Introduction

This book is about structural information theory (SIT) and its application to visual form perception. Here, by way of general introduction, we highlight several unique characteristics of perception and we give a sketch of the scientific roots of SIT.

The uniqueness of perception

Almost all textbooks introduce perception by showing visual illusions. Indeed, visual illusions are salient phenomena. The core issues in perception are less salient, however. In fact, they are rather inaccessible and often confusing. This may be illustrated as follows. In every research domain – be it biology, physics, psychology, you name it – perception is the mediating instrument for making observations. The goal is to establish properties of objects. An observation may, for instance, establish that a leaf is green. Notice that this proposition merely deals with the relationship between a leaf and its colour. What is meant by a leaf and by green is supposed to be known. In perception research, however, perception is both mediating instrument and topic of study (Rock, 1983). As a topic of study, perception is the process that starts from an assembly of patches of light at various positions on the retina. This process assesses which patches are grouped together to constitute a leaf, for instance. In other words, the objects we perceive belong to the output of perception and not to the input. The goal of perception is not to establish properties of given objects but to establish objects from properties of the given retinal image.

Hence, in perception research, the two roles of perception (i.e., mediating instrument and topic of study) are virtually opposed to each other. Nevertheless, often, they are hardly distinguished. Usually, only one role is attributed to perception, namely, that of mediating instrument. This role is relevant at the conscious level involved in the everyday human communication of propositions. At this level, there is no sensation of the actual visual input which is an assembly of unstructured patches on the retina. There is also no experience of the perception process. The process

Figure I.1 A Maxwell demon, during his attempt to remove the milk from a milk-coffee mixture. This cumbersome job is similar to that of perception, in that both jobs turn chaos into order. Yet, there is also a crucial difference: to complete the job, the demon needs a lot of time, whereas perception needs only a few milliseconds (Leeuwenberg, 2003a).

is too rapid and too effortless for this. As a consequence, the perceived objects are taken as the actual components of the external scene we look at, even though they are in fact just mental constructs that result from the perception process in our brain.

Perception also exhibits a paradoxical feature, namely, with respect to temporal aspects. Within a few milliseconds, perception turns a mess of unstructured retinal patches into an ordered structure, whereas an analogous physical process from chaos to order may need millions of years. Such a process is illustrated in Figure I.1. It presents a little creature, a so-called Maxwell demon, in his attempt to remove the milk from a milk-coffee mixture. A milk-coffee mixture is obtained within the wink of an eye by pouring milk into black coffee; this is a process from order to chaos. To do the inverse, that is, to go from chaos to order, the Maxwell demon has to pick out the milk molecules one by one.

The above-mentioned confusions and paradoxes are not new. They already inspired Aristotle (350 BC/1957) to make the following, rather pessimistic, prophecy:

> In a shorter time more will be known about the most remote world, namely that of the stars, than about the most nearby topic, namely perception.

In our view, Aristotle's prophecy was right on target. Before 1850, perception was hardly acknowledged as a topic of study. Even nowadays,

it still has all the features of a young science. It is still approached by numerous independent and loose theories, which are plausible in some respects but untenable in others. Yet, in the mid-twentieth century, there were also developments which triggered the idea that perception should not be merely a topic of psychology and that one should not merely focus on discovering remarkable phenomena (Palmer, 1999). The insight was born that perception research should involve, apart from psychology, also physiology, mathematics, physics, and artificial intelligence. In other words, one realized that only divergent, interdisciplinary, research may lead to a convergent, coherent, understanding of perception as the unconscious process from the unstructured patches on our retina to the structured world we consciously perceive.

Scientific roots of SIT

The first modern approach to perception was given by the Gestalt psychologists (Koffka, 1935; Köhler, 1920; Wertheimer, 1923). SIT stems from their approach and attempts to integrate their findings. To give a gist of this, we first present the basic Gestaltist claims.

The Gestaltists claimed that perception is not a trivial process of copying stimuli but, rather, a complex bottom-up process of grouping retinal stimulus elements into a few stimulus segments. They proposed about 113 grouping cues, the so-called Gestalt laws of perceptual organization (Pomerantz and Kubovy, 1986). An example is the law of proximity. It states that nearby elements in a retinal stimulus tend to be grouped together to constitute one perceived object. Implicit to the application of a grouping cue is that a small change in the stimulus may lead to a dramatic reorganization of the stimulus. Furthermore, the strengths of the grouping cues are not fixed *a priori*. This implies that the perceptual interplay between simultaneously present cues depends on the stimulus at hand. That is, different cues may lead to different segmentations, and one cue may be decisive in one stimulus, but another cue may be decisive in another stimulus.

This stimulus-dependent interplay between cues is expressed by the Gestaltists' claim that 'the whole is different from the sum of its parts'. To specify this 'whole' (i.e., the percept, or Gestalt), they proposed a governing selection principle, namely the so-called law of Prägnanz. It states that the visual system tends to select the stimulus organization that is most 'simple', 'stable', 'balanced', and 'harmonious' (Koffka, 1935). This is reminiscent of the minimum principle in physics, which implies that physical systems tend to settle into stable minimum-energy states.

Figure I.2 Structural versus metrical information. (A) An elephant structure with the metrical proportions of a human. (B) A human structure with the metrical proportions of an elephant (Leeuwenberg, 2003a).

Finally, both the Gestalt cues and the governing principle are claimed to be autonomous and innate, that is, they are not affected by knowledge represented at higher cognitive levels. In line with this is the empirical observation that grouping based on familiarity is actually the weakest Gestalt cue. A general consideration is that perception is an input processor that aims at acquiring knowledge about the structure of the world around us. This perceptually acquired knowledge may, subsequently, be enriched by knowledge represented at higher cognitive levels (e.g., before actions are undertaken), but this is a post-perceptual issue. In this respect, notice that knowledge is a factor external to stimuli whereas, as a rule, the Gestalt cues refer to internal geometrical attributes of stimuli.

Basically, SIT shares the above-mentioned Gestaltist claims. SIT assumes cues for grouping stimulus elements, but instead of 113 cues, it assumes just three cues. These three cues refer to geometrical regularities such as repetition and bilateral symmetry (see Chapter 5 for a theoretical foundation). This restriction is mainly due to SIT's focus on structural rather than metrical pattern aspects. Structural aspects deal with categories such as present versus absent features, whereas metrical aspects refer to quantitative variations within categories. Figure I.2 illustrates the difference. The figure at the left presents an elephant structure with the metrical proportions of a human. The figure at the right presents a human structure with the metrical proportions of an elephant.

SIT's focus on geometrical structures indicates that it shares the Gestaltist claim about the knowledge independence of perception. This claim about the autonomy of perception not only applies to ontogenetic knowledge (i.e., knowledge acquired during one's life), but also to phylogenetic knowledge (i.e., knowledge acquired during the evolution). This contrasts with the Helmholtzian likelihood principle, which assumes that

perception is guided by such knowledge. That is, it states that the visual system selects the stimulus organization that agrees most probably with the distal (i.e., actual) object that gave rise to the proximal (i.e., retinal) stimulus. It is true that the likelihood principle is appealing in that it would yield veridical (i.e., truthful) percepts, but it requires probabilties that are hardly quantifiable, if at all (see Chapter 5 for arguments that SIT's approach yields veridical percepts just as well).

The Gestaltists were not concerned with combining grouping cues of different strengths, but SIT is. SIT focuses on stimulus descriptions which specify the contributions of cues to candidate stimulus organizations. Such a description not only represents a stimulus in the form of a reconstruction recipe, but also represents a candidate organization of the stimulus and thereby a class of stimuli with the same structure. This is different from, but yet reminiscent of, Garner's (1962) ground-breaking idea that the visual system, when presented with a stimulus, infers a class of structurally similar stimuli.

Like the Gestalt approach, SIT assumes a governing principle, or criterion, for selecting the perceptually preferred stimulus organization. A difference is that SIT conceives of 'simplicity' as the pivotal concept that includes 'stability', 'balance', and 'harmony'. SIT's simplicity principle agrees with the descriptive minimum principle proposed by Hochberg and McAlister (1953) which, in turn, can be seen as as information-theoretic translation of Koffka's (1935) law of Prägnanz. The simplicity principle implies that the visual system tends to select the stimulus organization that can be described using a minimum of structural information parameters. This structural information load, or complexity, of descriptions can be quantified in a fairly objective way, so that SIT enables falsifiable predictions about perceptually preferred stimulus organizations.

In hindsight, SIT can be seen as a perception-tailored version of the domain-independent mathematical approach called algorithmic information theory (AIT, or the theory of Kolmogorov complexity; see Li and Vitányi, 1997). Historically, however, SIT and AIT developed independently since the 1960s (they interacted only since the 1990s; see Chater, 1996; van der Helm, 2000), and both can be seen as viable alternatives for Shannon's (1948) classical information theory which had been developed in communication theory. Whereas Shannon's approach, just as the above-mentioned likelihood principle, requires probabilities that are often hardly quantifiable, both SIT and AIT resort to descriptive complexities which, as mentioned, can be quantified in a fairly objective way. Furthermore, SIT's simplicity principle corresponds, in AIT, to the so-called minimum description length principle. In fact, both principles can be seen as modern formalizations of William of Occam's (±1290–1349)

idea, known as Occam's razor, that the simplest interpretation of data is most likely the best and most favoured one.

There are also crucial differences between SIT and AIT, however. First, SIT makes the perceptually relevant distinction between structural and metrical information (see Figure I.2), whereas AIT does not. Second, SIT encodes for a restricted set of perceptually relevant regularities whereas AIT allows any imaginable regularity. Third, in SIT, the perceptually relevant outcome of an encoding is the stimulus organization induced by a simplest code and this organization establishes the objects we perceive, whereas in AIT, the only relevant outcome is the complexity of a simplest code.

In modern cognitive science, also connectionist and dynamic systems approaches trace their origin back to the Gestaltist ideas (cf. Sundqvist, 2003). In contrast to these approaches, which focus on internal cognitive and neural mechanisms of the perceptual process, SIT focuses on characteristics of the outcomes of this process. In fact, also SIT assumes that the outcome (i.e., the mental representation of a stimulus, or its percept, or its Gestalt) is reflected by a relatively stable cognitive state during an otherwise dynamical neural process. Dynamic systems approaches rightfully focus on the transitions from one neural state to the next, and connectionist approaches rightfully focus on the cognitive mechanisms leading to relatively stable cognitive states, but SIT prefers to focus on the perceptual nature of such relatively stable cognitive states. This may clarify why SIT's selection criterion is not stated in terms of process mechanisms but in terms of process outcomes. A pragmatic reason is that, empirically, these outcomes are better accessible than the internal mechanisms. A more fundamental reason is that, before modeling a process, one should have a clear picture of the outcomes that should result from this process.

The latter indicates that SIT is primarily a theory at what Marr (1982) called the computational level of description, that is, the level at which the goal of information processing systems is described. SIT focuses less on the algorithmic level, at which the method (i.e., the cognitive mechanisms) is described, and even less on the implementational level, at which the means (i.e., the neural mechanisms) are described. Of course, eventually, perception research should arrive at compatible descriptions at all three levels, explaining how the goal is obtained by a method allowed by the means (see Chapter 5 for steps in this direction). The purpose of this book is to give an overview of SIT's contribution to this scientific endeavour.

Part I

Towards a theory of visual form

In Part I, we discuss a number of visual form phenomena. Our intention is to show how structural information theory (SIT) assumptions may emerge step-by step from explanations of these phenomena.

In Chapter 1, the role of the input and output of perception is considered. An extreme position about the role of the input is to assume pure bottom-up effects in the sense that patterns are represented just stimulus analogously. An extreme position about the role of the output is to assume pure top-down effects in the sense that perception is completely guided by acquired knowledge. Both positions are criticized. The conclusion is that there is a stage of pattern interpretation preceding pattern recognition.

In Chapter 2, we deal with the question of which attributes of patterns are described by their representations. These representations are supposed to reveal visual pattern interpretations and segmentations. Four kinds of attributes are considered and compared with each other, namely, dimensions, features, transformations, and Gestalt properties. It is argued that only the latter attributes contribute to candidate pattern representations, and that they require a criterion to select appropriate representations.

Chapter 3 starts from the relevance of Gestalt cues, and focuses on their visual role. To this end, we compare two kinds of criteria for the selection of the actually preferred pattern representation. One kind of criteria applies to the selection process itself and the other kind applies to its output, that is, to the final pattern representation. Arguments are presented against process criteria and in favour of representational criteria.

In Chapter 4, we compare two models that assume representational criteria. One model derives object descriptions from object components, and the other model derives object components from object descriptions. Arguments are presented against the former model. We also contrast two representational selection criteria, namely, the likelihood principle and the simplicity principle. Arguments are presented against the former principle.

In Chapter 5, we summarize the insights that emerge from the preceding chapters, and we present the basic assumptions in SIT's coding model. We further give an overview of the theoretical foundations underlying SIT, regarding the veridicality of simplest codes, regarding the regularities to be extracted in the coding model, and regarding the implementation of the coding process in the brain.

1 Borders of perception

Introduction

Perception can be seen as the process that bridges the gap between incoming stimuli and already stored knowledge. The question here is to what extent perception shares properties of these two ends of the bridge. If it does so to an extreme extent, mental pattern representations are stimulus analogous, or biased by knowledge acquired by earlier observations, or both. In order to arrive at an appropriate global definition of perception, we consider pros and cons of each option separately.

1.1 The stimulus

In favour of stimulus-analogous coding

Reasoning involved in solving riddles is time consuming, requires mental effort, is under conscious control, and can be improved by training. In contrast, perception is rapid, effortless, automatic, and rigid. This multiple contrast may suggest that perception and reasoning are opposed in every respect. This may lead to the conclusion that reasoning is a process of interpretation and classification, whereas perception merely is a registration process that records and stores incoming information the way photographs do. Indeed, seeing is not felt as having to choose, for instance, whether a dark colour stems from a dark paint or from a shadow, or whether two stone parts stem from one stone occluded by a branch or from two separate stones. Chairs, tables, and doors are not experienced as mental constructs but as objects belonging to the external reality the perceiver looks at. After all, they remain present and tangible when closing one's eyes. This introspective argument could be taken to support a stimulus-analogous character of mental representations.

Furthermore, there is the phenomenon that different projections of the same object are not always recognizable as stemming from the same object. Figure 1.1 gives an example. It presents eight views of a tubular

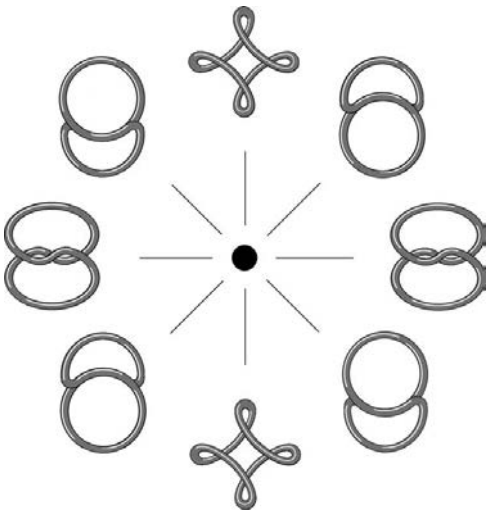

Figure 1.1 A tubular object at the centre of a roundabout, as seen from eight different directions. Even visually trained subjects are unable to infer one view from the other and to infer that views from opposite directions are identical (sculpture by Anneke van Bergen).

object standing at the centre of a roundabout in Beuningen, a Dutch village near Nijmegen. The height, width, and depth of the object are about the same, and the depicted projections agree with the views one gets from different directions; notice that views from opposite directions are identical. The views at the top and at the bottom of Figure 1.1 probably reveal the 3-D structure of the object most perspicuously. With some effort, these views can be inferred from each other. However, even visually trained people are not able to infer the other six views from these views. Also this viewpoint dependency in object recognition could be taken as supporting a stimulus-analogous character of pattern representations (Tarr, 1995).

It is true that less complex objects can usually be recognized viewpoint independently, yet one condition in the experiment by Shepard and Metzler (1971) seems to support that also simple objects are represented stimulus analogously. In each trial, they presented the projections of two objects in different orientations. In terms of not only the perceived 3-D shapes but also the presented 2-D projections, the two objects were either equal or mirrored. The task was to judge whether the two objects are equal or different. For instance, Figure 1.2A depicts a pair of equal objects, being equally handed like two left-hand gloves. Furthermore,