local spectral moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*, 333-343.

Taylor, J. (1962). *The behavioral basis of perception.* New Haven, CT, US: Yale

University Press.

Titchener, E. B. (1906). *Experimental psychology.* New York: Macmillan.

Tittle, J. S., Norman, J. F., Perotti, V. J., & Phillips, F. (1998). The perception of scale-dependent and scale-independent surface structure from binocular disparity, texture, and shading. *Perception*, *27*(2), 147-66.

Tittle, J. S., Perotti, V. J., & Norman, J. F. (1997). Integration of binocular stereopsis and structure from motion in the discrimination of noisy surfaces. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(4), 1035-1049.

Tittle, J. S., Todd, J. T., Perotti, V. J., & Norman, J. F. (1995). Systematic distortion of perceived three-dimensional structure from motion and binocular stereopsis. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 663-678.

Todd, J. T. (2004). The visual perception of 3d shape. *Trends in cognitive Sciences*, *8*, 115-121.

Todd, J. T., Akerstrom, R., Reichel, F., & Hayes, W. (1988). Apparent rotation in three-dimensional space: Effects of temporal, spatial, and structural factors. *Perception and Psychophysics*, *43*, 179-188.

Todd, J. T., & Akerstrom, R. A. (1987). Perception of three-dimensional form from patterns of optical texture. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 242-255.

Todd, J. T., & Bressan, P. (1990). The perception of 3-dimensional affine structure from minimal apparent motion sequences. *Perception and Psychophysics*, *48*(5), 419-430.

Todd, J. T., & D., R. F. (1989). Ordinal structure in the visual perception and cognition of smoothly curved surfaces. *Psychological Review*, *96*, 643-657.

Todd, J. T., & Mingolla, E. (1983). Perception of surface curvature and direction of illumination from patterns of shading. *J. Exp. Psych.: Hum. Percept. and Perf.*, *9*, 583-595.

Todd, J. T., & Norman, J. (1991). The visual perception of smoothly curved surfaces from minimal apparent motion sequences. *Perception and Psychophysics*, *50*, 509-523.

Todd, J. T., & Norman, J. F. (2003). The visual perception of 3-d shape from multiple cues: Are observers capable of perceiving metric structure. *Perception and Psychophysics*, *65*(1), 31-47.

Todd, J. T., & Perotti, V. J. (1999). The visual perception of surface orientation from optical motion. *Perception and Psychophysics*, *61*, 1577-1589.

Todd, J. T., Tittle, J. S., & Norman, J. (1995). Distortions of three-dimensional space in the perceptual analysis of motion and stereo. *Perception*, 75-.

Ullman. (1979). The interpretation of structure from motion. *Proc R Soc Lond B Biol Sci.*, *203*, 405-26.

Wallach, H., & O'Connell, D. N. (1953). The kinetic depth effect. *Journal of*

Experimental Psychology, *73*, 117-129.

Watt, S. J., Akeley, K., Ernst, M. O., & Banks, M. S. (2005). Focus cues affect perceived depth. *Journal of Vision*, *5*, 834-862.

Westheimer, G., & McKee, S. P. (1979). What prior uniocular processing is necessary for stereopsis? *Invest. Ophtalmol. Visual Sci.*, *18*, 614-621.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: Ii. bootstrap-based confidence intervals and sampling. *Perception and Psychophysics*, *63*, 1314-1329.

Witkin, A. P. (1981). Recovering surface shape and orientation from texture. *Artificial Intelligence*, *17*, 17-45.

Yonas, A. (1979). Attached and cast shadows. In *Perception and pictural representation* (p. 100-109). Praeger.

Young, M. J., Landy, M. S., & Maloney, L. T. (1993). A perturbation analysis of depth perception from combinations of texture and motion cues. *Vision Research*, *33*(18), 2685-2696.

Yuille, A. L., & Bülthoff, H. H. (1994). Bayesian decision theory and psychophysics. *Advances in Neural Information Processing Systems.*