Arbib Chapter: Early Vision

A.L. Yuille and D.K. Kersten

No Institute Given

Abstract. In this chapter, we present a modern perspective and conceptualization of early vision in terms of computational models described using the mathematics of filtering, probabilities and graphical models. This mathematical framework has the advantage that it is increasingly being used in computer vision, in the modeling of neurons and neural circuits, in models of human visual behavior, and in the analysis of neural data by statistical and machine learning techniques. Hence it enables us to describe vision from multiple perspectives from a unified framework.

Introduction

This chapter has six sections. Section (1) gives an overview of early vision and introduces the basic concepts. In section (2) we describe linear filter models of early vision and the types of computations which they can perform. Section (3) introduces probability models for neurons and for computational tasks. In section (4) we describe neural network models that take into account spatial context and can perform complex tasks. Section (5) discusses how different visual cues can be combined taking into account their dependencies. Finally, we summarize the chapter in section (6) and briefly sketch the relations to high-level vision. The first five sections are linked to demonstrations which we introduce at the start of each section.

1 Basic Concepts and Overview

We start by introducing the study of vision in section (1.1). Next we discuss visual tasks and modules in section (1.2). Then we discuss the visual system and the brain in section (1.3).

For this section, we suggest that the reader explore the fascinating demonstrations of visual phenomena on website: `http://michaelbach.de/ot/`. This website was created and maintained by Prof. Michael Bach who accompanies the demos with descriptions and explanations of the phenomena. We particularly draw attention to: (i) Hidden Figures, (ii) Rotating Face Masks, (iii) Ames Window, (iii) Neon Color Spreading, (iv) Dress Code Enigma, (v) Adelsons Checker-Shadow Illusion, and (vi) Biological Motion. In addition, we encourage the reader to familiarize themselves with IPython Notebook in preparation for the interactive demos in later sections by going to website: `http://www.nature.com/news/interactive-notebooks-sharing-the-code-1.16261`.

1.1 What is Vision? Why is It Hard? How is it Studied?

What is Vision? Vision is the process of extracting information from retinal input, and we will here focus on images of visual scenes (whether natural or artificial). These images may capture a moment, or extend over time, as in a video. A human with normal vision can rapidly glance at an image and detect the objects in it and also decide what the object is made of, its distance and orientation relative to other objects, how it is moving (especially but not only if the image is dynamic), if it might collide with other objects, and whether to catch it or avoid it. Figure (1)(A) illustrates how much information we can get from a single image, despite the fact that images are locally highly ambiguous, as shown in figure (1)(B). In short, humans can estimate a rough approximation to the three-dimensional scene that has generated the image. But this is only one aspect of vision. In addition, humans have the ability to rapidly attend to different regions of the scene and ignore the rest. Vision is also used to enable actions, such as grasping objects or determining where to put your feet while hiking. In

summary, vision performs a range of *visual tasks* which extract information from the scene in order to achieve goals.

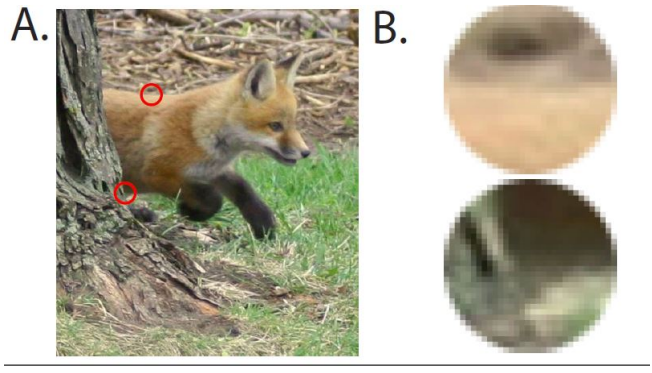

Fig. 1. (A) Humans can extract a lot of information from a single image. E.g., "There is a young fox emerging from behind the base of a tree not far from the view point, it is heading right, stepping through short grass, and moving quickly. Its body fur is fluffy, reddish-brown, light in color, but with some variation. It has darker colored front legs and a dark patch above the mouth. Most of the body hairs flow from front to back." B. Images are locally ambiguous. These two patches correspond to small parts of the fox's back and the side of the tree, see red circles in (A), but are highly ambiguous without context.

Vision is extremely difficult. This is perhaps surprising because humans find it very easy. You simply open your eyes and you can understand the scene without any apparent effort. But this is only possible because a very large part of your brain is involved in doing vision. It is estimated that roughly forty percent of neurons in the cortex are involved in visual processing. Despite decades of work, current machine vision systems perform significantly worse than humans, except for a few highly specialized tasks. But probably most other animals, except monkeys and our other close relatives, get far less information from vision judging by the much smaller numbers of neurons they devote to vision (e.g., the visual cortex of a mouse contains fewer neurons than a human or a monkey, by many orders of magnitude).

Why is Vision hard? The Complexity Problem and Natural/Ecological Constraints. To get some appreciation of the difficulty of vision, consider how the image of the Fox, shown in Figure (1), appears to the retina. Figure (2) shows the magnitude of the intensity of the image as a function of spatial position. These intensity magnitudes are proportional to the number of photons, or magnitude of light rays, that are imaged at different positions in the retina. The human visual system must somehow decode these intensity patterns and

determine that they are caused by a fox emerging from behind a tree. But based on these intensity patterns, it is hard to perform *visual tasks* such as *segmenting the image* into regions corresponding to different objects, performing *object recognition* to determine that a region of the image corresponds to a fox and another region to a tree, or performing *depth estimation* to determine the positions of the objects in the visual scene. These tasks are particularly complex because the intensity patterns will change significantly if we make small changes to the visual scene. The patterns will vary greatly if we alter the pose of the fox, the lighting conditions, the viewpoint of the observer, and how much the fox is occluded by the tree.

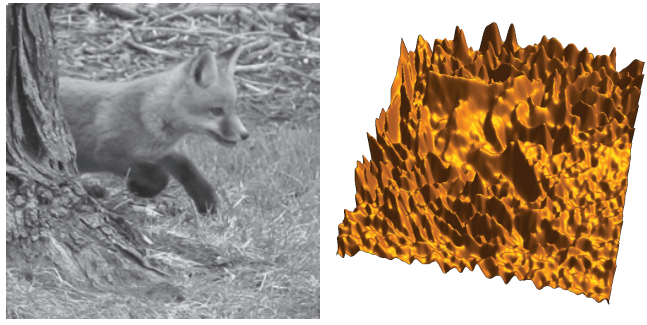

Fig. 2. Why is vision hard? The raw input to the Fox image (left panel) is the intensity values plotted as a function of spatial position (right panel). These intensity patterns vary depending on the pose of the fox, the lighting conditions, and other factors. The human visual system must decode this raw input, which is extremely difficult.

Hence the main challenge of vision is due to the enormous complexity of natural images, and their local ambiguities shown in figure (1)(B). The number of possible images, or intensity patterns, that can be described by a small image array with 100×100 positions, or *image pixels*, is $(256)^{10,000}$ which is astronomically large [77]. These images are caused by very large number of possible objects (estimated from 20,000-200,000) which can be arranged in a scene in an extremely large number of ways (in terms of the pose of the objects, their positions relative to each other), and be illuminated in an enormous number of different ways. Computer graphics studies how images can be generated if the scene is known. But vision must be capable of performing the much harder *inverse inference* task of determining the scene from the image. From this perspective, it is almost miraculous that humans can simply open their eyes and recognize objects and visual scenes within a few hundred milliseconds (which is roughly the time to blink your eyes).

In order to perform these visual tasks the visual system must be able to detect and exploit regularities in image patterns. These regularities are due to the structure of the world that we live in which cause the images which we ob-

serve. They include the assumption that surfaces are generally spatially smooth, that objects tend to move rigidly, that most scenes contain a ground plane and objects touch the ground plane at contact points. These assumptions have been called ecological, or natural, constraints [43,109]. Many of these constraints are "generic", in the sense that they are independent of the specific objects and object configurations in the scene, and so can be applied to perform some visual tasks such as grasping an object without needing to recognize what the object is. It is speculated that humans have learnt to exploit the structure of natural images and the world through evolution [45], early development [76], or by learning later in life [49]. Vision science researchers can learn these image regularities by applying machine learning methods to image datasets. Figure (3) gives an example which illustrates both the ability of humans to exploit constraints about the three-dimensional world to perform inverse inference and the mistakes which can arise when the constraints are violated.

Fig. 3. This image gives the illusion of a flying carpet where the woman on the towel is perceived to be floating above the beach. This illusion shows that humans use constraints – about ground planes, shadows, and contact points – to interpret images. But in this case, the constraints are violated because we incorrectly think that the shadow is being caused by the towel. Instead the shadow is cast by a flag which is outside the image.

How is Vision Studied? Broadly speaking, vision can be studied in three related ways: (i) at the "behavioral" level by studying how well humans, and animals, can perform visual tasks, (ii) at the "neural" level to understand the neural mechanisms (by electrode recording or by non-invasive methods like fMRI), and (iii) at the "computational" level by designing mathematical models and computer vision algorithms that can perform visual tasks. We distinguish between those mathematical models which attempt to describe how humans or animals

see by accounting for behavioral or neural data, and those, called *computer vision*, whose goals are to perform visual tasks without attempting to model how humans, or animals, perform them. There is a complicated symbiosis between these two approaches to mathematical modeling of vision (which are done by different research communities). What they have in common is the need to address similar visual tasks and to deal with the complexity of image patterns. We note that computer vision researchers, even those who have no interest in biology, can nevertheless yield insights into human vision by developing algorithms which can perform the same visual tasks as humans in similar environments. This leads to a strategy of understanding the brain by reverse engineering.

The visual system is so complex that vision scientists must make simplifications to break it down into manageable pieces. These include: (i) studying visual tasks in isolation instead of addressing the complete visual system, (ii) simplifying the visual stimuli, (iii) simplifying the models of neurons and neural circuits, and (iv) simplifying the overall structure of the visual areas of the brain and how they interact with each other. Although these simplifications are necessary they raise concerns which we will keep returning to in this chapter.

Firstly, vision researchers break vision down into different visual tasks which can be studied separately. These tasks include image segmentation, depth estimation, and object recognition. They are performed by *modules* which output *representations*. Modules, however, may not be localized to distinct parts of the brain. These simplifications assume that modules are semi-independent, which can be questioned (see next section). They also raise the question of how the modules interact with each other. Marr's influential framework for vision [109], see figure (4)(left), gives one way to address these issues. Marr proposed that the human visual system uses modules which compute a sequence of representations of the image which start with a *primal sketch* of the image, proceed to a $2-1/2$-*D sketch* which represents the three-dimensional structure of the scene, and concludes with a *3-D* representation of objects. Hence the modules interact by outputting representations which are used as inputs to other modules. Marr's framework was never fully developed but it does capture some important aspects of the visual system. It also illustrates the important classification of visual tasks into: (i) *low-level* vision which processes the image (e.g., produces the primal sketch), (ii) *mid-level* vision which estimate the structure and properties of geometric surfaces (e.g., produces the 21/2-D sketch), and (iii) *high-level* vision which recognize objects and analyzes scenes. This classification is shown in figure (4)(right).

Secondly, the set of all possible stimuli is astronomically large. Hence it is impractical to study visual systems behaviorally by their response to all stimuli. In addition, when studying specific visual tasks it is sensible to use stimuli which only contain information, or *visual cues*, which are specific for these tasks. Also good experimental design requires, if possible, controlled stimuli so that the difficulty of performing a specific task can be quantified in terms of varying a small number of variables (e.g., by making the images darker or brighter). For these reasons the study of vision is often simplified by using synthetic stimuli.

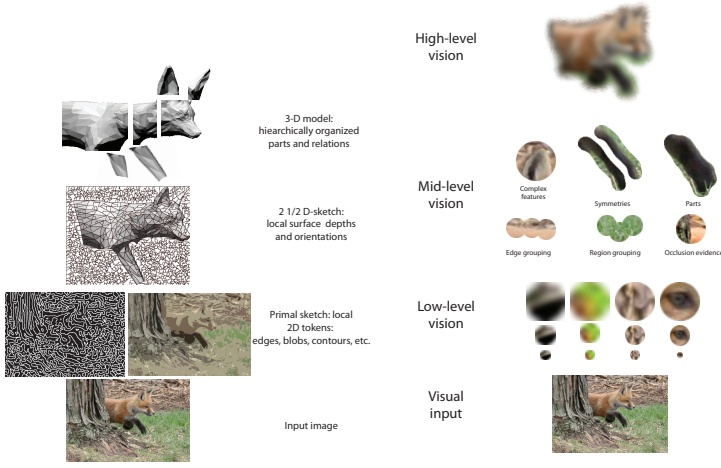

Fig. 4. Marr's framework for vision (left panel) consists a series of representations. Visual tasks can be classified into low-, mid-, and high-level tasks (right panel). This classification roughly relates to Marr's framework.

For example, the ability of humans to perceive depth from Julesz's random dot stereograms [71] demonstrate that humans can perceive depth when objects are not present, see figure (5). Too much reliance on synthetic stimuli, however, can be misleading and there is concern that experimental findings on synthetic stimuli may not generalize to human, or mammalian, abilities in more natural situations [19], [188]. After all, human and animal visual systems have evolved to perform complex tasks on complex stimuli, and so understanding them requires probing it with stimuli which captures this complexity.

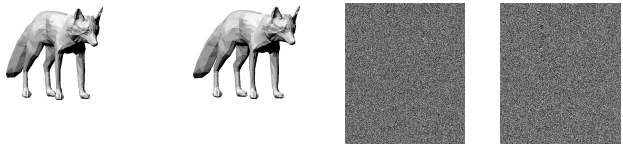

Fig. 5. Binocular stereo of the fox for real images and for Julesz random dots stereograms. The left two images are a stereo pair (the left and right images) of a fox so that, when fused (e.g., by a stereo viewer presenting the left and right images to the right and left eyes, respectively), they yield the three-dimensional shape of the fox. The right two images are stereo pairs of random dot images of a fox. When fused, they also give the three-dimensional shape of the fox.

Thirdly, simplifications must be made when modeling neurons and neural circuits. The detailed biophysical properties of single neurons are very complicated and so simpler models are used, particularly when modeling visual tasks. The most commonly used model is *integrate-and-fire*, when the neuron fires an action potential if a weighted linear sum of its inputs is above a threshold. We will use this model in this chapter but it should be acknowledged that real neurons are more complicated. Their output may be a highly non-linear function of the input and they may signal information by a sophisticated "neural code" involving the precise timing of action potentials. Moreover, recent studies also show that there are large varieties of neurons which differ in their anatomy and function. There is also evidence that neural circuits can behave differently in different situations.

Fourthly, the numbers of neurons involved in visual perception is extremely large. This means that simplifications need to be done when studying the overall structure of visual areas of the brain and how they relate to each other. Although the broad structure of some visual areas are known we are a long way from having wiring diagrams describing the connections between neurons within each visual area (although there are research programs to obtain such wiring diagrams for the mouse, whose visual cortex is many orders of magnitude smaller than that of a human or monkey).

1.2 Visual Tasks and Modules in Early and High Level Vision

This section gives a brief introduction to visual tasks, modules, and the distinctions between low-, mid-, and high-level vision. We discuss these tasks only at the behavioural level and postpone discussion of the brain until the next section. This chapter is about early vision, which we define to be low- and mid-level vision. But we also discuss how these early levels relate to high-level vision, which is described in the chapter by Lewis and Poggio.

Low-level visual tasks involve estimating local properties of the image. They include finding the boundary of an object (without deciding what the object is) and also include estimating the motion flow. Mid-level visual tasks estimate properties of geometrical surfaces, the shape and position of surfaces in the visual scene, and their depth ordering. High-level visual tasks estimate properties of objects, scene structures, the relationships between objects, and the actions of the objects. In addition, each level provides information which is passed on to the next level, as illustrated by Marr's theory. Another way to think of this organization is in terms of the knowledge available at different levels. For example, low-level vision can be performed by a system which knows only about regularities of image patterns (e.g., that images typically consist of regions where the intensity changes slowly which are separated by *edges* where the intensity changes rapidly). Mid-level vision processing knows about properties of geometric surfaces (e.g., that they tend to be spatially smooth) and that they can partially occlude each other. High-level knows about objects, the relationships between them, and actions. Hence the flow of information from low- to high-level vision is from generic to specific.

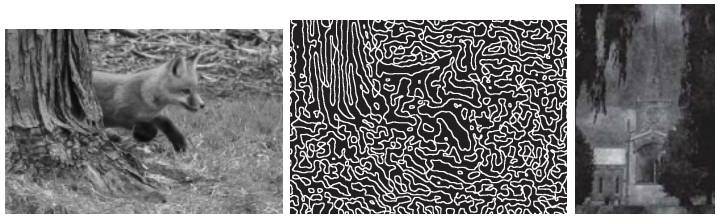

Fig. 6. The edges of the fox image (left panel) detected by low-level processing (center panel). Some edges lie on the boundary of objects, like the fox and the tree, while others are due to properties of the textures (e.g., the grass or the bark of the tree). It is difficult to distinguish between these different types of edges. The church steeple (right panel), and the position of its edges, is obvious if you view the whole image. But it is almost impossible to see locally because there is no strong local evidence for the edges of the steeple. This shows that sometimes edge detection is impossible except when done in conjunction with object detection. One possibility is that low-level vision proposes many possible edges which are validated, or rejected, by object models, as discussed in section (4).

Low-Level Vision. Low-level vision can be roughly defined to be the visual processing which can be done without explicit knowledge of objects and the three-dimensional structure of the world (although this additional knowledge, if available, will improve performance) Typical low-level vision tasks include determining differences within the image, such as detecting edges and performing segmentation, see figure (6). Low-level vision also involves extracting representations of image patterns which can be used for higher-level processing. As discussed later in the chapter, low-level vision can exploit statistical properties of images which are true for most images (e.g., that images tend to be piecewise smooth).

Low-level vision also includes estimating the local motion of images. This involves finding the correspondence between points in images taken at different times. This is done by matching regions which have similar intensity properties. But this has problems at places where there are many, equally good, matches. Figure (7) illustrates the *aperture problem* which is solved by making assumptions that the motion is usually slow and smooth.

Although these low-level visual tasks can be studied in isolation there is much evidence that they interact with other higher level tasks. For example, it is usually impossible to detect the edges of the objects in an image without making mistakes using low-level processing alone. Instead a better strategy is for low-level to *propose* a set of possible positions for edges, which can be validated by object models, see figure (6)(right). But there are many visual phenomena indicating the complexity of the interactions between low-level and high-level vision, such as the dalmatian dog illusion where the extreme ambiguity of low-level cues for edges make it very hard to detect the dalmatian, see `http://michaelbach.de/ot/cog_dalmatian/`.

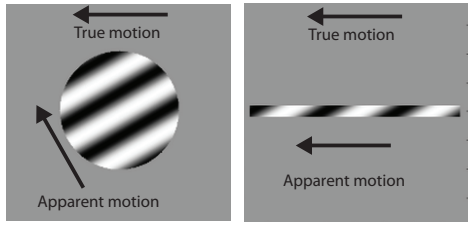

Fig. 7. These images show black and white bars whose *true motion* is leftwards viewed through two apertures (circular and rectangular). But the motion is locally ambiguous because we can only directly observe the motion component normal to the bars (we cannot detect if there is any motion tangential to the bars) and so the observed stimuli is consistent with many possible motions. The human visual system uses constraints to resolve these ambiguities. For these stimuli, humans assume that the motion is as slow as possible and hence is perpendicular to the bars (i.e. assuming that the unobservable tangential component is zero), as indicated by the *apparent motion*. More generally, as discussed in section (4.4) humans tend to assume that motion is slow and smooth. See demo 4e.

Mid-Level Vision. Mid-level vision builds on low-level vision. It corresponds, roughly speaking, to vision processing that knows about geometry, materials, and lighting. But it does not know about specific objects or scene structures. For example, mid-level vision knows about surfaces of red metal, but does not know about red cars. Mid-level vision includes many visual tasks, often formulated in terms of modules, and this section introduces some of them,

Mid-level vision includes inferences about depth ordering of surfaces and, in particular, knowing that surfaces can partially occlude each other. The Kanizsa figure (8) is perceived as three black disks which are partially occluded by a triangular surface. But close inspection shows that there is no direct evidence for the existence of the triangle, i.e. there are no intensity edges at the boundary of the triangle. The Kanizsa triangle is an example of a *Gestalt* grouping phenomenon, many of which can be explained in terms of a human tendency to interpret images as simple geometric structures [73].

Another example of how the human visual system uses geometry was shown earlier in figure (3). The human visual system assumes that most images contain a ground plane with objects standing on it (e.g., a man standing on a lawn). By the laws of *perspective projection*, if there are parallel lines in the image (such as the tracks of a railway line) then the projection of these lines in the image will converge at a *vanishing point*. Humans can use vanishing points to estimate the orientation of the ground plane. The contact points of the objects with the plane specify the positions of the object in the scene. More information about the scene can be extracted if the image contains several vanishing points corresponding to surfaces which are orthogonal in space.

Binocular stereo is another vision module which estimates the depth and orientation of surfaces. Humans have the ability to get depth from two eyes – hence the popularity of so-called 3D movies. This is illustrated in figure (5). It

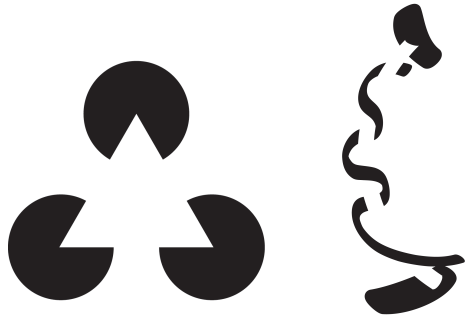

Fig. 8. The Kanizsa triangle (left panel) is perceived as a white triangular surface which partially occludes three black disks, it shows the tendency of humans to interpret images in terms of geometric structures. Another example of this (right panel) shows that human tend to "explain away" gaps in the black band by positing a white band which lies sometimes above and sometimes below the black band. For an interesting variant, see `http://www.michaelbach.de/ot/cog-kanizsa/`, which shows how the effect can disappear if other cues are present.

requires solving a *correspondence problem* between features in the two eyes which are caused/imaged by the same point in space. If correspondence can be performed, then the depth can be estimated by trigonometry. The correspondence problem is difficult, see section (4.3) and seems to assume that the geometric surfaces being viewed as spatially smooth.. The correspondence problem is made easier because of the *epipolar line constraint*, see figure (9), which means that corresponding points requires only searching in a one dimensional direction. But knowledge of the epipolar lines requires knowing the direction of gaze of the cameras (maybe done by feedback from muscles controlling the eyes, or by *calibration*). Note, that partial occlusion can happen, where part of the scene is only visible to one eye. Da Vinci was the first to point out that this was a useful visual cue.

Humans can also get three-dimensional shape information of surfaces from shading, texture, and even contours. These are often studied as separate modules. Typically models of shape from shading assume that surfaces can be modeled as matte (i.e. dull and un-shiny) and hence their reflectance properties (i.e. their tendency to reflect light) given by Lambert's law [9]. This enables us to predict the image intensity of an object in terms of its geometry and the lighting conditions. In some conditions this can be inverted to estimate the shape of the object from the intensity patterns, which is called *shape from shading* [9]. Shape from texture arises if a surface has a regular pattern of texture. This pattern will be distorted due to the shape of the object, which enables the shape of the surface to be estimated from the intensity patterns by *shape from texture* [83]. Finally, certain contours naturally suggest shapes, which is called *shape from contour* [83].

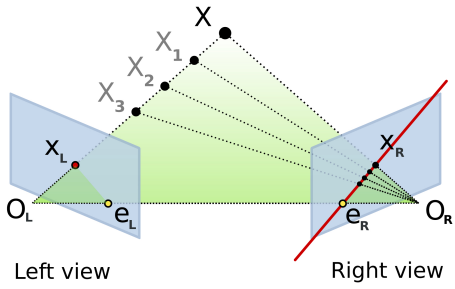

Fig. 9. Stereopsis and epipolar lines. A point $\boldsymbol{x}$ in three-dimensional space gets projected onto positions x_L and x_R in the left and right eyes. This uses a pinhole camera model of each eye where the eye is specified as a plane (in grey) and O_L and O_R represent the centers of projection. Observe that all points on the plane defined by O_L, O_R and $\boldsymbol{x}$ get projected onto straight lines $\boldsymbol{e}_L$ and $\boldsymbol{e}_R$, the corresponding *epipolar lines*, in the two eyes. To illustrate this we show the projections of a few points x_1, x_2, x_3 onto the right eye. If we alter the position of the point $\boldsymbol{x}$ in space then we will get a family of corresponding epipolar lines. The *epipolar line constraint* states that points on an epipolar line in one eye can only be matched to a point in the other eye on the corresponding epipolar line. In this picture the eyes are fixating at a point. But if the eyes are parallel to each other the fixation point will be at infinity and the epipolar lines will be parallel to each other in the two eyes (or images).

In addition to detecting the shape of surfaces humans are also able to estimate their material properties, see [16]. This enables us to tell whether a teapot is metallic (e.g., by being shiny) or matte. This ability has many uses. For example, if a surface is metallic then it could be part of an airplane but it cannot be part of an animal. If a patch on a road is very shiny then there is a possibility that it is ice, at least in winter, and hence it better not to step on it. In general, the ability to detect material properties and textures is very useful when performing actions – e.g., is this object slippery? can I walk on it? will it make a sound if I do?– and as a pre-cursor for object recognition and scene understanding.

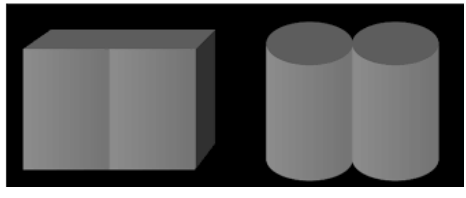

Fig. 10. The shapes of the contours affect the perception of the shading patterns. The intensity patterns are the same in both stimuli but the shape of the boundaries makes the perception of shading different in the two cases [84].

These vision modules depend on specific properties of images which are called *cues*, for example the shadow on the sand in figure (3). These modules, and the visual cues they rely on, can be modeled and studied separately. But many lines of evidence suggest that they are coupled. For example, Knill and Kersten [84] illustrate that shape from contour can alter the perception of shading patterns and material properties, see figure (10). In addition, careful studies of how images are formed in terms of the surface of the scene (the objects, their geometry, how they reflect light, the lighting conditions, and the viewpoint) it becomes clear that the *cues* are often coupled, see section (5.1). Indeed under certain lighting conditions it is extremely difficult to tell the color of an object being viewed, see `http://michaelbach.de/ot/col-dress/`. In addition, there is interaction between these modules and high-level vision. Humans perceive an inverted (i.e. concave) face mask to be a normal convex face even when binocular cues are present, see `http://michaelbach.de/ot/fcs_hollow-face/`. In general, object recognition trumps mid-level vision although Kanizsa shows some exceptions [73], see also `http://michaelbach.de/ot/sze-AmesBallerina/`.

1.3 The Visual System and the Brain

This section briefly overviews what we know about how the brain does vision. It reviews the areas of the brain which perform visual processing, the relationships between them, the structures of these areas, and the visual tasks they perform. Due to the complexity of the visual system our knowledge of these issues, though considerable, is limited. It is based on a combination of anatomical studies, electro-physiological recordings, and non-invasive imaging methods such as functional Magnetic Resonance Imaging (fMRI).

Retina and Lateral Geniculate Nucleus. Visual neural processing begins with the retina which transmits information to the visual cortex via the Lateral Geniculate Nucleus (LGN). The anatomy and electro-physiology of the retina and LGN have been studied in detail [116,47,15]. Although we will not discuss it further in this chapter, another pathway from the retina leads to the superior colliculus in the midbrain. This plays an important role in, e.g., eye movements (see Chapters 20, Saccades and Smooth Pursuit Eye Movements, and 23, Integrative Functions of the Corticostriatal System).

The retina converts intensity patterns – the light rays which reach the retina – into patterns of neural activity. This starts with *photo-receptors* which are directly activated by light. The photo-receptors have been studied in detail and it can be shown that they are extremely efficient at "capturing" photons [139]. The remaining set of neurons in the retina, in particular *ganglion cells*, process the photo-receptors output and encode it for transmission via the *optic nerve* to the rest of the brain (note, it is believed that this flow of visual information only goes one way although the brain does influence the muscles which control the eyes).

The retina is generally agreed to function as a sophisticated camera which captures the information in the incoming images and encodes it so that it can be

transmitted to the visual cortex (although the control of eye movements means that the retina is not a passive device like a typical camera and instead is actively searching for information). To do this, the retina must face two challenges: (i) the enormous variability of intensity in natural images, (ii) the ability to encode the images so they can be transmitted efficiently and robustly. Neural models of the retina are largely motivated by these two issues.

Firstly, the intensity of natural light varies enormously from faint starlight to bright sunlight dealing with intensity magnitudes ranging from 1 to 10^9. Moreover, the changes of intensity within specific images can also vary hugely [27]. But neurons have limited ranges of response and hence they cannot encode these huge ranges of intensity. Hence many theories of the retina, see section (2), propose that the ganglion cells perform *gain control* and filter the images so that they capture only the *local contrast* – the differences of intensity between nearby parts of the retina – and hence reduce the need to represent the entire intensity range. Observe that digital cameras must perform a similar function, since they convert incoming light patterns into digital representations where the intensity only takes 256 values from 0 to 255 (in each color channel).

Secondly, the retina must encode the image information so that it can be transmitted through the optic nerve to the rest of the brain for further processing. The image information is transmitted through a relatively smaller number of fibers in the optic nerve (compared with the number of photo-receptor cells). Information theory offers guidelines for how information can be encoded efficiently based on statistical knowledge of the stimuli. Researchers have applied this theory to predict retinal properties with some success but this work is out of scope of this chapter and we refer to the detailed description in [190].

In addition, the human and animal visual system does not passively receive images and process them. Instead it is active and by using eye and head movements, and attentional processes, it seeks out information useful for performing tasks. The retina only has high spatial precision at the center in the *fovea* and so high-resolution images can only be gathered there. Neurons in the periphery are more sensitive to motion and presumably alert the visual system to places to foveate to. This work is also out of the scope of this chapter and is addressed in [190].

The studies of the retina illustrate the "simplification issues" which will reoccur throughout our chapter. At the computational level, the theories for describing how the retina deals with intensity are simpler than the engineering methods used by computer vision and image processing researchers that deal with the same challenges. At the experimental level, many of the findings about retinal neurons are based on simplified models of neurons obtained from studying their responses to synthetic stimuli. Moreover, although there is considerable knowledge of the anatomy there has only recently been highly detailed studies of the wiring diagrams and characterization of the fifty or more anatomical types of neurons [111]. It is also unclear why so many neurons are required to perform the two functions using current theories. Indeed it has been argued that the retina is considerably "smarter" than current theories suggest [114,47]) and may require

detecting motion, expansion, extrapolation, and more generally adapting to the complexity of image patterns.

The output from the retina is transmitted to the Lateral Geniculate Nucleus (LGN) and then to the visual cortex where it arrives in *visual area V1*. LGN is generally believed to have the limited function of serving as a way station on the route to the visual cortex. Hence current models of LGN neurons are fairly simple, as discussed in section (2). But there is reason to believe that LGN is more complex. For example, there is substantial feedback from V1 to LGN [15], and connections between LGN and other areas than V1 [153,123].

Cortical Visual Areas and the relationships between them. The visual cortex can be decomposed into a number of visual areas based on anatomical and electrophysiological measurements [165]. Like all areas of the cortex, these regions have standard six layer structure. These visual areas V1,V2,V4, Medial Temporal (MT), Medial Superior Temporal MST, and the Inferior Temporal Cortex (IT) are illustrated in figure (11). There are numerous interactions between these visual areas, see figure (11), but it is common to concentrate on two *hierarchical streams*: (I) The *ventral stream* which consists of V1, V2, V4, (the functional organization of V3 has been under some debate) and then the infero-temporal (IT) areas of extrastriate cortex. This pathway is believed to perform object detection and scene understanding. (II) The *dorsal stream* goes from V1, MT to the parietal cortex. It is believed that this is used for analysis of the movements and positions of objects as the relate to navigation and actions [117]. Although the distinction between ventral and dorsal pathways is well-established [97] it is also probable that the true situation is more complicated [149].

The size of the visual areas varies greatly. The first two areas, V1 and V2 are enormous and together account for roughly seventy percent of the number of neurons in the visual cortex (hence thirty percent of the neurons in the entire cortex). The size of V1 is much bigger, by a factor of at least two hundred, than the number of fibers that leave the eye. Indeed it has been estimated that this is more by a factor of several hundred than the amount needed to represent the information conveyed by the LGN [97], consistent with the idea that the purpose of V1 is to start interpreting the image instead of simply encoding it. Another major feature of the hierarchy of visual areas is that their size get progressively smaller as one rises in the hierarchy. For example, V4 is much smaller than V2, and visual areas within IT are considerably smaller than V4.

The early visual areas have regular structural organizations. These are studied by using electrophysiology to classify the receptive fields of neurons by probing their responses to synthetic stimuli with different *perceptual dimensions* such as position, orientation, color, texture, shape, sensitivity to input from both eyes, and motion. Individual neurons often show preferences, i.e. respond more strongly, to specific input stimuli and hence are said to be *tuned* to these stimuli. In particular, the positions of stimuli on the retina which cause the neuron to fire is called its *receptive field*. Neighboring neurons in early visual areas usually

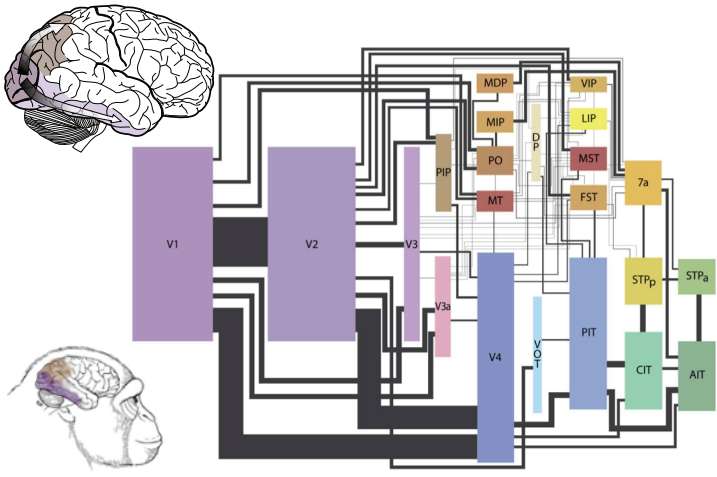

Fig. 11. Left panel (top and bottom) illustrate the monkey visual cortex. The right panel is a schematic of connections between visual cortical areas in the macaque monkey brain. The colored rectangles represent visual areas (see [34]). The black lines show the connections between areas, with the thickness proportional to estimates of the number of feedforward fibers. Areas in cool and warm tones belong to the ventral and dorsal streams, respectively. (Figure reprinted with permission from [172]; see also [98]).

respond to similar regions of the image. Hence these areas are roughly *retinotopic* in the sense that their spatial organization is similar to that of the image at the retina, but with a nonlinear (log-polar) spatial transformation [151]. This retinotopic structure is strongest in V1 and V2 and gets weaker at high visual areas (see following paragraphs). But although neurons often have small receptive fields, at least in V1 and V2, their tuning to other "perceptual dimensions" is often broader and neurons can responds to a range of stimuli. It is helpful to broadly classify neurons by which specific "dimension", if any, those neurons responds to, but this can become problematic, particularly in higher areas [143,141]. Even in V1 most neurons respond to several dimensions [97]. Mapping has also been done using optical techniques [105,81] which also show that most early visual areas are organized retinotopically, although this is strongest in V1 and V2. Finally, we note that until the last few years it has only been possible to record from a neuron for a limited amount of time (e.g., from an hour to at most a couple of days). This meant it was only possible to test the response of a neuron to a limited number of stimuli and stimuli dimensions. Hence some of these findings may be revised when it becomes possible to test the response of neurons to a larger and more representative set of stimuli, see discussion in section (2.3).

Other salient structures of V1 include *hypercolumns* (∼1-2 mm) consisting of: (i) a regular array of orientation columns, perpendicular to the cortical surface, in which orientation selectivity of neurons is approximately the same. The

orientation tuning varies slowly parallel to the cortical surface; (ii) ocular dominance columns (where the proportion of input from both eyes is constant within each column, but varies smoothly between columns), and (iii) a lattice of cytochrome oxidase blobs – sensitive to color [64,103]. Sections (2,3,4) describes computational models which exploit these structural properties. From a more abstract perspective, the organizational structures of hypercolumns can be partly explained by the need to map stimulus dimensions (e.g. retinal position, orientation, etc.) onto two-dimensional cortical surface while attempting to make the map as smooth as possible (this is not possible, on topological grounds, so discontinuities occur) [29]. Cells in V2 have larger receptive fields than in V1 but have other properties. For example, many cells in V2 seem capable of performing binocular integration to estimate depth. In addition, they appear to include neurons which respond to features which are linked spatially [168,132], which are sensitive to illusory contours[169,93], and which relate to detecting occlusion and establishing border ownership [191,33]. These issues are discussed in section (4).

A notable property of these visual areas is their hierarchical organization, which relates to the low-,mid-, and high-levels discussed in section (1.2). Broadly speaking, V1 and MT seem to be involved in low-level processing, V2,V4, MST in mid-level vision, and high-level vision in IT. Hence early vision is believed to be mostly performed in V1,V2,V4, MT, and MST. There is also a strong tendency for receptive fields to be larger as we ascend the visual hierarchy. Compared to those in V1, the receptive fields are 2-3 times bigger in V2, 4-5 times larger in V3/VP, and 7-10 times larger in MT. But, conversely, the receptive fields become increasing specific to stimuli, and stimuli of greater complexity, as we move up the ventral stream. In summary, the receptive fields become more invariant to position and more specific to structure as we proceed up the ventral stream from V1 to V2 to IT [144][104].

Experimental Methods. Many of the findings discussed above are based on electro-physiological studies of monkeys and non-invasive studies of monkeys and humans. These visual systems are fairly similar based on studies relating fMRI responses in humans and those in monkeys (where electro-physiology is performed). Relationships have been found in early visual areas V1, V2, V3 [173], but are not yet fully established at higher areas [174]. But it is noteworthy that the *face area* discovered in humans by fMRI studies [74] corresponds to an analogous face area in monkeys which can be studied by fMRI and electrophysiology [162].

The interpretations of experimental findings are limited because of the simplifications discussed in earlier sections. Non-invasive studies like fMRI suffer from limited spatial and temporal resolution and currently can only observe coarse properties of the visual system. Electro-physiology is restricted to recording from a small number of neurons in response to a limited range of stimuli. See [19] for a detailed discussion of the problems of interpreting electro-physiological results in the early visual cortex. In general, even in V1 it is not easy to predict the response of neurons to natural stimuli. This may, however, be partially

because neurons in the visual cortex interact with each other and may be better predicted by the activity of other neurons than by the image itself. All the visual areas contain lateral (i.e. sideways) structure (in addition to hierarchy) and neural connections within V1 extend as much as 8 mm. As will be discussed later in this chapter, it can be shown that these lateral connections may implement *non-classical receptive fields* which relate to visual tasks such as linking, or grouping, features [44]. There is considerable interest, and progress, in developing experimental methods which can probe the properties of neural circuits in much greater detail, such as opto-genetics, which may revolutionize our understanding the the early visual system.

2 Linear and Complex Filters

This section covers five topics. Firstly we introduce simple linear models of neurons, describe how they are used to model the receptive fields of neurons in the retina, LGN, and *simple cells* in V1, and discuss the visual tasks they can be used for. We also discuss extensions of these models to deal with *complex cells* in V1. Secondly, we discuss an alternative perspective which thinks of these cells as representing images and introduces over-complete bases and sparse encoding. Thirdly, we discuss how these receptive fields can be learnt from images by unsupervised algorithms or estimated from neural data by regression. Fourthly, we describe how these receptive field models can be used for binocular stereo and for motion estimation. The section also introduces background material on linear filter theory and fourier analysis.

This section includes three interactive demos: (2a) Linear filters and convolution. (2b) Gabor filters. (2c) Oja's Rule and Principal Component Analysis.

2.1 Linear Models of Neurons

This section introduces linear models of neurons. These models are used to describe the receptive field properties in the retina, the LGN, and simple cells in visual area V1. We introduce background material on linear filtering and fourier analysis which helps to understand the properties of these neural models.

Linear Models of Simplified Cells. This section *introduces* a model of a simplified cell, see figure (12). The cell receives inputs $\mathbf{I} = (I_1, I_2, ..., I_N)$ from *dendrites* which are weighted by *synaptic strengths* $\mathbf{w} = (w_1, w_2, ..., w_N)$, these are summed together at the *soma* (cell body) to obtain $\mathbf{w} \cdot \mathbf{I} = \sum_{i=1}^{N} w_i I_i$. The cells output a response $f(\mathbf{w} \cdot \mathbf{I})$ along its *axon*, indicated by the firing rate of the neuron. $f(.)$ is monotonic non-linear function, which takes value 0 if the input is small, then increases linearly in the *linear regime* until it saturates at a maximum value. A typical choice of $f(.)$ is the sigmoid function $f(\mathbf{w} \cdot \mathbf{I}) = \sigma(\mathbf{w} \cdot \mathbf{I} - T)$, where T is a threshold and $\sigma(.)$ is a soft-threshold as shown in figure (12), which we will see in later sections. Interactive demo (2a) illustrates linear filters and convolution.

In this section, we ignore $f(.)$ (except when explicitly stated) and study the behavior of the model in the linear regime. Cells in the retina and Lateral Geniculate Nucleus (LGN) are often modeled without the non-linear function $f(.)$, but adding instead a constant C to the output, to account for spontaneous firing of the cell, and yielding an output $\mathbf{w} \cdot \mathbf{I} + C$, see [190].

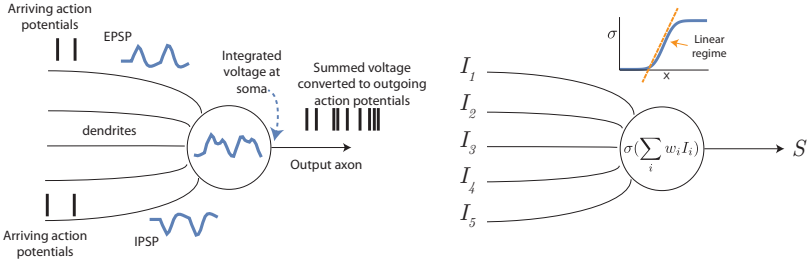

Fig. 12. Left Panel: A neuron receives input – action potentials from other neurons – at its dendrites which generate excitatory and inhibitory postsynaptic potentials (EPSPs and IPSPs respectively) whose voltages are integrated at the soma and converted to outgoing action potentials. Right panel: a simplified model of a neuron. There are inputs $(I_1, ..., I_5)$ at the dendrites, with synaptic strengths $w_1, ..., w_5$, these are summed at the soma, $\sum_i w_i I_i$, and the output S is given by a sigmoid function $\sigma(\sum_i w_i I_i)$. The sigmoid function $\sigma()$ (top right) has a linear regime (brown line) and low- and high-thresholds.

Hence in this section, we model a simplified cell by:

$$S = \mathbf{w} \cdot \mathbf{I} = \sum_{i=1}^{N} w_i I_i.$$

This model is linear in two respects. Firstly it is linear in the input $\boldsymbol{I}$ so that if we double the input $\mathbf{I} \mapsto 2\mathbf{I}$, then the output doubles also $S \mapsto 2S$. Secondly, it is linear in the weights $\mathbf{w}$. Most importantly, it obeys the *principle of superposition* so that if S^1, S^2 are the outputs to input $\mathbf{I}^1, \mathbf{I}^2$ respectively, then the output to input $\lambda_1 \mathbf{I}^1 + \lambda_2 \mathbf{I}^2$ is $\lambda_1 S_1 + \lambda_2 S_2$. This result is important for characterizing the response of simple neural cells, since it implies that we can determine the output of the cell to any stimulus by observing its response to a limited set of input stimuli $\mathbf{I}$, and we will return to this issue later in this section. Note that this property still remains if we re-introduce the non-linear function $f(.)$, provided the function is known.

The retinotopic organization of the early visual system has two implications for these cells. *Firstly*, the weights of the cell depend on its retinotopic position $\boldsymbol{x} = (x_1, x_2)$ and the positions $\boldsymbol{y} = (y_1, y_2)$ of its dendrites. We replace the input I_i by $I(\boldsymbol{y})$ and the weights w_i by $w(\boldsymbol{x} - \boldsymbol{y})$. The *receptive field* $w(\boldsymbol{x} - \boldsymbol{y})$ will typically be zero unless $|\boldsymbol{x} - \boldsymbol{y}|$ is small (this can be mapped by determining if image properties at position $\boldsymbol{y}$ in the image cause the cell to fire). Hence the

neuron can be modelled by:

$$S(\boldsymbol{x}) = \sum_{\boldsymbol{y}} w(\boldsymbol{x} - \boldsymbol{y}) I(\boldsymbol{y}) = \mathbf{w} * I,$$

Secondly, retinotopy implies that there are cells with similar properties (e.g., the same weights $\boldsymbol{w}$) arranged roughly evenly in spatial position (apart from the log-polar transformations [151]). This can be thought of as having "copies" of the same cell at all positions in space. In terms of linear filter theory, see later this section, these sets of cells are *convolving* the image $\boldsymbol{I}$ by a filter $\boldsymbol{w}$.

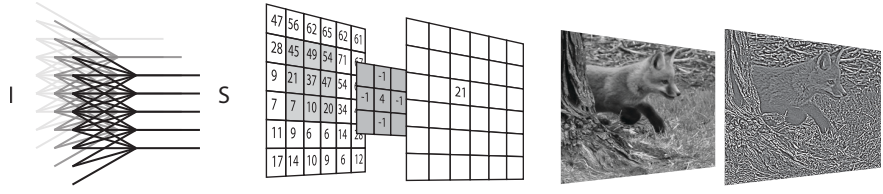

Fig. 13. This figure shows the input-output of a center surround cell (e.g., Laplacian of a Gaussian) in three different ways. First in terms of the inputs and outputs of neurons (left). Second in terms of the digitized input image, the filter, and the digitized output (center). The output at each pixel is given by the product of the filter to the appropriate intensity values in the input image, e.g., $4 \times 37 - 1 \times 49 - 1 \times 47 - 1 \times 10 - 1 \times 21 = 21$. Thirdly, in terms of the input and output images (right).

Center-Surround Cells: Retina and LGN. The receptive fields of the ganglion cells in the retina and cells in the Lateral Geniculate Nucleus (LGN) can be determined by measuring the firing rate of the neurons in terms of its response to different input stimuli $\boldsymbol{I}$ and estimating a model for the response, as discussed in the next section (we refer to [139] for a description of the photo-receptors). The experimental findings are that many simple cells have a characteristic receptive field called *center-surround*. But these findings are done using synthetic stimuli, as we will briefly describe, and their response may be more complex if they are studied using natural stimuli, see section (2.3).

There are two different types: on-center and off-center. The receptive field weights $w(\boldsymbol{x} - \boldsymbol{y})$ are radially symmetric and take the form of a Mexican hat or inverted Mexican hat, for on-center and off-center cells respectively [109], see figure (14)(center right). These cell responses are usually thresholded, e.g., by the sigmoid function, so that they usually only give positive responses. The weights $w(\boldsymbol{x} - \boldsymbol{y})$ can be approximated by the *Laplacian of a Gaussian* (LOG) or by its negative:

$$w_{LOG}(\boldsymbol{x}) = -\{\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}\} G(\boldsymbol{x} : \mathbf{0}, \sigma^2) \qquad (1)$$

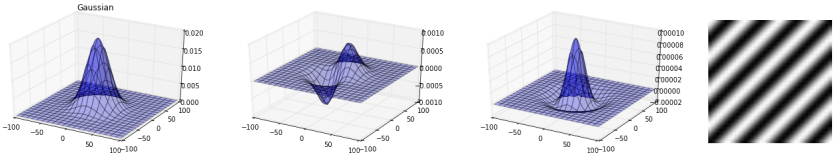

Fig. 14. A Gaussian filter (far left). The first derivative of a Gaussian (left). The laplacian of a Gaussian or Mexican hat (right). A sinusoid (far right).

where $G(\boldsymbol{x} : \mathbf{0}, \sigma^2) = \frac{1}{2\pi\sigma} \exp\{-(x_1^2 + x_2^2)/(2\sigma^2)\}$, see figure (14)(far left).

These cells have two important properties: (I) They are radially symmetric in the sense that $w_{LOG}(.)$ is invariant to rotation, e.g. suppose we express position $\boldsymbol{x}$ in terms of radial components: $x_1 = r\cos\theta, x_2 = r\sin\theta$, then $w_{LOG}(r\cos\theta, r\sin\theta)$ is independent of θ. (II) The receptive field weights $w(.)$ sum up to zero. More precisely,

$$\sum_{\boldsymbol{x}} w_{LOG}(\boldsymbol{x}) = 0.$$

Note that center-surround cells are often modelled as the *differences of two Gaussians* $w_{DOG}(\boldsymbol{x}) = A_1 G(\boldsymbol{x} : \mathbf{0}, \sigma_1^2) - A_2 G(\boldsymbol{x} : \mathbf{0}, \sigma_2^2)$, where σ_1, σ_2 take different values [190]. This gives a similar model, if $|\sigma_1 - \sigma_2|$ and $|A_1 - A_2|$ are small.

The purpose of these center-surround cells is believed to help deal with the large dynamic range of images. Suppose we can express the image locally as $I(\boldsymbol{x}) = C(\boldsymbol{x}) + B$ where $C(\boldsymbol{x})$ is the *contrast*, which describes the local details of the image, and B is the *background*. Then filtering an image by a center-surround cell, whose receptive field sums to 0, removes the background term and preserves part of the contrast. More precisely, using equation (2.1):

$$S(\boldsymbol{x}) = \sum_{\boldsymbol{y}} w_{LOG}(\boldsymbol{x}-\boldsymbol{y}) I(\boldsymbol{y}) = \sum_{\boldsymbol{y}} w_{LOG}(\boldsymbol{x}-\boldsymbol{y})(C(\boldsymbol{y})+B) = \sum_{\boldsymbol{y}} w_{LOG}(\boldsymbol{x}-\boldsymbol{y}) C(\boldsymbol{y}).$$

Receptive fields of this type can also help efficiently encode the information at the retina in order to transmit it efficiently to the visual cortex. This can be studied using information theory and the statistics of natural images to predict properties of receptive fields and how they change in different environments [6]. This theory is beyond the scope of our chapter and we refer to the detailed exposition in [190].

These models of cells in both the retina and the LGN are well studied. Although many of their properties were estimated using synthetic input data it has been shown that in some cases the input image can be estimated from the response of cells in either the retina or the LGN using these types of models [175,24,19]. But other authors [47] argue that the retina is more complex and that, in particular, the neurons may act more as *feature detectors*, see section (2.2), instead of as spatial-temporal filters as described in this section. In particular, [47] describes many finding suggesting that the retina is more com-

plex that the linear filtering model described above. It is known, for example, that if the light levels go down then the receptive field size becomes larger [190].

Note that we are also ignoring the temporal behavior of the cells and a more realistic model models the output as $S(\boldsymbol{x},t) = \sum_{\boldsymbol{y},\tau} w(\boldsymbol{x}-\boldsymbol{y},t-\tau)I(\boldsymbol{y},\tau)$, where $w(\boldsymbol{x}-\boldsymbol{y},t-\tau)$ is a space-time filter. Broadly speaking there are two types of cells with very different temporal characteristics. These are: (i) the M-cells whose receptive fields are spatially large but temporally small (faster) and which project to the dorsal stream, and (ii) the P-cells whose receptive fields are spatially smaller but temporally larger (slower) and which project to the ventral stream. We also do not model the dependence of the cells on the wavelength of the input light. This would require a model $S(\boldsymbol{x}) = \int d\lambda w(\boldsymbol{x}-\boldsymbol{y})w_c(\lambda)I(\boldsymbol{x},\lambda)$, where λ denotes the wavelength and $w_c(\lambda)$ specifies the sensitivity of the cell to color (i.e. there are three types of sensitivity to color $w_{red}(.), w_{green}(.), w_{blue}(.)$). These types of models are described in [190].

Studying tuning by response to sinusoid stimuli. How do we know the receptive field of a neuron? One way is to study its response to a class of stimuli while varying the stimulus parameters (like the perceptual dimensions mentioned in the first section). In particular, we can find how well the neuron is *tuned* to particular stimulus parameters (typically the neuron will prefer a specific value of the parameter and its response will decrease as the parameter changes). This is a classic way to study receptive field properties [64]. In this section, we analyze tuning when the stimuli are sinusoid gratings.

We probe the receptive field of a neuron by stimulating it by a sinusoid grating with intensity $I(\boldsymbol{x}) = A\cos(\boldsymbol{\omega}\cdot\boldsymbol{x}+\rho)+I_0$, where A is the *amplitude*, ρ is the *phase*, $\boldsymbol{\omega}$ is the *frequency* and I_0 is the mean light level. The frequency is vector valued and specifies the orientation of the stimulus, by the unit vector $\hat{\boldsymbol{\omega}} = \boldsymbol{\omega}/|\boldsymbol{\omega}|$, and the period of the oscillation $|\boldsymbol{\omega}|$. The phase ρ shifts the center of the sinusoid, hence if $\rho = 0$ the center occurs at $\boldsymbol{x} = 0$. We can re-express $A\cos(\boldsymbol{\omega}\cdot\boldsymbol{x}+\rho) = A\cos(\boldsymbol{\omega}\cdot(\boldsymbol{x}-\boldsymbol{x}_0))$, where $\boldsymbol{x}_0 = -\rho\boldsymbol{\omega}/|\boldsymbol{\omega}|^2$ is the shift in position.

Then we can analyze the response of the neuron by assuming a functional form for the receptive field. For example, suppose we guess that the neuron is a center-surround cell and its receptive field is a laplacian-of-a-gaussian $w_{LOG}(\boldsymbol{x})$, given by equation (1). Then we can probe the neuron with sinusoid stimuli to determine if it is center-surround and, if so, to determine the parameter σ^2. In order to do this, we calculate the response of laplacian-of-a-Gaussian to a sinusoid function. The laplacian-of-a-Gaussian is a symmetric filter, so its fourier transform is symmetric, and its response to the sinusoid is $A(\cos\rho)\hat{w}_{LOG}(\boldsymbol{\omega})$, where the fourier transform of $w(LOG)(.)$ is $\hat{w}_{LOG}(\boldsymbol{\omega}) = (\boldsymbol{\omega}\cdot\boldsymbol{\omega})\exp\{-(\sigma^2\boldsymbol{\omega}\cdot\boldsymbol{\omega})/2\}$. Hence the predicted response is:

$$\int d\boldsymbol{x} w_{LOG}(\boldsymbol{x})A\cos(\boldsymbol{\omega}\cdot\boldsymbol{x}+\rho) = A(\cos\rho)(\boldsymbol{\omega}\cdot\boldsymbol{\omega})\exp\{-(\sigma^2\boldsymbol{\omega}\cdot\boldsymbol{\omega})/2\}.$$

From this we can deduce three properties: (i) the response is biggest if the center of the sinusoid is aligned to the center of the cell, i.e. $\rho = 0$, falling to zero at

$\rho = \pi/2$, and (ii) the cell responds best to frequencies with $|\boldsymbol{\omega} \cdot \boldsymbol{\omega}| = 2\sigma^{-2}$ (this is obtained by maximizing the response with respect to $|\boldsymbol{\omega}|$), (iii) the cell is insensitive to the orientation of the stimuli. If the real neuron obeys these properties, which we find by stimulating it with sinusoids and measuring its firing rate, then we deduce that it is center-surround and estimate its parameter σ^2. If not, for example if the neuron is sensitive to the orientation of the sinusoid, then we know the cell is not center-surround and needs a different model (see next section). In short, important properties of the receptive field can be obtained by stimulating the cell with sinusoid gratings. This determines the sensitivity of the cell to the orientation θ, the frequency $\boldsymbol{\omega}$, and the phase ρ. As we will show in the next section, simple cells in V1 are *tuned* to the orientation θ and frequency $\boldsymbol{\omega}$ of the local image patch.

Studying the tuning to stimuli parameters gives a way to characterize the receptive field properties of a cell. But it is not sufficient to determine the receptive field uniquely. To do this requires studying the response of the neuron to many types of inputs (an advantaged of studying tuning is that it can be done with comparatively few stimuli). Later, in section (2.3), we discuss how the more advanced ways of estimating receptive fields by regression which require making fewer assumptions about the receptive field.

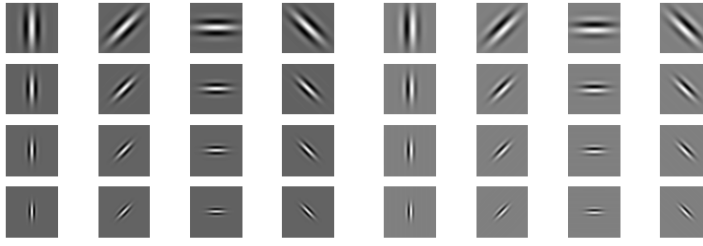

Fig. 15. A family of Gabor receptive fields. The panels show cosine-Gabors (left panel) and sine-Gabors (right panel) at different orientations (rows) and different scales (columns). Observe that the cosine-Gabors have biggest responses at their centers (because $\cos 0 = 1$) while the sine-Gabors have small responses there (because $\sin 0 = 0$).

Visual Cortex area V1: Oriented Receptive Fields. Now we turn to the receptive field properties of *simple cells* in area V1. These cells were first systematically studied by Hubel and Wiesel [65][66] who showed that they were *tuned* to the orientation of edges and size of bars of light. (Later studies used gratings to show tuning to orientation and spatial frequency, which roughly relates to bar size.) Hubel and Wiesel also showed that these cells were spatially organized with hypercolumns and retinotopic organization. Further electrophysiological studies by Roner and Pollen [133] and Jones and Palmer [70] showed

that the receptive field properties of these cells could be approximately modelled by *Gabor filters* [25] which are the product of Gaussians and sinusoids, and can be thought of as performing local fourier transforms, see figure (15). It was also reported that the receptive fields occur in quadrature pairs [133] so that neighboring cells are ninety degrees out of phase (e.g., a cosine Gabor is paired with a sine Gabor). These are illustrated in figure (16). Good fits, however, to the receptive fields can also be obtained using the derivatives of Gaussian filters [181], see figure (14)(center left). Interactive demo (2b) illustrates Gabor functions and their properties.

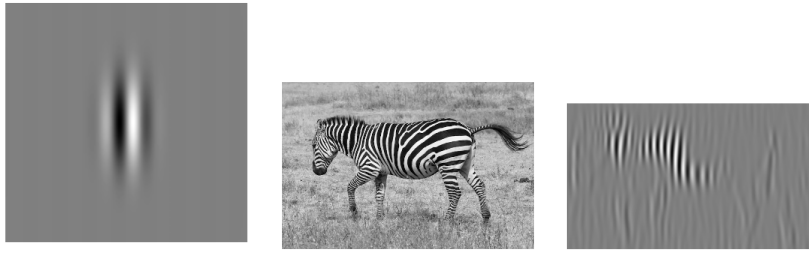

Fig. 16. A Gabor functions aligned to the vertical axis (left). The image of a zebra (center). The response of the vertical Gabor filter on the zebra image (right).

Gabor functions can be expressed as the product of a Gaussian $G(\boldsymbol{x}; \mathbf{0}, \Sigma) = \frac{1}{2\pi|\Sigma|} \exp\{-(1/2)\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x}\}$ (Σ is the covariance of the Gaussian) by a sinusoid $\exp\{i\boldsymbol{\omega} \cdot \boldsymbol{x}\} = \cos \boldsymbol{\omega} \cdot \boldsymbol{x} + i \sin \boldsymbol{\omega} \cdot \boldsymbol{x}$. This gives two basic types of Gabors: (i) cosine-Gabors $G_{\cos}(\boldsymbol{x}) = G(\boldsymbol{x}; \mathbf{0}, \Sigma) \cos \boldsymbol{\omega} \cdot \boldsymbol{x}$ and (ii) sine-Gabors $G_{sin}(\boldsymbol{x}) = G(\boldsymbol{x}; \mathbf{0}, \Sigma) \sin \boldsymbol{\omega} \cdot \boldsymbol{x}$. This form a *quadrature pair*, because sin(.) and cos(.) are ninety degrees out of phase, which will be discussed and used later in this chapter (also phases can vary). The intuition for Gabor filters is they give a good trade-off between *localization* in position and in frequency. The Gaussian has good localization in position, in the sense that its response is very small if $|\boldsymbol{x}| > 2\sigma$. The sinusoid has perfect localization in frequency (due to the orthogonality of sinusoids) but is unable to localize in position (because a sinusoid does not tend to zero for large $\boldsymbol{x}$). More rigorously, Gabor derived the Gabor function by optimizing a criterion that balanced optimality in frequency with optimality in position.

A systematic study in the late 1990's concluded that many simple cells in V1 could be modeled by a family of Gabor filters with specific relationships between the parameters of the gaussian and the sinusoid, Σ and $\boldsymbol{\omega}$ [94]. The experimental data at that time suggested that only a restricted class of Gabor filters were implemented. To understand this restriction, express the frequency of the sinusoid as $\boldsymbol{\omega} = \omega(\cos\theta, \sin\theta)$, where θ specifies the orientation of the sinusoid (and its propagating direction) and ω the magnitude of its frequency.

Then the Σ of the gaussian is proportional to $(1/4)(\cos\theta, \sin\theta)(\cos\theta, \sin\theta)^T + (-\sin\theta, \cos\theta)(-\sin\theta, \cos\theta)^T$ (T denotes vector transform). So the aspect ratio of the gaussian (the ratio between its major and minor axes) is fixed at 4. The sinusoid $\exp(i\boldsymbol{x}\cdot\boldsymbol{\omega})$ has its "propagating direction" along the shorter axis of the Gaussian, so the gaussian smooths more in the direction perpendicular to the propagating direction, by a factor of $1/2 = \sqrt{1/4}$. This family of Gabor filters in illustrated in figure (15).

This family is specified as follows:

$$\psi(\boldsymbol{x};\omega,\theta,K) = \frac{\omega^2}{4\pi K^2}\exp\{-(\omega^2/8K^2)\{4(\boldsymbol{x}\cdot(\cos\theta,\sin\theta))^2 + (\boldsymbol{x}\cdot(-\sin\theta,\cos\theta))^2\} \times \exp\{i\omega\boldsymbol{x}\cdot(\cos\theta,\sin\theta)\}\exp\{(K^2/2)\}\} \quad (2)$$

In this family the orientation is specified by θ, the frequency by ω, and the variance is proportional to K^2. This is normalized so that $\int d\boldsymbol{x}\{\psi(\boldsymbol{x};\omega,\theta,K)\}^2 = 1$. $K \approx \pi$ for a frequency bandwidth of one octave, $K \approx 2.5$ for a frequency bandwidth of 1.5 octaves ("octaves" are the log ratio of the frequency – see [190]). This family can also be scaled to give a form:

$$\psi_a(\boldsymbol{x};\omega,\theta,K) = \frac{1}{a}\psi_a(\boldsymbol{x}/a;\omega,\theta,K)$$

Gabor functions have many interesting properties. They respond well to periodic structure – i.e. certain types of texture – provided the period of the Gabor filter is similar to the period if the texture structure, see figure (16). In addition, oriented sinusoid Gabors act like derivative filters, with their biggest responses on edges with the same orientation, and hence may function as components of edge detectors (see next section). More generally, *filterbanks of Gabor filters* offer local representations of image properties which can used used for many tasks including texture modeling, motion estimation, and binocular stereo (see later this section).

We can also study the tuning of Gabor cells by stimulating them with a family of stimuli of form $A\cos(\boldsymbol{\omega}\cdot\boldsymbol{x}+\rho)$ and varying $\boldsymbol{\omega}$ and ρ (similar to what we did for center surround cells in the previous section). We define $\omega_x = \boldsymbol{\omega}\cdot(\cos\theta,\sin\theta)$ and $\omega_y = \boldsymbol{\omega}\cdot(-\sin\theta,\cos\theta)$ to be the projections of the input sinusoid in the favored direction of the cell (i.e. $\boldsymbol{\omega}$) and in the orthogonal direction (i.e. $\omega_y = 0$ if the input sinusoid aligns perfectly with the orientation of the cell). The responses of the cosine-Gabor G_{cos} and the sine-Gabor G_{sin} are given by:

$$\frac{A}{2}\cos\rho\exp\{-2K^2\omega_y^2/\omega^2\} \times\{\exp\{-(K^2/2\omega^2)(\omega+\omega_x)^2\} + \exp\{-(K^2/2\omega^2)(\omega-\omega_x)^2\}\}\exp\{K^2/2\}, \quad (3)$$

$$\frac{A}{2}\sin\rho\exp\{-2K^2\omega_y^2/\omega^2\} \times\{\exp\{-(K^2/2\omega^2)(\omega+\omega_x)^2\} - \exp\{-(K^2/2\omega^2)(\omega-\omega_x)^2\}\}\exp\{K^2/2\}. \quad (4)$$

These equations show that the cosine-Gabor cell is tuned to $\rho = 0$ and the tuning falls off as $\cos\rho$. The cell also favors sinusoid stimuli which are aligned to it (i.e. $\omega_y = 0$), and whose frequency $\omega_x = \pm\omega$. By contrast, the sine-Gabor prefers stimuli with $\rho = \pi/2$ (naturally, since this converts the stimulus into a sine function which aligns to the sine-Gabor), and has similar tuning to the frequency with $\omega_y = 0$ and $\omega_x = \pm\omega$.

Complex Cells. Complex cells are sensitive to orientation but they are less sensitive than simple cells to the spatial position of the stimuli. A standard theory of the ventral stream, see chapter by Lewis and Poggio, suggests that visual processing proceeds up this stream using receptive fields similar to simple and complex cells, which are increasingly tuned to more complex structures and are less sensitive to the precise positions of the stimuli. Complex cells are the second stage after simple cells, forming a simple-complex cell module which gets repeated up the hierarchy.

We describe here the *energy model* where the complex cell receives input from two simple cells which are ninety degrees out of phase (i.e. cosine-Gabors and sine-Gabors). This is partly motivated by quadrature cells [70] and because, see the following paragraphs, these cells are less sensitive than simple cells to the specific position of the stimuli. Note the word "energy" is based on analogy to physical systems.

More precisely, the energy model of a complex cell gives response:

$$S(\boldsymbol{x}) = \{\psi_{\sin} * I(\boldsymbol{x})\}^2 + \{\psi_{\cos} * I(\boldsymbol{x})\}^2. \tag{5}$$

where $*$ indicates convolution (2.1).

We can study the tuning of complex cells by measuring their response to stimuli $A\cos(\boldsymbol{\omega}\cdot\boldsymbol{x} + \rho)$ and varying $\boldsymbol{\omega}$ and ρ (as before). The findings show that these cells are, like simple cells, also tuned to orientation, frequency, and phase. But their tuning, particularly to phase, is less precise. Hence complex cells are less sensitive to the precise position of the stimuli. The response is given by:

$$\begin{aligned} &\frac{A^2}{4}\exp\{K^2\}\exp\{-4K^2\omega_y^2/\omega^2\} \\ &\{\exp\{-(K^2/\omega^2)(\omega+\omega_x)^2\} + \exp\{-(K^2/\omega^2)(\omega-\omega_x)^2\} \\ &+2\cos 2\rho\exp\{-(K^2/\omega^2)(\omega+\omega_x)^2\}\exp\{-(K^2/\omega^2)(\omega-\omega_x)^2\}\}. \end{aligned} \tag{6}$$

Observe that the dependence on the phase ρ is much small (the dominant term in the second line is independent of ρ).

Complex cells have several possible functions including stereo and motion, as we will show in later sections. They will also respond to oriented edges and can be used for edge detection. In addition, they can represent the local "energy" of textures.

This complex cell model has nice theoretical properties and has been extensively studied and used to construct models (see later in the lecture). But there

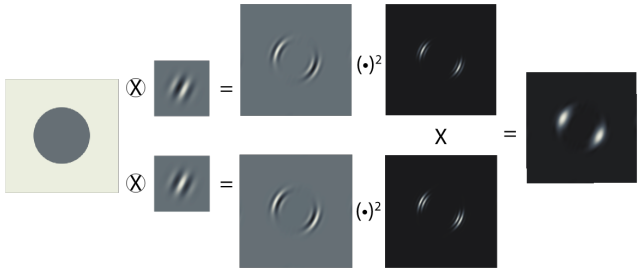

Fig. 17. A complex cell can be modelled as a quadrature pair of Gabor filters. The stimulus is a grey circle on a white background (far left). A quadrature pair of Gabor filters is applied to the stimulus giving the largest responses when the orientation of the Gabors matches the orientation of the edge of the circle. The responses of the Gabors are squared and then summed to yield the final output (far right).

are other models where complex cells are built from simple cells in alternative ways, see the chapter by Lewis and Poggio, but where the complex cells retain their basic property of being tuned to orientation and frequency but being less sensitive to the position of the stimuli. Some researchers question whether complex cells receive input from single cells arguing that the computations could be done by non-linear neurons which exploit the complexity of the dendridic tree [115]. Other researchers argue [113] that there is no sharp dichotomy between simple and complex cells but instead there is an continuum of cells with variable sensitivity to position. Observe that the change between simple and complex cells illustrates a general property of receptive fields in the Ventral stream – namely that receptive fields become more tuned to complex stimuli and less sensitive to the position of the stimulus (e.g., a cell tuned to faces is fairly insensitive to where the face lies in the visual field).

Linear Filtering, Basis Functions, and Fourier Analysis. We now take a step backwards and put these models into the context of the extensive mathematical literature on linear filtering and fourier analysis. This is an advanced section which gives greater understanding but is not required for a basic introduction. Note that in this chapter we will often switch between treating $\boldsymbol{x}$ as a discrete variable and using summation, $\sum_{\boldsymbol{x}}$, or treating it as a continuous variable and using integration, $\int d\boldsymbol{x}$. Similar notation is used in [190].

As discussed in the previous section, simple cell models apply *linear filters* to images and cells at different spatial locations perform *convolution* $*$ by applying the same filter $\boldsymbol{w}$ across the image:

$$S(\boldsymbol{x}) = \boldsymbol{w} * I(\boldsymbol{x}) = \sum_{\boldsymbol{y}} w(\boldsymbol{x} - \boldsymbol{y}) I(\boldsymbol{y}).$$

It is also convenient to approximate this (take the continuum limit) and express this as an integral:

$$S(\boldsymbol{x}) = \int_{\boldsymbol{y}} w(\boldsymbol{x}-\boldsymbol{y})I(\boldsymbol{y})d\boldsymbol{y}.$$

This continuum limit is a good approximation, if the summation $\sum_{\boldsymbol{y}}$ is over a dense set of positions $\boldsymbol{y}$, and enable certain type of analysis (e.g., showing that a center-surround cell model sums, approximately, to zero). Throughout this chapter we will often switch from continuous to discrete filters whenever it simplifies the analysis.

Convolving an image by a linear filter produces an output image $S(\boldsymbol{x})$ whose form depends on the type of filter $\boldsymbol{w}$. For example, if $w(\boldsymbol{x})$ is a Gaussian function $G(\boldsymbol{x};\sigma) = \frac{1}{2\pi\sigma^2}\exp\{-(x_1^2+x_2^2)/(2\sigma^2)\}$ then convolution effectively just smooths the image by taking a linear weighted average. If $\boldsymbol{w}$ is a derivative of the Gaussian in the x_1 direction, $w(\boldsymbol{x}) = \frac{d}{dx_1}G(\boldsymbol{x};\sigma)$, then this filter gives a large response to *edges*, positions $\boldsymbol{y}$ where the intensity $I(\boldsymbol{y})$ changes abruptly, and has small responses in places where the image intensity changes slowly.

We can better understand images, and linear filtering, by using *functional analysis*. This states that an image, or any signal, can be expressed uniquely as a weighted sum of *basis functions*:

$$I(\boldsymbol{x}) = \sum_i \alpha_i b_i(\boldsymbol{x}), \tag{7}$$

where the $b_i(\boldsymbol{x})$ are *basis functions* and the $\{\alpha_i\}$ are *coefficients*. These basis functions are usually chosen to be *orthonormal*, so that $\sum_{\boldsymbol{x}} b_i(\boldsymbol{x})b_j(\boldsymbol{x}) = \delta_{ij}$ ($= 1$ if $i = j$ and $= 0$ if $i \neq j$). If the basis functions are orthogonal then the coefficients α can be obtained by:

$$\alpha_i = \sum_{\boldsymbol{x}} I(\boldsymbol{x})b_i(\boldsymbol{x}). \tag{8}$$

The *principle of superposition* states that we can determine the output S as a weighted combination of the outputs of the basis functions:

$$S(\boldsymbol{x}) = \sum_i \alpha_i S_i(\boldsymbol{x}), \quad \text{where } S_i(\boldsymbol{x}) = \sum_{\boldsymbol{y}} w(\boldsymbol{x}-\boldsymbol{y})b_i(\boldsymbol{y}). \tag{9}$$

This implies that if we know the response $S_i(.)$ to each basis function $b_i(.)$, then we can predict the response to any input. This is an attractive property which, if it holds, enables us to measure the receptive field of a linear neuron, or a thresholded linear neuron, from a limited set of stimuli.

Fourier analysis deals with a special class of basis functions. These are sinusoids, i.e. of form $\sin\omega x, \cos\omega x$. Then the α's are the *fourier transform* of the image. If we restrict ourselves to an image defined on a lattice (i.e. so that x_1, x_2 each take a finite number of values, as on a digital camera) then this is the *discrete fourier transform*. But if we allow x_1, x_2 to take continuous values, then we get the *fourier transform*:

$$I(\boldsymbol{x}) = \frac{1}{2\pi}\int \hat{I}(\boldsymbol{\omega})\exp\{-i\boldsymbol{\omega}\cdot\boldsymbol{x}\}d\boldsymbol{\omega} \tag{10}$$

$$\hat{I}(\boldsymbol{\omega}) = \frac{1}{2\pi}\int I(\boldsymbol{x})\exp\{i\boldsymbol{\omega}\cdot\boldsymbol{x}\}d\boldsymbol{x} \tag{11}$$

Here $\exp\{i\boldsymbol{\omega}\cdot\boldsymbol{x}\} = \cos(\boldsymbol{\omega}\cdot\boldsymbol{x}) + i\sin(\boldsymbol{\omega}\cdot\boldsymbol{x})$. Note that if $I(.)$ is symmetric, $I(\boldsymbol{x}) = I(-\boldsymbol{x})$, then $\hat{I}(\boldsymbol{\omega})$ is also symmetric $\hat{I}(-\boldsymbol{\omega}) = \hat{I}(\boldsymbol{\omega})$. Observe that equations (10,11) correspond to equations (7,8) for special choices of the basis functions (and changing from discrete to continuous $\boldsymbol{x}$).

Fourier analysis is particularly important because it gives us a way to represent non-local structure of images in terms of *frequencies* ω. The high frequencies (large $|\boldsymbol{\omega}|$) represent image patterns which change rapidly while the lower frequencies (small $|\boldsymbol{\omega}|$) represent slowly changing patterns. In particular, if an image pattern is *periodic*, like the stripes on a zebra, then it can be expressed in form:

$$I(\boldsymbol{x}) = \sum_n A_n \cos(2\pi n\boldsymbol{\omega}_0\cdot\boldsymbol{x}),$$

where $\boldsymbol{\omega}_0$ is the basic frequency and n denote integers. Then the Fourier transform is only non-zero at integer multiples of the basic frequency $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. Hence periodic image patterns, such as *textures*, have very simple descriptions in Fourier space.

If we blur the image, by convolving with a Gaussian $G(\boldsymbol{x};\sigma)$, to obtain $G * I(\boldsymbol{x})$, then the high frequencies of the image $\boldsymbol{I}$ will be smoothed out. This follows from the *convolution theorem* which states that fourier transform of $G * I(\boldsymbol{x})$ equals the product of the fourier transforms of G and $\boldsymbol{I}$, and the fact that the fourier transform of a Gaussian is also a Gaussian $\exp\{-|\boldsymbol{\omega}|^2(\sigma^2/2)\}$. Hence we can express the convolved image as a weighted combination of sinusoids where the high frequency weights are decreased by $\exp\{-|\boldsymbol{\omega}|^2(\sigma^2/2)\}$:

$$\boldsymbol{I}(\boldsymbol{x}) = \frac{1}{2\pi}\int \hat{I}(\boldsymbol{\omega})\exp\{-i\boldsymbol{\omega}\cdot\boldsymbol{x}\}\exp\{-|\boldsymbol{\omega}|^2(\sigma^2/2)\}d\boldsymbol{\omega}.$$

Hence if we increase the amount of blurring, by increasing the variance σ^2 of the Gaussian, we will make the coefficients of the high frequencies increasingly small. Blurring the image can be obtained by defocusing your eyes so that the image is seen out of focus. The receptive fields of cells occurs at a range of different scales, corresponding to convolving with Gaussians of different variances.

The superposition principle, combined with the use of basis functions, shows that we can determine the receptive fields of linear neurons by stimulating them with sinusoids. Sinusoids can be used as basis functions and superposition can be used to predict the response to stimuli which have not been seen yet (i.e. as superpositions of those stimuli to which the response is known). This, however, is rarely done. We will return to studying ways to determine receptive fields from experiments in section (2.3).

2.2 Sparsity, Matched Filters, and Natural Images

This section considers receptive field models from different perspectives. This includes the use of *sparsity* to suggest receptive field properties based on the statistics of natural images and also the idea of *matched filters* which revert to an older idea of receptive fields as feature detectors [99]. Sparsity was proposed by Barlow [7] as a general principle for modeling the brain based on the observation that typically only a small number of neurons are active. It was developed as a way to predict receptive field properties by Olshausen and Field [127]. It is natural to ask whether the receptive fields of cells encode basis functions which somehow capture the typical structure of images and represent it in a form which is suitable for later processing.

Our starting point is the idea that images, and particularly local regions of images. can be represented as a linear combination of basis functions $I(\boldsymbol{x}) = \sum_i \alpha_i \boldsymbol{b}_i(\boldsymbol{x})$, see equation (7).

Over-Complete Bases and Sparsity. This section introduces the idea of *over-complete* basis functions and *sparsity*. To motivate this idea, consider an image which consists partly of regions where the intensity varies spatially smoothly and others where the intensity is more jagged and consists of a number of bright spots, or *impulses*. The smoothly varying regions of the image can be represented by fourier analysis efficiently, in the sense that we can approximate the intensity by only a small number of weighted sinusoids (in other words, the fourier transform of the image is peaked at a limited number of frequencies). By contrast, the impulses are not well described by fourier analysis because the fourier transform is not zero for all frequencies (the fourier transform of an impulse at $\boldsymbol{x}_0$ is $\exp\{i\boldsymbol{\omega}\cdot\boldsymbol{x}_0\}$, so the amplitude spectrum is constant at all frequencies). Instead it would be better to represent the spikes in terms of a basis of impulse functions, but this representation would be very inefficient for the smoothly varying parts of the image. In short, different types of basis functions are suitable for different regions of the image. This suggests a strategy where we seek a representation in terms of an over-complete set of basis functions, in this case sinusoids and impulse functions, and a criterion which selects an efficient representation so that only a small number of basis functions are activated for each image. This requirement is called *sparsity*.

More formally, we represent an image, or local image region, by:

$$I(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i b_i(\boldsymbol{x}),$$

where the $\{b_i\}$ are the basis functions (which are the same for all images, and could include sinusoids and impulse functions) and the $\{\alpha_i\}$ are the coefficients of the bases (which depend on the image). The number N of bases is much bigger than the dimension of the image, and hence the bases are *over-complete*. This differs from fourier analysis where the data (e.g., an image) is expressed in terms of a set of basis functions which are mutually orthogonal, which enables

the coefficients α to for each image to be estimated by $\alpha_i = \sum_{\boldsymbol{x}} b_i(\boldsymbol{x}) \cdot I(\boldsymbol{x})$. Over-completeness implies that there are many ways to represent the image in terms of these basis functions (by different choices of the α's) and we need an additional criterion to select the α's. The *sparsity* criterion proposes that we favor representations which make $\sum_{i=1}^{N} |\alpha_i|$ small, which penalize the weights of the basis functions and encourages most coefficients to be 0.

More precisely, we represent an image $\boldsymbol{I}$ by the approximation $\sum_{i=1}^{N} \hat{\alpha}_i \boldsymbol{b}_i$, where the $\{\hat{\alpha}_i\}$ are chosen to minimize the function:

$$E(\alpha) = \sum_{\boldsymbol{x}} (I(\boldsymbol{x}) - \sum_{i=1}^{N} \alpha_i b_i(\boldsymbol{x}))^2 + \lambda \sum_{i=1}^{N} |\alpha_i|. \tag{12}$$

The first penalizes the error of the approximation and the second term, whose strength is weighted by a parameter λ, penalizes the coefficients $\{\alpha_i\}$. The solution $\hat{\alpha} = \arg\min_{\alpha} E(\alpha)$ can not not be specified in closed form (unlike the case for orthogonal basis function), but $E(\alpha)$ is a *convex* function of α and efficient algorithms exist for minimizing it to estimate $\hat{\alpha}$. The results of these algorithms can, for example, decompose an image into a sum of sinusoids and a sum of impulse functions.

These ideas give an alternative way to think about the receptive fields of cells in V1. Firstly, observe that V1 has far more cells than the retina or the LGN and so it is has enough neural machinery to implement over-complete bases. Secondly, over-complete bases can be designed for specific image structures of interest (e.g., impulse functions or edges) which enables us to start interpreting the image instead of simply representing it. Thirdly, it relates to the observation that cells in V1 fire *sparsely*, which suggests [7] that they are tuned to specific stimuli and may relate to metabolic processes(firing a neuron takes energy which needs to be replenished). Hence the idea that the visual cortex seeks to obtain sparse, and hence presumably more easily interpretable representations, has intuitive appeal.

How does this discussion of over-completeness and sparsity relate to our previous description of V1 cells in terms of Gabor filters? Gabor filters have some of the properties that this approach requires. Families of Gabor filters are built by taking a basic functions and performing transformations on it which give an over-complete basis. Hence they do not specify a unique representation of an image (i.e. any image can be represented many different ways in terms of Gabor functions). These issues, and the relations of Gabors to wavelets, are discussed in more detail in [94].

Sparsity and Natural Images. Sparsity can also be used to derive the properties of receptive fields of cell in V1 if we assume that these cells are designed to be able to represent properties of natural images [127], see figure (18)(Left). Hence instead of hypothesizing models of receptive fields (e.g., Gabor filters) we can try to predict these receptive fields from studying images. These predictions do give some justification for Gabor functions but they also suggest other receptive field models which have also been experimentally observed.

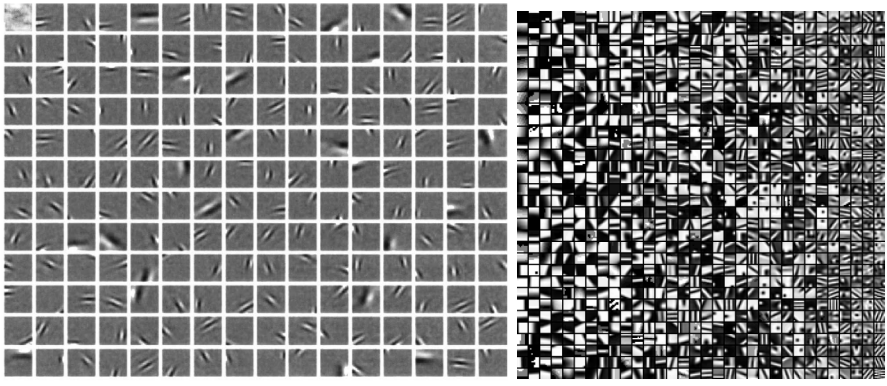

Fig. 18. Left: The receptive fields learnt using sparsity [127]. Figure reprinted with permission from [127]. Right: receptive fields learnt by matched filters, e.g., see [128].

This requires learning the basis functions $\{\boldsymbol{b}_i\}$ from a set of natural images $\{\boldsymbol{I}^\mu : \mu \in \Lambda\}$. This can be found be extending equation (12) to obtain a criteria $E(b, \alpha)$ for fitting basis functions b and coefficients α to the set of images:

$$E(b, \alpha) = \sum_{\mu \in \Lambda} (I^\mu(\boldsymbol{x}) - \sum_{i=1}^{N} \alpha_i^\mu b_i(\boldsymbol{x}))^2 + \lambda \sum_{\mu \in \Lambda} \sum_{i=1}^{N} |\alpha_i|.$$

We estimate the basis functions $\hat{b}$ and the coefficients $\hat{\alpha}$ by minimizing $E(b, \alpha)$ to obtain:

$$(\hat{b}, \hat{\alpha}) = \arg\min_{(b,\alpha)} E(b, \alpha).$$

Note that the basis functions are the same for all images but the coefficients vary for each image (hence they are indexed by the image μ as well as the basis coefficient i). This minimization is non-convex but there are efficient algorithms to perform it.

This criterion has been applied to natural images (where the $\boldsymbol{I}$ represent small image regions) and the resulting basis functions, see figure (18)(left), include filters which look like Gabor functions but they also include other types of filters which are also observed in experiments [127].

We note that there are other methods for predicting receptive field properties from natural images using a similar image model, $I(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i b_i(\boldsymbol{x})$, but imposing different assumptions on the form of the bases. In particular, independent component analysis (ICA) gives similar receptive field models [166]. Hyvarinen [67] explains this by showing that both types of models – L1 sparsity and ICA – both encourage the α_i to be strongly peaked at 0, but can occasionally have large non-zero values.

What happens if we remove the sparsity requirement and instead find the basis functions that minimize $\sum_{\mu \in \Lambda} (I^\mu(\boldsymbol{x}) - \sum_{i=1}^{N} \alpha_i^\mu b_i(\boldsymbol{x}))^2$? The basis functions will be the eigenvectors of the correlation matrix of the images and can

be found by principal component analysis (PCA). Code for performing PCA is supplied in interactive demo (2c). It can be shown that the principal components of images will typically be sinusoids (provided the images are sufficiently representative of natural images). We return to this issue in section (2.3) when we describe unsupervised ways to learn receptive fields of neurons.

Matched Filter Interpretation. This section gives an alternative way, and earlier, model for receptive fields. The idea is the cells are feature detectors [99] which can be modelled by a set of matched filters, so that the cells indicated the type of image patch which is present. This can be thought of as an extreme form of sparsity, because any image patch can be represented by a single matched filter (instead of a linear combination of them). Examples of matched filters are shown in figure (18)(right). We now describe the details of this approach.

Suppose we have a filter $\boldsymbol{W}$ and an input image patch $\boldsymbol{I}_p$. We want to find the best fit of the filter to the image by allowing us to transform the filter by $\boldsymbol{W} \mapsto a\boldsymbol{W} + b\boldsymbol{e}$, where $\boldsymbol{e} = (1/\sqrt{N})(1, ..., 1)$. This corresponds to scaling the filter by a and adding a constant vector b. If $\boldsymbol{W}$ is a derivative filter then, by definition, $\boldsymbol{W} \cdot \boldsymbol{e} = 0$. We normalize $\boldsymbol{W}$ and $\boldsymbol{e}$ so that $\boldsymbol{W} \cdot \boldsymbol{W} = \boldsymbol{e} \cdot \boldsymbol{e} = 1$.

The goal is to find the best scaling/contrast a and background b to minimize the match:

$$E(a, b) = |\boldsymbol{I}_p - a\boldsymbol{W} - b\boldsymbol{e}|^2.$$

The solution $\hat{a}, \hat{b}$ are given by (take derivatives of E with respect to a and b, recalling that $\boldsymbol{W}$ and $\boldsymbol{e}$ are normalized):

$$\hat{a} = \boldsymbol{W} \cdot I_p, \quad \hat{b} = \boldsymbol{e} \cdot \boldsymbol{I}_p.$$

In this interpretation, the filter response is just the best estimate of the contrast a. The estimate of the background b is just the mean value of the image. Finally, the energy $E(\hat{a}, \hat{b})$ is a measure of how well the filter "matches" the input image. Receptive fields learnt by matched filters are shown in figure (18)(right).

The idea of a matched filter leads naturally to the idea of having a "dictionary" of filters $\{\boldsymbol{W}^\mu : \mu \in \Lambda\}$, where different filters $\boldsymbol{W}^\mu$ are tuned to different types of image patches. In other words, the input image patch is encoded by the filter that best matches it. The dictionary of matched filters could be implemented by a set of cells (e.g., orientation columns). In this interpretation, the magnitude of the dot product $\boldsymbol{W} \cdot \boldsymbol{I}$ is less important than deciding which filter best matches the input $\boldsymbol{I}_p$. Matched filters can be thought of an extreme case of sparsity. In the previous sections, an image was represented by a linear combination of basis functions whose weights were penalizes by the *L1-norm*, $\sum_i |\alpha_i|$. By comparison, matched filters represent an image by a single basis function. This gives an ever sparser representation of the image, but at the possible cost of a much larger image dictionary. Matched filters can be thought of as *feature detectors* because they respond only to very specific inputs.

2.3 Learning

We cover two topics in this section: (I) how can these receptive fields be learnt using biologically plausible mechanisms? (II) How we estimate receptive fields from experimental data.

Unsupervised learning of the receptive fields. Where do the receptive fields of simple cells come from? The most plausible explanation is that they are learnt in an unsupervised self-organizing manner but it is debatable whether this happens before birth, due to spontaneous activity in the brain, or in response to natural stimuli observed after birth. The methods in this section could work either way, since the main property they rely on is that images are shift-invariant. This section is based on computational studies performed in the 1980's [101][100],[187], see [190] for other references. These studies are based on modifications of the Hebb learning rule which has some experimental support. Interactive demo (2c) illustrates principal component analysis and Oja's rule [126].

The basic findings are that center-surround, orientation selective, quadrature pairs, and disparity sensitive cells (precursors to cells which can estimate depth from binocular stereo) could all be obtained by variants of the same learning rule.

We first describe a simple unsupervised learning model of this type for a single cells [126]. The output $S(t)$ of the cell is a function of time t and is a weighted sum of the inputs $I_i(t)$, where the weights $\omega_i(t)$ are functions of time are updated by Oja's rule [126]:

$$S(t) = \sum_j w_j(t) I_j(t),$$

$$\frac{dw_i(t)}{dt} = S(t)\{I_i(t) - S(t)w_i(t)\}. \tag{13}$$

The first term of this update is the classic Hebb's term which increases the strength of a weight w_i if its input $I_i(t)$ is positively correlated with the output $S(t)$ (i.e. $< S(t)I_i(t) >> 0$), while the second term decreases the value of all weights by an amount proportional to their strength.

This can be expressed as a single update equation (substituting for $S(t)$):

$$\frac{dw_i(t)}{dt} = \sum_j w_j I_i(t) I_j(t) - \sum_{jk} w_i w_j w_k I_j(t) I_k(t). \tag{14}$$

Next we make the key assumption that the weights w_i change at a slower rate than the input images. This enables us to replace the terms $I_i(t)I_j(t)$ in equation (14) by their expectation $K_{ij} =< I_i(t)I_j(t) >$ which is the correlation function of the input. This gives an update rule:

$$\frac{dw_i(t)}{dt} = \sum_j w_j K_{ij} - \sum_{jk} w_i w_j w_k K_{jk}. \tag{15}$$

The fixed points of this equation, the values of w such that $\frac{dw_i(t)}{dt} = 0$, can be shown to be eigenvectors of the correlation function K_{ij}. A slight modification of this update rule [187] gives a rule that is guaranteed to converge to the global minimum of the cost function:

$$E(\boldsymbol{w}) = -(1/2)\sum_{i,j} K_{ij} w_i w_j + (k/4)(\sum_i w_i^2)^2$$

It can be shown that this global minimum corresponds to the biggest eigenvalue of K_{ij}. It follows that if the correlation function K_{ij} decreases spatially, then the biggest eigenvalue is at frequency 0 so the cell is not tuned to any frequency. But if the correlation function has the shape of a Mexican hat, then the biggest eigenvalue has a non-zero frequency which implies that the cell is orientated. In fact, there are a set of eigenvalues with the same frequency modulus, each pointing in a different direction. Small change to the input, or noise in the simulation, will determine which orientation is selected which is like symmetry breaking in physics [187]. The correlation function of natural images does decrease spatially, but Linsker showed that correlation functions similar to the Mexican hat arise if this learning procedure is applied to a sequence of layers [101][100].

This analysis yields receptive fields which are sinusoids, and hence have no spatial fall-off which is unrealistic. But receptive fields of neurons are limited by the geometrical positions of the dendrites. If these constraints are included then the algorithms converge to receptive fields which are similar to Gabor functions. These methods can be expanded to other other properties of cells in the visual system. For example, quadrature pairs can be learnt by introducing inhibition between neighboring output neurons [187].

How to empirically estimate receptive field models by regression. This section describes how to estimate the receptive field properties of cells from electrical recordings of neurons by estimating the best model using *regression*. This is a standard method in statistics and machine learning. In requires making very few assumptions about the form of the receptive field (i.e. no need to assume laplacian-of-gasussian, or Gabor, or sum of squared Gabors).

We first recall, as discussed earlier, that the receptive field properties of neurons are traditionally found in simpler ways. For example, by choosing a set of basic stimuli, such as oriented bars, and then observe how the response of the cell changes as we move or rotate the bars. This approach can be extended by probing the receptive field response to different *perceptual dimensions*, such as position, orientation, color, contrast, motion, and whether it is sensitive to stimuli from one eye or both. This gives a classification of the type of the receptive field but does not specify its receptive field weights $\boldsymbol{w}$ unless strong assumptions are made (e.g., that the receptive field is a Gabor function).

The regression method described in this section makes fewer assumptions about the forms of the receptive field but it does require more data. More formally, regression assumes that we have a stimulus dataset of $\mathcal{S} = \{(S^\mu, \boldsymbol{I}^\mu) : \mu = 1, ..., N\}$ of inputs $\boldsymbol{I}^\mu$ and outputs S^μ (e.g., the firing rates). We assume a

model $S = g(\boldsymbol{I} : \boldsymbol{w})$ where $\boldsymbol{w}$ specifies the parameters of the model and $g(.)$ is a non-linear function. For example, we could choose the non-linear function to be a sigmoid $g(\boldsymbol{I} : \boldsymbol{w}) = \sigma(\boldsymbol{w} \cdot \boldsymbol{I})$, where $\sigma(.)$ is a sigmoid function. Alternatively $g(.)$ could specify a Gabor filter where $\boldsymbol{w}$ specifies the parameters of the Gabor (e.g., frequency and scale).

Regression requires minimizing a cost function such as:

$$F(\boldsymbol{w}) = \frac{1}{|\mathcal{S}|} \sum_{\mu \in \mathcal{S}}^{N} E(S^{\mu} - g(I^{\mu}; \boldsymbol{w}))$$

where $E(.)$ is a penalty function such as $(S^{\mu} - g(I^{\mu}; T))^2$.

This minimization can be done by standard computer packages. It outputs an estimate of the model parameters $\hat{\boldsymbol{w}}$ and an error measure $F(\hat{\boldsymbol{w}}) = \frac{1}{|\mathcal{S}|} \sum_{\mu \in \mathcal{S}} E(S^{\mu} - g(I^{\mu}; \hat{\boldsymbol{w}}))$. Hence it not only outputs an estimate of the receptive field but it also produces an error measure, which can yield the variance of the output.

In practice, there are several complications. It is unrealistic to show the neuron all possible stimuli because there are so many possible image stimuli. Hence researchers have to choose a restricted set of stimuli. If neurons are linear, or a non-linear function of a linear filter, then this should not matter because we can exploit the superposition principle and estimate the receptive field from a limited number of stimuli. But in reality, linearity is only an approximation, and in practice the choice of stimuli can matter considerably. One concern is that the stimulus set does not contain the types of stimuli that the neuron is most sensitive to, in which case regression will output unreliable estimates. Also, if the linear assumption is only partially correct then there is no guarantee that the receptive field learnt on one set of stimuli will predict the behavior well on another set of stimuli.

The complexities are illustrated by recent work by Talebi and Baker [158]. They describe in detail how receptive fields of neurons are estimated by regression techniques and also addresses the issue of how the results of regression differ depending on the stimulus set. Their findings (for neurons in the cortex of anesthetized cats) show that estimates of the receptive fields of neurons can depend heavily on the set of stimuli. They estimate receptive fields using three different stimulus sets: (i) white noise (WN), (ii) oriented bars (B), and (iii) natural images (NI). This gives three estimates for the receptive fields $\boldsymbol{w}_{WN}, \boldsymbol{w}_B, \boldsymbol{w}_{NI}$ obtained using the three stimulus sets $\mathcal{S}_{WN}, \mathcal{S}_B, \mathcal{S}_{NI}$. For each dataset, they can compute the prediction errors F_{WN}, F_B, F_{NI} which are the errors for that dataset, e.g., $F_{WN}(\hat{\boldsymbol{w}_{WN}}) = \frac{1}{|\mathcal{S}_{WN}|} \sum_{\mu \in \mathcal{S}_{WN}} E(S^{\mu} - g(I^{\mu}; \hat{\boldsymbol{w}_{WN}}))$. These quantities show how well the models can fit each stimulus set. But, more interestingly, they can study how well the estimated receptive field from one stimulus set can predict the other datasets. This involves computing quantities such as $F_{WN}(\hat{\boldsymbol{w}_B})$, $F_{WN}(\hat{\boldsymbol{w}_{NI}})$, $F_B(\hat{\boldsymbol{w}_{WN}})$, $F_{WN}(\hat{\boldsymbol{w}_{NI}})$, $F_{NI}(\hat{\boldsymbol{w}_{WN}})$, $F_{WN}(\hat{\boldsymbol{w}_B})$. Their results show, perhaps not surprisingly, that the receptive fields estimated on the natural image stimulus set were much better at predicting the responses on the other two stimulus sets.

2.4 Local Models for Binocular Stereo and Motion

This section shows how these linear filter models of receptive fields can be used to perform local estimates of binocular stereo and motion. These involve having filterbanks, or populations of filters, which are tuned to different properties of the stimuli so that estimates of depth and motion can be extracted from the population [190].

Stereo Disparity Models. Recall that we introduced binocular stereo in section (1.2). Depth is estimated by triangulation provided we can solve the *correspondence problem* by finding which points in the left and right eyes correspond to the same point in three-dimensional space. This reduces to estimating the displacement, or *disparity*, between the images in the left and right eyes. This section introduces the disparity energy model which estimates disparity based on local properties of the image. In section (4) we will discuss how non-local context can be used to improve disparity estimation.

The disparity energy model is formulated using Gabor filters and has some claim to biological plausibility. We follow the presentation in Qian [135] which is based on experimental findings on the receptive field property of cells by Ohzawa *et al.* [125]. The model assumes that we have a large set of cells, receiving input from both images, and which are tuned to different image frequencies and spatial phases. The disparity of the image can be computed from the response of these filters.

We give the presentation in one-dimension for simplicity. This is allowed because of the epipolar line constraint, see figure (9). It assumes that the cell receives input from both left and right eyes with receptive fields $f_l(x) = \exp\{-x^2/(2\sigma^2)\}\cos(\omega x + \rho_l)$ and $f_r(x) = \exp\{-x^2/(2\sigma^2)\}\cos(\omega x + \rho_r)$. In other words, they are Gabors where the Gaussian has variance σ^2, tuned to frequency ω and with phases ρ_l, ρ_r. The linear response is:

$$r = \int dx\{f_l(x)I_l(x) + f_r(x)I_r(x)\}. \tag{16}$$

This filter is tuned to spatial frequency ω. Recall, from earlier sections, that we can express the image by a Fourier expansion. The filter is most sensitive to the image component at this frequency. Hence we can represent the image (approximately) by $I(\boldsymbol{x}) = \rho\cos(\omega x + \theta)$.

Now suppose that the right image is a displaced version of the left image $I_r(x) = I_l(x + D(x))$, where $D(x)$ is the disparity. We assume that the disparity varies slowly so that we can approximate it locally as a constant D (over the size of the Gaussian, 2σ), see figure (19)(left). Then we make another approximation by ignoring the Gaussian to calculate r (this is similar to ignoring the spatial fall-off when we studied unsupervised learning algorithms).

This yields a response:

$$r_1 = \rho\{\cos(\theta - \rho_l) + \cos(\theta - \rho_r - \omega D)\}. \tag{17}$$

which can be re-expressed (using trigonometry identities) by:

$$r_1 = 2\rho\cos(\theta - \frac{\rho_l + \rho_r}{2} - \frac{\omega D}{2})\cos(\frac{\rho_l - \rho_r}{2} - \omega\frac{D}{2}). \tag{18}$$

Hence the response of the cell depends on the disparity. In particular, when $\rho_l - \rho_r = \omega D$ then the second cosine takes its biggest value of 1. But the cell's response also depends on image properties, i.e., the image phase θ which is an argument of the first cosine. This means it is unable to detect disparity by itself. But disparity can be estimated from a population of quadrature cells of this type tuned to different frequencies.

To see this, suppose that we consider quadrature pairs of the two cells tuned to the same ω. Where one cell has phases ρ_l, ρ_r and the other has phases ρ'_l, ρ'_r, where $(\rho_l - \rho_r) = (\rho'_l - \rho'_r)$ and $\rho'_l + \rho'_r = \rho_l + \rho_r + \frac{\pi}{2}$. Then the second cell has response $r_2 = 2\rho\cos(\theta - \frac{\rho_l+\rho_r}{2} - \frac{\omega D}{2})\cos(\frac{\rho_l-\rho_r}{2} - \omega\frac{D}{2}) = 2\rho\sin(\theta - \frac{\rho_l+\rho_r}{2} - \frac{\omega D}{2})\cos(\frac{\rho_l-\rho_r}{2}$. Hence if we square and add the responses of the two cells we obtain:

$$r_1^2 + r_2^2 = \cos^2(\frac{\rho_l - \rho_r}{2} - \omega\frac{D}{2}). \tag{19}$$

This response depends only on the disparity D and the image frequency ω. It takes largest values when $\rho_l - \rho_r = \omega D$. Hence we can estimate D from a population of quadrature cells tuned to different phases ρ_l, ρ_r and frequencies ω, see figure (19).

A neural network for estimating D consists of two steps (there are many variants of this approach). In step (I) we define a set of disparity cells tuned to disparities $\{D_i : i = 1, ..., N\}$. The disparity cell tuned to disparity D_i receives input $\cos^2(\frac{\rho_l-\rho_r}{2} - \omega\frac{D_i}{2})$ from each quadrature pair (ρ_l, ρ_r, ω) and sums these inputs together to compute a vote $v(D_i)$:

$$v(D_i) = \sum_{\rho_l,\rho_r,\omega} \cos^2(\frac{\rho_l - \rho_r}{2} - \omega\frac{D_i}{2}). \tag{20}$$

Step (II) uses a winner-take-all network to compute the disparity with the biggest vote by solving $\hat{D} = \arg\max_{i=1,...,N} v(D_i)$, so that $v(\hat{D}) \geq v(D_i)$ for $i = 1, ..., N$. There are many varieties of winner-take-all networks, see [107].

This theory of binocular stereo gives an example where the information of interest, in this case the disparity, is represented by the activity of a population of neurons. We will give another example for motion in the next section. There is plenty of evidence that the brain represents information by neural populations [42],[112]. There has also been much theoretical studies of how populations of neurons could encode knowledge and perform computations [134,106].

Motion Measurement: Spatio-Temporal Filters. We now discuss how related models can be applied to estimate motion for sequences of images. Spatio-temporal filters are biologically plausible ways to measure motion which agree with properties of cells in the visual cortex. The standard model suggests two

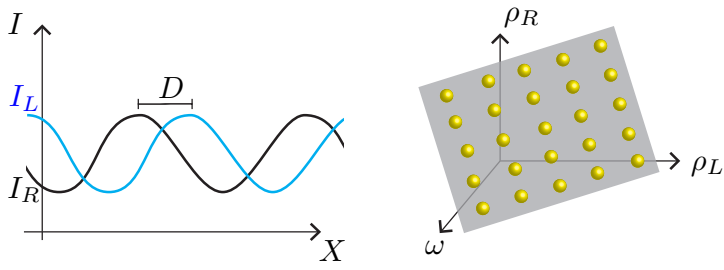

Fig. 19. Left Panel: The disparity D between the images in the two eyes corresponds to a change of phase if we approximate the intensities by sinusoids, see text. Right Panel: The local disparity D is encoded by the feature response of cells tuned to frequencies which obey $\rho_l - \rho_r = \omega D$.

classes of cells where the first are spatio-temporal filters which are sensitive to the directions of motion while the second combine outputs of these filters to estimate the motion itself [3],[57], [150].

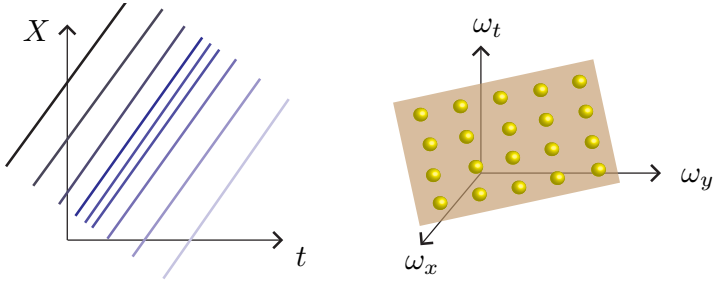

Fig. 20. Left Panel: This figure shows the space-time illustration of a signal traveling with constant velocity $I(X,t) = F(X - tv)$. This means that the intensity $I(X,t)$ is constant on the lines $X - tv = constant$. Right Panel: A stimuli moving with velocity $\boldsymbol{v}$ will activate spatial-temporal filters $\boldsymbol{\omega}, \omega_t$ which lie on the plane $\boldsymbol{v} \cdot \boldsymbol{\omega} + \omega_t = 0$. Hence the velocity can be estimated from the population of activity of the filters.

Measuring the motion velocity assumes that locally the intensity can be modeled as a linear translating pattern, see figure (20)(left panel):

$$I(\boldsymbol{x}, t) = F(\boldsymbol{x} - \boldsymbol{v}t). \tag{21}$$

Differentiating with respect to $\boldsymbol{x}$ and t (using $\boldsymbol{\nabla} I = \boldsymbol{\nabla} F$ and $\frac{\partial I}{\partial t} = -\boldsymbol{v} \cdot \boldsymbol{\nabla} F$), gives the *optical flow equation*:

$$\boldsymbol{v} \cdot \boldsymbol{\nabla} I + \frac{\partial I}{\partial t} = 0. \tag{22}$$

This enables us to estimate one component of the motion $\boldsymbol{v}$ but suffers from the aperture problem discussed in section (1). One way to resolve this ambiguity is by applying a population of filters $\{G^\mu(\boldsymbol{x},t) : \mu = 1,...,M\}$ indexed by μ (for example, the $G^\mu(.)$ could be Gaussian filters). These filters introduce local context:

$$G^\mu * I(\boldsymbol{x},t) = \int G^\mu(\boldsymbol{x}-\boldsymbol{y},t-s)I(\boldsymbol{y},s)dsd\boldsymbol{y}. \tag{23}$$

Hence each filter gives a constraint on the velocity,

$$\boldsymbol{v}\cdot\boldsymbol{\nabla}G^\mu * I + \frac{\partial G^\mu * I}{\partial t} = 0. \tag{24}$$

We can then get an estimate of the velocity $\boldsymbol{v}$ by minimizing the cost function:

$$E(\boldsymbol{v}) = \sum_{\mu=1}^{M}(\boldsymbol{v}\cdot\boldsymbol{\nabla}G^\mu * I + \frac{\partial G^\mu * I}{\partial t})^2.$$

This minimization can be done using a similar neural network to that used for estimating disparity for stereo in the previous section. We have a set of cells tuned to different velocities $\{\boldsymbol{v}_i : i = 1,...,N\}$. The cell tuned to velocity $\boldsymbol{v}_i$ receives input $(\boldsymbol{v}\cdot\boldsymbol{\nabla}G^\mu * I + \frac{\partial G^\mu * I}{\partial t})^2$ from each filter μ and sums the responses to obtain $E(\boldsymbol{v}_i)$. Then we use a variant of winner-take-all to compute $\hat{\boldsymbol{v}} = \arg\min_{i=1,...,N} E(\boldsymbol{v}_i)$.

This approach assumes that there is enough local information to resolve the motion ambiguity which may not be the case. For example, for the stimuli in figure (7) we can only locally estimate one component of the motion because of the aperture problem. To resolve this ambiguity we need to use more spatial or temporal context as described in section (4.4).

An alternative way to analyze this problem is by applying fourier analysis to equation (21).

$$\hat{I}(\boldsymbol{\omega},\omega_t) = \frac{1}{2\pi}\int\int\int \exp\{i(\boldsymbol{\omega}\cdot\boldsymbol{x}+\omega_t t)\}I(\boldsymbol{x},t)d\boldsymbol{x}dt$$

$$\hat{I}(\boldsymbol{\omega},\omega_t) = \frac{1}{2\pi}\int\int\int \exp\{i(\boldsymbol{v}\cdot\boldsymbol{x}+\omega_t)t\}\exp\{i\boldsymbol{\omega}\cdot(\boldsymbol{x}-\boldsymbol{v}t)\}F(\boldsymbol{x}-\boldsymbol{v}t)d\boldsymbol{x}dt$$

$$\hat{I}(\boldsymbol{\omega},\omega_t) = \frac{1}{2\pi}\int \exp\{i(\boldsymbol{v}\cdot\boldsymbol{\omega}+\omega_t)t\}dt\int\int \exp\{i\boldsymbol{\omega}\cdot\overline{\boldsymbol{x}}\}F(\overline{\boldsymbol{x}})d\overline{\boldsymbol{x}}$$

$$\hat{I}(\boldsymbol{\omega},\omega_t) = \delta(\boldsymbol{v}\cdot\boldsymbol{\omega}+\omega_t)\hat{F}(\boldsymbol{\omega})$$

Where $\overline{\boldsymbol{x}} = \boldsymbol{x}-\boldsymbol{v}t$ is a change of variables in the integral.

This shows that if we have filters $\exp\{i(\boldsymbol{x}\boldsymbol{\omega}+\omega_t t)\}$ tuned to spatial-temporal frequencies $\boldsymbol{\omega},\omega_t$ then the only filters which respond are those whose frequencies obey the equation $\boldsymbol{v}\cdot\boldsymbol{\omega}+\omega_t = 0$ and hence lie on a plane in frequency space.

Hence we can determine $\boldsymbol{v}$ from a population of filters by observing which filters are activated and finding the best fit plane, see figure (20)(right panel).

In practice, we cannot use filters tuned to frequency because these are not bounded in space and time. But instead we can impose bounds in space and time. For example, converting the filters to spatio-temporal Gabors by multiplying sinusoids by Gaussians (to impose the bounds). The analysis for this case is more complicated because the fourier transforms of the filters are no longer sharply localized in fourier space. But it can be shown [57] that if the filters are spatio-temporal Gabors then the most active filters are those whose spatial-temporal tuning is centered on the plane $\boldsymbol{v} \cdot \boldsymbol{\omega} + \omega_t = 0$. Hence the plane in frequency space can be estimated from a population of spatio-temporal filters and the velocity locally estimated.

This gives rise to a two stage model of motion estimation where the first population of neurons where each neuron (i.e. filter) is sensitive to the spatio-temporal frequency of the input image but not directly to the motion. The population, however, implicity encodes the motion as described above. The second population of neurons extract the motion information from the first population and hence these neurons are tuned directly to motion. This is consistent with the experimental findings [3],[57], [150]. Similar models arise in related work on the fly and beetle visual systems [58,14].

3 Probabilities and Decision Theory

The previous section has given examples about how we can combine the response of many features/filters to perform tasks like stereo or motion estimation. The filter responses were combined by specific mechanisms, e.g., looking for the set of filters which have maximal response. This section describes a more principled approach based on probabilities and decision theory. The section also illustrates the importance of knowing whether filter responses, hence visual cues for the task, are dependent or independent.

We introduce the probabilities of filter responses by describing a classical experimental finding about natural image statistics. Intuitively the intensities of neighboring pixels tend to be similar. This intuition can be captured by taking derivative filters of the image, i.e., $\frac{dI}{dx}$ or $\frac{d^2I}{dx^2}$, and plotting their probability distribution, or histogram. Surprisingly these probability distributions are very similar from image to image [155]. This can be verified from interactive demo (3a) on natural image statistics. Interactive demo (3b) explores the statistics of edge detection and illustrates decision theory.

3.1 Edge Detectors/ Texture Detectors and Decisions.

Consider the tasks of deciding whether an *image patch* at position x contains an *edge* by which we mean the boundary of an object or a strong texture boundary (e.g., like the writing on a tea shirt). The previous section showed that some Gabor filters are tuned (i.e. respond strongly) to edges at specific orientations.

But such filters will also response to other stimuli such as texture patterns, so how can we decide if their response is due to an edge? The simplest way is to *threshold* the response so that an edge, at a specific orientation, is signalled if the filter response is larger than a certain threshold value. But what should that threshold be? How do we do a trade-off to balance *false negative* errors, where we fail to detect a true edge in the image, with *false positive* errors where we incorrectly label a pixel as an edge? Also each filter in a filterbank contains some evidence about the presence of an edge, so how can we combine their evidence in an optimal manner? How can we formulate the intuition that the evidence from some filters give *independent* evidence while others do not.

Decision theory gives a way to address these issues. The theory was developed as a way to make decisions in the presence of uncertainty. In this section we develop the key ideas of decision theory by addressing the specific task of edge detection. In the next section we give a more general treatment. We will only treat the case when we are detecting edges based on local evidence in the image. Later in this chapter we will extend to when we can use non-local, or contextual information. Interactive demo (3b) on decision theory and edge detection illustrates most of the ideas in these two sections.

To start with, we consider the evidence for the presence of an edge using a single filter $f(.)$ only. We assume we have a benchmarked dataset so that at each pixel we have intensity $I(x)$ and a variable $y(x) \in \{\pm 1\}$ (where $y = 1$ indicates an edge, and $y = -1$ does the opposite). We apply the filter to the image to get a set of filter responses $f(I(x))$. If the filter is tuned to edges, then the response $f(I(x))$ is likely to be higher if an edge is present than if not. This requires selecting a filter $f(x)$, such as the modulus of the gradient of the intensity $|\nabla I(x)| = \sqrt{\frac{dI}{dx}^2 + \frac{dI}{dy}^2}$ (since $|\nabla I(x)|$ is likely to be large on edges and small off edges).

To quantify this, we use the benchmarked dataset to learn *conditional probability distributions* for the filter response $f(I)$ conditioned on whether there is an edge or not:

$$P(f(I)|y = 1), P(f(I)|y = -1).$$

Each distribution is estimated by computing the *histogram* of the filter response by counting the numbers of times the response occurs within one of N equally spaced bins and normalizing by dividing by the total number of responses. The histograms for $P(f(I)|y = 1)$ and $P(f(I)|y = -1)$ are computed from the filter responses on the points labeled as edges $\{f(I(x)) : y(x) = 1\}$ and not-edges $\{f(I(x)) : y(x) = 1\}$ respectively. Typical conditional distributions are shown in figure (21).

We can now perform edge detection on an image, see figure (22). At each pixel x we compute $f(I(x))$ and calculate the conditional distributions $P(f(I(x))|y = 1)$ and $P(f(I(x))|y = -1)$. These distributions give local evidence for the presence of edges at each pixel. Note, however, that local evidence for edges is often highly ambiguous, see figure (23). Spatial context can supply additional information to help improve edge detection, as discussed in a later section, and so can high-level knowledge (e.g., by recognizing the objects in the image).

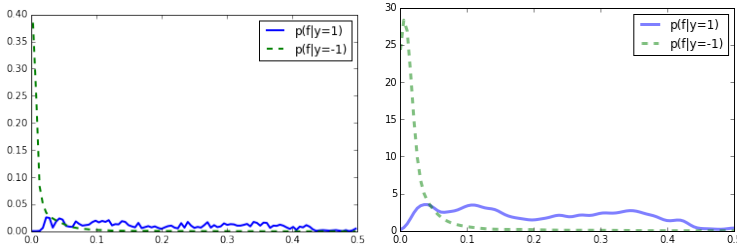

Fig. 21. The probability of filter responses conditioned on whether the filter is *on* or *off* an edge – $P(f|y=1), P(f|y=-1)$, where $f(x) = |\nabla I(x)|$. Left panel: the probability distributions learnt from a dataset of images. Right panel: the smoothed distributions after fitting the data to a parametric model.

Fig. 22. The input image and its groundtruth edges (far left and left). The derivative dI/dx of the image in the x direction (center). The probabilities of the local filter responses $P(\boldsymbol{f}(I(x))|y = 1)$ (right) and $P(\boldsymbol{f}(I(x))|y = -1)$ (far right) have their biggest responses on the boundaries and off the boundaries respectively, hence the log-likelihood ratio $\log \frac{P(\boldsymbol{f}(I(x))|y=1)}{P(\boldsymbol{f}(I(x))|y=-1)}$ gives evidence for the presence of edges.

Fig. 23. The local ambiguity of edges. An observer has no difficulty in detecting all the boundary of the horse if the full image is available (left). But it is much more difficult to detect edges locally (other panels).

The log-likelihood ratio $\log \frac{P(f(I(x))|y=1)}{P(f(I(x))|y=-1)}$ gives evidence for the presence of an edge in image I at position x. This ratio takes large positive values if $P(f(I(x))|y = 1) > P(f(I(x))|y = -1)$ (i.e. if the probability of the filter response is higher given an edge is present) and large negative values if $P(f(I(x))|y = -1) > P(f(I(x))|y = 1)$. So a natural decision criterion is decide that an edge is present if the log-likelihood ratio is greater than zero and that otherwise there is no edge. This can be formulated as a *decision rule* $\alpha(x)$:

$$\alpha(x) = 1, \text{ if } \log \frac{P(f(I(x))|y=1)}{P(f(I(x))|y=-1)} > 0, \;\; \alpha(x) = -1, \text{ if } \log \frac{P(f(I(x))|y=1)}{P(f(I(x))|y=-1)} < 0.$$

This can expressed, more compactly, as

$$\alpha(x) = \arg \max_{y \in \{\pm 1\}} y \log \frac{P(f(I(x))|y=1)}{P(f(I(x))|y=-1)}.$$

Note that this rule gives perfect results (i.e. is one hundred percent correct) if the two distributions do not overlap, i.e. if $P(f(I(x))|y = 1)P(f(I(x))|y = -1) = 0$ for all I. In this case it is impossible to confuse the filter responses to the different types of stimuli. But this situation is very unlikely to happen. Now consider a more general *log-likelihood ratio test* which depends on a threshold T, this gives a rule:

$$\alpha_T(x) = \arg \max_{y \in \{\pm 1\}} y\{\log \frac{P(f(I(x))|y=1)}{P(f(I(x))|y=-1)} - T\}.$$

By varying T we get different types of mistakes. We can distinguish between the *false positives* which are non-edge stimuli which the decision rule mistakenly decides to be an edge, and *false negatives* which are edge stimuli which are mistakenly classified as not being edges. Increasing the threshold T reduces the number of false positives but at the cost of increasing the number of false negatives, while decreasing T has the opposite effect.

Making a decision requires a trade-off between these two types of errors. Bayes decision theory says this tradeoff should depend on two issues. Firstly, the *prior* probability that the image patch is an edge. Statistically most image patches do not contain edges, so we would get a small number of total errors (false positives and false negatives) by simply deciding that every image patch is non-edge. This would encourage us to increase the threshold T (to $-\infty$ so that every image patch would be classified as non-edge). Secondly, we need to consider the *loss* if we make a mistake. If our goal is to detect edges, then we may be willing to tolerate many false positives provided we keep the number of false negatives small. This means we choose a decision rule, by reducing the threshold T, so that we detect all the real edges but also output "false edges", which we hope to remove later by using contextual cues (see the next section). Later we show how this approach can be justified using the framework of decision theory. (In the next section, we see that the log-likelihood ratio is justified as local *evidence* for the presence of an edge even if we are making the decision using non-local context).

Now we consider combining several different filters $\{f_i(.)|i = 1, ..., M\}$ to detect an edge. Generalizing the analysis above, we must learn probability distributions for the *joint* response of all the filters $P(f_1, f_2, \ldots|y) = P(\{f_i(I(x))\}|y)$ *conditioned* on whether the image patch I at x is an edge $y = 1$ or not an edge $y = -1$. This leads to a decision rule:

$$\alpha_T(I(x)) = \arg \max_{y \in \{\pm 1\}} y\{\log \frac{P(\{f_i(I(x))\}|y = 1)}{P(\{f_i(I(x))\}|y = -1)} - T\}.$$

Methods of this type have been applied to edge detection and give good results [87] but they have two related drawbacks. Firstly, the joint distributions require a large amount of data to learn particularly if we represent the distributions by histograms. Secondly, the joint distributions are "black-boxes" and give no insight into how the decision is made. Intuitively, filters tuned to edges at the same orientations will all respond strongly if there is an edge. But this type of intuition is not obvious just from studying the joint distributions. So it is better to try to get a deeper understanding of how the different filters contribute to making this decision, by studying whether they are *statistically independent*.

The response of the filters is statistically independent if:

$$P(\{f_i(I(x))\}|y) = \prod_i P(f_i(I(x))|y) \text{ for each y}$$

This implies that the distributions $P(f_i(I(x))|y)$ can be learnt separately (which decreases the amount of data) and also implies that the log-likelihood test can be expressed in the following form:

$$\alpha_T(x) = \arg \max_{y \in \{\pm 1\}} y\{\sum_i \log \frac{P(f_i(I(x))|y = 1)}{P(f_i(I(x))|y = -1)} - T\}$$

Hence the decision rule corresponds to summing the evidence (the log-likelihood ratio) for all of the filters to determine whether it is above or below the threshold T. This means that each filter gives a "vote", which can be positive or negative, and the decision is based on the sum of these votes. This process is very simple so it is easy to see which filters are responsible for the decision.

Unfortunately, very few filters are statistically independent. For example, the response of each filter will depend on the total brightness of the image patch and so all of them will response more to a "strong" edge than to a "weak" edge. This prevents them from being independent, but it suggests a weaker independence condition known as *conditional independence*. Suppose we add an additional filter $f_0(I(x))$ which, for example, measures the overall brightness. Then it is possible that the other filters are statistically independent conditioned on the value of $f_0(I(x))$:

$$P(\{f_i(I(x))\}, f_0(I(x))|y) = P(f_0(I(x))|y) \prod_i P(f_i(I(x))|f_0(I(x)), y)$$

This requires only representing(learning) the distributions $P(f_i(I(x))|f_0(I(x)), y)$ and $P(f_0(I(x))|y)$. It also leads to a simply decision rule:

$$\alpha_T(x) = \arg\max_{y\in\{\pm 1\}} y\{\log \frac{P(f_0(I(x))|y=1)}{P(f_0(I(x))|y=-1)} + \sum_i \log \frac{P(f_i(I(x))|f_0(I(x)), y=1)}{P(f_i(I(x))|f_0(I(x)), y=-1)} - T\}$$

It has been argued [137] that methods of this type can be implemented by neurons and may be responsible for edge detection. Note that the arguments here are general and do not depend on the type of filters $f_i(.)$ or whether they are linear or non-linear. It has, for example, been suggested that edge detection is performed using the energy model of complex cells [118].

The same approach can be applied to other visual tasks. For example, consider using local filter responses to classify whether the local image patch at x is "sky", "vegetation", "water", "road", or "other", see figure (24). We denote these by a variable $y \in \mathcal{Y}$ (e.g., where $\mathcal{Y} = \{''sky'', '' vegetation'', '' water'', '' road'', or ''other''\}$. We choose a set of filters $\{f_i(I(x))\}$ which are sensitive to texture and color properties of image patches. Then, as before, we learn distributions $P(\{f_i(I(x))\}|y)$ for $y \in \mathcal{Y}$. We select a decision rule of form:

$$\alpha(I(x)) = \arg\max_{y\in\mathcal{Y}} P(\{f_i(I(x))\}|y)T_y,$$

where T_y is a set of thresholds (which can be derived from decision theory).

Fig. 24. Classifying local image patches. The image shows the groundtruth, see [119]. Certain classes – sky, grass, water – can be classified approximately from small image patches.

Experiments on images show that this method can locally estimate the local image class with reasonable error rates for these types of classes [88] and computer vision researchers have improved these types of results using more sophisticated filters. It is unknown, however, whether human or monkey visual systems do make these types of classification.

We stress that the theories described in this section model edge detection *without context*. There are two types of context we will consider in this chapter. The first, discussed in section (4), uses spatial context and is low- or mid-level since it depends only on *generic* properties of images and surfaces. It exploits the idea that edges in natural images are often geometrically regular and co-linear. The second type of context, is high-level and is object specific. For example, if we detect a face in an image then our knowledge about faces enables us to

detect the boundaries of a face better than if we rely only on local edge cues. This second type of context is out of scope of this chapter but is briefly discussed in section (6).

3.2 Bayes Decision Theory and Ideal Observers

We now describe Bayesian analysis in more detail. This is advanced material which can be skipped on a first reading. Bayes decision theory is a framework for making optimal decisions in the presence of uncertainty. We represent the input by $x \in \mathcal{X}$ and the output by $y \in \mathcal{Y}$ (e.g., for *edge detection* x is the filter response $f(I)$, and $y \in \{\pm 1\}$ indicates if an edge is present or not). We assume that there is a probability distribution $P(x, y)$ which generates the input and output. This can be expressed in terms of a *prior* $P(y)$ and a *likelihood* $P(x|y)$ by the identity $P(x, y) = P(x|y)P(y)$. A decision rule is expressed as $\hat{y} = \alpha(x)$. We specify a *loss function* $L(\alpha(x); y)$ which is the cost of making decision $\alpha(x)$ if the real decision should be y.

The *risk* is specified by:

$$R(\alpha) = \sum_{x,y} P(x, y) L(\alpha(x), y)$$

The *Bayes rule* $\hat{\alpha} = \arg\min_\alpha R(\alpha)$. The *Bayes risk* is $\min_\alpha R(\alpha) = R(\hat{\alpha})$ (except for a few highly unusual special cases).

The Bayes rule is the best decision rule you can make (subject to this criterion) and the Bayes risk is the best performance. Hence Bayes Decision Theory can specify the optimal way to estimate y from input x. There are several important special cases. If the loss function penalizes all errors by the same amount, i.e., $L(\alpha(x), y) = K_1$ if $\alpha(x) \neq y$ and $L(\alpha(x), y) = K_2$ if $\alpha(x) = y$ (with $K_1 > K_2$), then the Bayes rule corresponds to the *maximal a posteriori* estimator $\alpha(x) = \arg\max P(y|x)$, where $P(y|x) = \frac{P(x|y)P(y)}{P(x)}$ is the *posterior* distribution of y conditioned on y. If, in addition, the prior is a uniform distribution, i.e. $P(y) = constant$, then Bayes rule reduces to the *maximum likelihood* estimate $\alpha(x) = \arg\max P(x|y)$.

For binary decision problems $y \in \{\pm 1\}$, the loss function is usually chosen to pay no penalty if the correct decision is made – i.e. $\alpha(x) = y$ – but has a penalty F_p for *false positives*, where $y = -1$ but $\alpha(x) = 1$, and F_n for *false negatives* where $y = 1$ but $\alpha(x) = -$ (it is assumed here that the *target* is $y = 1$ and the *distracter* os $y = -1$, so a false positive occurs if we decide that a distracter is a target, and a false negative, if we decide that a target is a distracter). It follows that we can express the Bayes rule in terms of a log-likelihood ratio test $\log \frac{P(x|y=1)}{P(x|y=-1)} > T$, where T depends on the prior $p(y)$ and the loss function $L(\alpha(x), y)$.

More specifically, the Bayes Risk is $R(\alpha) = \sum_x p(x) \sum_y L(\alpha(x), y) p(Y|x)$. Then we divide the data (x, y) into four sets: (i) the *true positives* $\{(x, y) :$ s.t. $\alpha(x) = y = 1\}$. the *true negatives* $\{(x, y) :$ s.t. $\alpha(x) = y = -1\}$, (iii) the *false positives* $\{(x, y) :$ s.t. $\alpha(x) = 1, y = -1\}$, and the *false negatives*

$\{(x, y) :$ s.t. $\alpha(x) == -1, y = 1\}$. These four cases correspond to loss function values $L(\alpha(x) = 1, y = 1) = T_p$, $L(\alpha(x) = -1, y = -1) = T_n$, $L(\alpha(x) = 1, y = -1) = F_p$, $L(\alpha(x) = -1, y = 1) = F_n$ respectively. Then the decision rule $\alpha_T(.)$ reduces to:

$$\log \frac{P(x|y=1)}{P(x|y=-1)} > \log \frac{T_n - F_p}{T_p - F_n} + \log \frac{P(y=-1)}{P(y=1)}.$$

The intuition is that the evidence in the log-likelihood must be bigger than our prior biases while taking into account the penalties paid for different types of mistakes.

The results in the previous section on edge detection and texture classification can be derived from decision theory. The priors $P(y)$ specify the probability that an image patch contains an edge (empirically $P(y = 1) \approx 0.05$ and $P(y = -1) \approx 0.95$). The loss function should be chosen to specify the cost of making different types of mistakes. For texture classification, the variable y takes values in a set $\mathcal{Y}$ which is called multi-class decision. The same theory applies to tasks for which we need to make a set of related, but non-local decisions, which we will address in the next section.

We now show that an important special case of *signal detection theory* [50] – often used as a framework to model how humans make decisions when performing visual, auditory, and other tasks – can be obtained as a special case of Bayes Decision Theory. We consider the two class case, where $y \in \{\pm 1\}$, and suppose that the likelihood functions are specified by Gaussian distributions, $P(x|y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\{-(x - \mu_y)^2/(2\sigma_y^2)\}$, which differ by their means (μ_1, μ_{-1}) and their variances $(\sigma_1^2, \sigma_{-1}^2)$. The Bayes rule can be expressed in terms of the log-likelihood ratio test:

$$\hat{\alpha}(x) = \arg\max_y y\{-(x - \mu_1)^2/(2\sigma_1^2) - \log \sigma_1 + (x - \mu_{-1})^2/(2\sigma_{-1}^2) + \log \sigma_2 - T\}.$$

This has requires determining whether the data x is above or below a quadratic polynomial curve in x. In the special case when the standard deviations are identical $\sigma_1^2 = \sigma_2^2$ (so we drop the subscripts $_{1,-1}$), this reduces to making a decision based on whether the data point x satisfies:

$$2x(\mu_1 - \mu_{-1}) + (\mu_1^2 - \mu_2^2) < 2T\sigma^2$$

This special case, with $\sigma_1^2 = \sigma_{-1}^2$ is much studied in signal detection theory [50]. It means that the decision is based on a single function $d' = \frac{\mu_1 - \mu_{-1}}{\sigma}$. This quantity is used to quantify human performance for psychophysical tasks.

This motivates the idea of an *ideal observer*. An observer like this has optimal performance which requires exploiting the statistical properties of the distribution $P(x, y)$ of the data. A classic example of ideal observer theory shows that under certain conditions photoreceptors in the retina are almost *optimal* at detecting the photons which reach them [8,131]. This takes into account the probability of the photoreceptors *firing* x if it receives a photon $P(x|y = 1)$, the probability that the photoreceptor fires spontaneously $P(x|y = -1)$.

Ideal observers can also be defined for other vision tasks [159,46,161,37]. The difficulty, however, is judging whether humans are adapted to doing the task. It is possible to define ideal observers where human performance is much worse than the ideal observers [177]. Why can this happen? The task may provide information for which humans are not adapted (e.g., visual inspection of circuit boards to find deficits). Also, the ideal observers know the distributions $p(x, y)$ which, for synthetic stimuli, are those chosen by the scientist performing the experiment and may have little similarity to the natural statistics of stimuli of the world, which human vision have probably adapted to.

Another important concept is the receiver operating characteristic (ROC) curve. This allows us to study decisions when we do not want to restrict ourselves to specific priors and loss functions. Instead, we plot the *true positive rate* as a function of the *false positive rate* by allowing the decision threshold T to vary. For each value T of the threshold, we have a decision rule $\alpha_T(.)$ which results in a fraction of *true positives* $\sum_{x:\alpha_T(x)=1} P(x|y=1)$ and *false positives* $\sum_{x:\alpha_T(x)=1} P(x|y=-1)$. This gives a single point on the ROC curve. We plot the curve by allowing T to vary. Observe that for very large T (as $T \mapsto \infty$), the true positive and false positive rates will tend to 0. While as T gets very small ($T \mapsto -\infty$) both rates will tend to 1. Hence the ROC illustrates the trade-off between the two rates.

Note that this section presented Bayes Decision Theory for binary classification. But the same framework can be extended in a straightforward manner if the output y take multiple values. In particular, it applies to cases where we have a set of decision variables defined on each lattice site of an image (see later sections).

3.3 Divisive Normalization

An important example is the use of probabilistic models [170] to account for divisive normalization. Divisive normalization is a mechanism whereby cells mutually inhibit one another, effectively normalizing their responses with respect to stimulus inputs. Originally developed to explain non-linear responses to contrast in V1 [59], divisive normalization has been proposed as a basic cortical computation that underlies various effects of context (see next section), as well as higher-level processes such as attention [20].

The probabilistic approach give a theoretical justification for divisive normalization in V1. The main idea is that filters with similar preferences for orientation representing nearby spatial locations in a scene have striking statistical dependencies, which can be removed by divisive normalization. Specifically, if we plot the statistics of two linear filters f_c, f_s (center and surround) then the magnitudes of f_c, f_s are coordinated in a straightforward way, which has a characteristic shape of a Bow-Tie.

This can be modeled by assuming there are hidden variables ν which affect both responses and hence induces correlation between the responses (as discussed earlier for edge detection). For example, ν could represent the local average

image intensity which could affect the response of both filters but, after the filter response could be made independent by conditioning on the average intensity. Suppose ν has a prior distribution $P(\nu) = \nu \exp\{-\nu^2/2\}$ for $\nu \geq 0$. We have a pair of filters $\{l_i : i = 1, 2\}$ which are related to gaussian models $\{g_1 : i = 1, 2\}$ (the analysis can be generalized to arbitrary number of filters [152]). The claim is that we can model the activation of the set of filter responses:

$$P(l_1, l_2) = \int d\nu P(\nu) \prod_{i=1}^{2} P(l_i|\nu, g_i))P(g_i), \tag{25}$$

where $P(l_i|\nu, g_i) = \delta(l_i - \nu g_i)$. In this model the filter responses are generated by independent processes, g_1, g_2, but then are multiplied by the common factor ν. This is illustrated in figure (25).

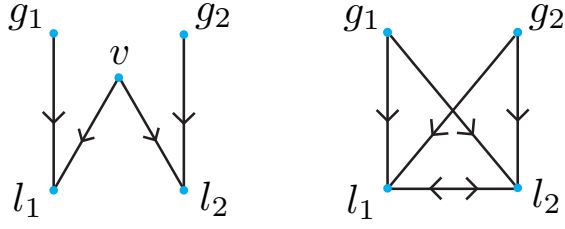

Fig. 25. Left Panel: The graphical structure of the divisive normalization model. The filter responses l_1, l_2 are generated from stimuli g_1, g_2 respectively and by the common factor ν. The distributions of l_1, l_2 are factorized if we condition on ν. Right Panel: But if we integrate out ν then almost all the variables become dependent as reflected by the complexity of the graph structure.

In particular, for each filter we can compute $P(g_i|l_1, l_2)$. After some algebra, this is computed to be:

$$P(g_1|l_1, l_2) = \frac{g_1^{-1} \exp\{-\frac{g_1^2 l^2}{2\sigma^2 l_1^2} - \frac{l_1^2}{2g_1^2}\}}{B(0, l/\sigma)}, \tag{26}$$

where $l = \sqrt{l_1^2 + l_2^2}$, and $B(.,.)$ is a Bessel's function (a well-known class of mathematical functions that arise in solving differential equations). To get intuition, note that $g_1 = l_1/\nu$ and $g_1 = l_1/\nu$. So if ν is small then $|l_1|$ and $|l_2|$ are likely to be small together, while if ν is large, then $|l_1|$ and $|l_2|$ are both likely to be large.

Assume that the goal of a model unit is to estimate the g_i from the observed filter responses $\{l_i : i = 1, 2\}$, which gives the non-linear response of the cell. It follows, from analysis above, that

$$E(g_1|l_1, l_2) \propto \text{sign}\{l_1\}\sqrt{|l_1|}\sqrt{\frac{|l_1|}{\sqrt{l_1^2 + l_2^2 + k}}}. \quad (27)$$

The $\sqrt{l_1^2 + l_2^2 + k}$ term sets the gain and performs the divisive normalization.

The model has also been applied to explain the classic tilt illusion in perception [152,136]. In the "simultaneous" tilt illusion, a set of vertically oriented lines appears to tilt right when surrounded by an annulus of lines tilted left–an effect called "repulsion". However, for large differences between the center orientation and surround (tilted left), the center vertical lines can appear to tilt left–an effect called "attraction". In their model, the population of neurons responding to the surround tilted lines contribute to divisive normalizing of the neurons responding to the center stimulus. This results in a change of their neural tuning curves which, together with the degree of coupling between center and surrounds, accounts for repulsion and attraction.

The suppressive effect of surround contrast on a central region is an example of spatial context. There are many phenomena like this whose effects vary in type and extent of their spatial interactions. The next section introduces a broader range of contextual phenomena and a set of computational techniques to model them.

4 Context and Spatial Interactions between Neurons

There is considerable evidence that low-level vision involves long-range spatial interactions so that human perception of local regions of an image can be strongly influenced by their spatial context. Psychophysicists have discovered many perceptual phenomena demonstrating spatial interactions. For example, local image regions which differ from their neighbors tend to "pop-out" and attract attention while, conversely, similar image features which form spatially smooth structures tend to get "grouped" together to form a coherent percept, see figure (26)(left panel). Image properties such as color tend to spread out, or fill-in regions, until they hit a boundary [55][148] as shown in figure (26)(right panel). In general, there is a tendency for low-level vision to group together similar image features and make breaks at places where the features change significantly. These perceptual phenomena are not surprising from a theoretical perspective since they correspond to low-level visual tasks such as segmentation and the detection of salient features. Segmenting an image into different regions is one of the first stages of object recognition (in the ventral stream) and a pre-cursor to estimating the three-dimensional structure of objects, or surfaces, in order to grasp them or avoid them (dorsal stream). Detection of salient features has many uses including bottom-up attention [68]. It has been suggested that many of these processes are performed in V1 [190] although possibly this involves feedback

and interactions between V1 and V2 [154]. Note that, as discussed in the first section, low-level vision interacts with high-level vision and it is very unlikely that tasks like segmentation are performed entirely by low-level processes.

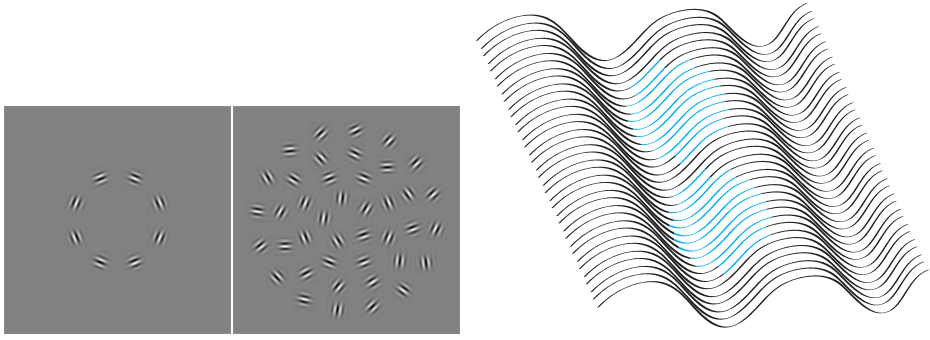

Fig. 26. Left Panel: association fields. The circular alignment of gabor patches (left panel) make it easier to see the circular form in the presence of clutter (right panel). Right Panel: The neon color illusion. A bluish color appears to fill in the white regions between the blue lines creating the appearance of blue transparent disks.
d

These psychophysical and theoretical studies are supported by single-electrode studies [90], [96] which show that the activities of neurons on monkey area V1 appear to involve spatial interactions with other neurons. When monkeys are shown stimuli like figure 33 their responses over the first 60 msec are similar to those predicted by classic models (e.g., previous sections) but their later activity spreads in from the boundaries, roughly similar to predictions of computational models [183]. There is also a considerable literature the related topic of *non-classical receptive fields* [75].

This section discusses neural networks models which address these phenomena. Although the models capture the essence of the phenomena they are simplifications in three respects. Firstly, they use simple models of neurons and it is currently not possible to compare them directly to real neural circuits. Secondly, these models are formulated in terms of lateral, or horizontal connections. Thirdly, the performance of these models on natural images is significantly worse than human's. Although there are more advanced computer vision models, built on similar principles, whose performance starts to approach human vision (unless high level cues are present, which humans can exploit).

We formulate these models in terms of probability distributions defined over graphs, where the nodes of the graph represent neurons. This differs from some of the standard "neural network" models for these types of phenomena, see [56]. but our approach has several advantages. Firstly, this enables us to use a coherent framework which unifies the models in this section with those we will present in later sections. Secondly, it puts the models in a form where they can be

directly related to a class of computer vision models. Thirdly, this probabilistic formulation is of increasing use in models of Artificial Intelligence, Cognitive Science, and in the machine learning and statistical techniques used to analyse experimental neuroscience data. Fourthly, it is possible to derive many of these neural network models as approximations to the probability models.

We start by first introducing probabilistic models of neurons and showing how our previous linear filter models can be derived as approximations. Next we introduce neural network models and show their relationship to probability models. Then the following section uses this material to derive some specific models for a range of visual tasks. This section has three interactive demos: (4a) Gibbs sampling. (4b) Mean Field Theory and Neural Models. (4c) Hopfield Networks and Stereopsis.

Single Neurons: Probabilistic Model and Integrate and Fire. In the previous section, we described neurons as linear filters and briefly mentioned thresholds and non-linearities. In this section we provide a more realistic model of a *stochastic neuron* where the neuron has a probability of firing an action potential. We will show how linear filters, thresholds, and non-linearities can be obtained as approximations to this stochastic model. This stochastic model is, in turn, an approximation and we refer to the literature for more realistic models such as assuming that the probability of firing is specified by a Poisson process [139]. For simplicity, we restrict ourselves to the simpler stochastic *integrate-and-fire* model which is easier to analyze and to relate to computational models.

In the integrate-and-fire model a neuron i receives input I_j at each dendrite j. These inputs are weighted by the synaptic strengths w_{ij} and sent along the dendrites to the soma. At the soma, these weighted inputs are summed linearly to yield summed linearly to yield $\sum_j w_{ij} I_j$. The probability of firing $s_i = 1$, or not firing $s_i = 0$, is given by:

$$P(s_i|\boldsymbol{I}) = \frac{\exp\{s_i(\sum_j w_{ij} I_j - T_i)\}}{1 + \exp\{\sum_j w_{ij} I_j - T_i\}}, \tag{28}$$

where T_i is a threshold.

To relate this stochastic model to our earlier linear models, we calculate the probability that the neuron fires. This is given by a sigmoid function:

In particular, the probability of firing ($s_i = 1$) is given by a sigmoid function:

$$\sum_{s_i=0}^{1} s_i P(s_i|\boldsymbol{I}) = \frac{1}{1 + \exp\{\sum_j w_{ij} I_j - T_i\}} = \sigma(\sum_j w_{ij} I_j - T_i). \tag{29}$$

Observe that this is also the *expected firing rate* $\sum_{s_i=0,1} s_i P(s_i|\boldsymbol{I})$ because

$$\sum_{s_i=0,1} s_i P(s_i|\boldsymbol{I}) = P(s_i = 1|\boldsymbol{I}) = \sigma(\sum_j w_{ij} I_j - T_i). \tag{30}$$

Hence, by computing the expected firing rate, we can obtain a deterministic approximation to a stochastic neuron. This is a sigmoid function of a linear

weighted sum of the input (minus a threshold). The sigmoid function is approximately linear for small inputs, saturates at value 1 for large positive inputs, and suppresses large negative inputs to 0. Hence there is a linear regime where the probability of firing is $\sum_j w_{ij} I_j - T_i$. This enables us to recover the linear models used in the previous section as an approximation.

Next we modify the model so that it deals with non-linear image features. This allows us to relate it to the types of computational models described in the previous section and will enable us to construct richer models of this type that can deal with spatial context.

To make this idea concrete, we consider detecting whether there is an edge or not at pixel x, or alternatively classifying whether a pixel is foreground or background. In either task, we formulate the problem as Bayes estimation where we have conditional distributions $P(f(I(x))|s)$ and priors $P(s)$ for $s \in \{0, 1\}$. The posterior distribution $P(s|f(I(x)))$ can be expressed in form:

$$P(s|f(I(x))) = \frac{1}{Z} \exp\{s(\log \frac{P(f(I(x))|s=1)}{P(f(I(x))|s=0)} + \log \frac{P(s=1)}{P(s=0)})\},$$

where Z is a normalization constant (chosen so that $\sum_{s=0}^{1} P(s|f(I(x))) = 1$). This shows that the posterior distribution for the presence of an edge, or a foreground object, can be expressed in the same form. The only difference is that the input is a nonlinear function of the image instead of the image itself. This claim can be justified by expressing $P(f(I(x))|s) = \{P(f(I(x))|s = 1)\}^s \{P(f(I(x))|s = 0)\}^{1-s}$, $P(s) = \{P(s = 1)\}^s \{P(s = 0)\}^{1-s}$, substituting these into the posterior $P(s|f(I(x))) = P(f(I(x))|s)P(s)/P(f(I(x)))$.

Probability Models with Context. Now we consider generalizing the model for foreground/background classification so that it can include spatial context. Intuitively neighboring pixels in the image are likely to belong to the same class, i.e. are likely to be either all background or all foreground. This is a form of prior knowledge, or natural statistic, which can be learnt by analyzing natural images.

We now specify neurons by spatial position $\boldsymbol{x}$ instead of index i. As above, we have distributions $P(f(I(\boldsymbol{x}))|s)$ for the features $f(I(\boldsymbol{x}))$ at position $\boldsymbol{x}$ conditioned on whether this is part of the foreground object $s(\boldsymbol{x}) = 1$, or not $s(\boldsymbol{x}) = 0$. We use the notation $\boldsymbol{S}$ to be the set of the states of all neurons $\{s(\boldsymbol{x})\}$. We also specify a prior distribution:

$$P(\boldsymbol{S}) = \frac{1}{Z} \exp\{-\gamma \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y} \in N(\boldsymbol{x})} \{s(\boldsymbol{x}) - s(\boldsymbol{y})\}^2\},$$

where γ is a constant. This prior uses a neighborhood $N(\boldsymbol{x})$ which specifies those spatial positions which directly interact with $\boldsymbol{x}$ in the model. In graphical terms, the positions $\boldsymbol{x}$ are the nodes $\mathcal{V}$ of a graph $\mathcal{G}$ and the edges $\mathcal{E}$ specify which nodes are connected. This is illustrated in figure (27)(far left panel).

Formally, the edges of the graph define the *Markov structure* of the probability distribution $P(\boldsymbol{S})$. It can be shown that the conditional distribution of the state $s(\boldsymbol{x})$ at one position depends *only* on the states of positions in its neighborhood $N(\boldsymbol{x})$. This is the *Markov condition*:

$$P(s(\boldsymbol{x})|\boldsymbol{S}/s(\boldsymbol{x})) = P(s(\boldsymbol{x})|\{s(\boldsymbol{y}) : \boldsymbol{y} \in N(\boldsymbol{x})\}),$$

where $\boldsymbol{S}/s(\boldsymbol{x})$ denotes all states in $\boldsymbol{S}$ except $s(\boldsymbol{x})$. In real vision applications this type of prior, including the size of the neighborhoods, can be estimated from the statistics of natural images.

Next, we define a probability model for the observed image features at positions $\boldsymbol{x}$ in the image. We use the same models as before, at each position $\boldsymbol{x}$:

$$P(f(I(\boldsymbol{x}))|s) = \{P(f(I(\boldsymbol{x}))|s=1)\}^s\{P(f(I(\boldsymbol{x}))|s=0)\}^{1-s}.$$

We combine these, using independence assumptions, to get a distribution:

$$P(f(\boldsymbol{I})|\boldsymbol{S}) = \prod_{\boldsymbol{x}} P(f(I(\boldsymbol{x}))|s) = \frac{1}{Z_l}\exp\{\sum_{\boldsymbol{x}} s(\boldsymbol{x})(\log\frac{P(f(I(\boldsymbol{x}))|s=1)}{P(f(I(\boldsymbol{x}))|s=0)}\},$$

where Z_l is a normalization term (which can be calculated directly).

These distributions $P(f(\boldsymbol{I})|\boldsymbol{S})$ and $P(\boldsymbol{S})$ can be combined to get the posterior distribution $P(\boldsymbol{S}|f(\boldsymbol{I}))$ which is of form:

$$P(\boldsymbol{S}|f(\boldsymbol{I})) = \frac{1}{Z_p}\exp\{-E(\boldsymbol{S})\},$$

where

$$E(\boldsymbol{S}) = -\sum_{\boldsymbol{x}} s(\boldsymbol{x})\log\frac{P(f(I(\boldsymbol{x}))|s=1)}{P(f(I(\boldsymbol{x}))|s=0)} + \sum_{\boldsymbol{x}}\sum_{\boldsymbol{y}\in N(\boldsymbol{x})}\gamma\{s(\boldsymbol{x})-s(\boldsymbol{y})\}^2.$$

The first term of $E(\boldsymbol{S})$ gives the local cues for foreground or background (the log-likelihood ratios of the features) while the second term adds the local context. This context encourages neighboring positions to be either all foreground or all background. Note that this method of specifying a distribution $P(\boldsymbol{S})$ in terms of a function $E(\boldsymbol{S})$ will keep re-occurring throughout this section.

This model specifies the posterior distribution for foreground-background classification using spatial context and, as we will show, similar methods can be applied to other visual tasks. But there remains the issue of how to estimate the most probable states, i.e. computing the Bayes estimator.

$$\hat{\boldsymbol{S}} = \arg\max P(\boldsymbol{S}|f(\boldsymbol{I})).$$

In the next two sections we will discussion neurally plausible algorithms which can do this. There are two types: (i) stochastic models which are natural extensions of the probabilistic neural models discussed earlier and, which the Statistics literature, are called *Gibbs* samplers [102], and (ii) neural network models which are based on simplified biophysics of neurons but which can also, in certain cases, be related to *mean field approximations* to the stochastic models.

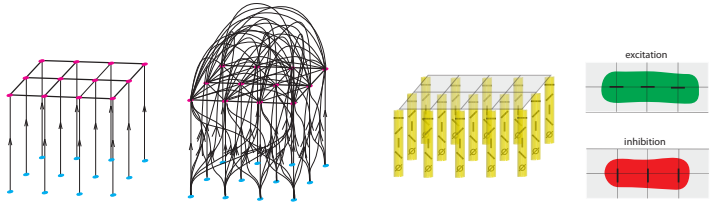

Fig. 27. Far Left Panel: The graphical structure of the Markov model with nearest neighbor connections. Left Panel: a fully connected graphical model. Right Panel: A hyper-column structure where neurons within each column are tuned to different orientations and inhibit each other. Far Right Panel: Edges have excitation (green) along the direction of the edge and inhibition (red) perpendicular to the edge.

Probabilistic models of groups of neurons. In this section we introduce a more general probability distribution. It is also specified by a model defined over a graph where the nodes correspond to neurons and the edges to connections between them. But we will not make any Markov restrictions on the edges and so this model can be fully connected, see figure (27)(left panel).

More specifically, we have set of M neurons with states $\boldsymbol{S} = (s_1, ..., s_M)$ and with input $\boldsymbol{I} = (I_1, ..., I_N)$. We specify a *Gibbs probability distribution* over the set of activity of all neurons $\boldsymbol{S} = (s_1, ..., s_n)$ as follows. First we define an energy function:

$$E(\boldsymbol{S}, \boldsymbol{I} : \boldsymbol{W}, \boldsymbol{\theta}) = -\sum_{ij} W_{ij} s_i I_j - (1/2) \sum_{kl} \theta_{kl} s_k s_l.$$

This energy contains two types of terms: (i) those of form $s_i I_j$ which give the interactions between the states of the neurons $\boldsymbol{S}$ and the input $\boldsymbol{I}$ and (ii) those which specify interactions between the neurons. This energy is used to specify a *Gibbs distribution*:

$$P(\boldsymbol{S}, \boldsymbol{I}) = \frac{1}{Z} \exp\{-E(\boldsymbol{S}, \boldsymbol{I} : \boldsymbol{W}, \boldsymbol{\theta})\}. \tag{31}$$

Here Z is a normalization constant chosen to ensure that $\sum_{\boldsymbol{S}} P(\boldsymbol{S}|\boldsymbol{I}) = 1$. We note that Gibbs distribution originally arose in statistical physics where they specify the probability distribution of a physical system in thermal equilibrium. Here the physical energy of the system is E and the distribution can be derived using the maximum entropy principle.

The weights $\{w_{ij}\}, \{\theta_{kl}\}$ specify the strength of the interactions between the neuron and the inputs, and between the neurons and each other. In particular, the *interaction term* $\sum_{kl} \theta_{kl} s_k s_l$ specifies the interactions between the neurons. If this term was not present, then the distribution simplifies and it can be expressed as a product of independent distributions:

$$P(\boldsymbol{S}|\boldsymbol{I}) = \frac{1}{Z} \exp\{\sum_{ij} w_{ij} s_i I_j\} = \prod_{i=1}^{n} P(s_i|\boldsymbol{I}). \tag{32}$$

Hence in this special case the neurons act independently and are driven purely by the input (i.e. there is no context). As a technical point, in this case the normalization factor can be computed directly as $Z = \prod_i Z_i$, where $Z_i = \sum_{s_i=0}^{1} \exp\{\sum_j w_{ij} s_i I_j\}$.

Observe that we can recover the foreground/background model in the previous section by specifying values of the weights. This requires that we identify the nodes i with positions $\boldsymbol{x}$.

Now we specify stochastic dynamics on this model. These dynamics have two purposes. Firstly, to describe the activities of sets of neurons interacting with each other. Secondly, to give algorithms for estimating properties such as the most probable configurations of the states $\boldsymbol{S}$, which can be used for visual tasks and for making decisions.

To specify stochastic dynamics, we generalize the stochastic neural model, see equation (28), to deal with a set of neurons. A neuron received input $\boldsymbol{S}$ from other neurons in addition to direct input from the stimulus $\boldsymbol{I}$. Consider only the activity of this neuron, fixing the states of all the others. Then the neurons will have total input of $\sum_j w_{ij} I_j$ plus input $\sum_k \theta_{ik} s_k$ from the other neurons. Then, extending equation (28), the probability that the cell i fires is:

$$P(s_i|\boldsymbol{I}, \boldsymbol{S}_{/i}) = \frac{1}{Z_i} \exp\{s_i(\sum_j w_{ij} I_j + \sum_{k \neq i} \theta_{ik} s_k)\}. \tag{33}$$

where the notation $\boldsymbol{S}/i$ means the states $\{s_j : j \neq i\}$ of all the neurons except the neuron we are considering. The term Z_i is defined so that the distribution is normalized, so it is given by $Z_i = 1 + \exp\{\sum_j w_{ij} I_j + \sum_{k \neq i} \theta_{ik} s_k\}$.

This gives the following dynamics for a group of neurons. At each time, a neuron is selected at random and fires with a probability specified by equation (33). This model assumes that no neurons ever fire at the same time and ignores the time for a spike fired from one neuron to reach other neurons. This is illustrated in interactive demo (4a).

How does this stochastic dynamics relate to the Gibbs distribution specified above? From the statistical perspective, this is an example of *Markov Chain Monte Carlo* (MCMC) sampling [102]. MCMC refers to a class of algorithms which explore the state space of $\boldsymbol{S}$ stochastically so that it will gradually move to configurations which have high probability $P(\boldsymbol{S}|\boldsymbol{I})$. More precisely, MCMC algorithms are guaranteed to give samples from the Gibbs distribution — $\boldsymbol{S}_1, ..., \boldsymbol{S}_M \; P(\boldsymbol{S}|\boldsymbol{I})$. The stochastic update rule in equation (33) is a special type of MCMC algorithm which is known as a *Gibbs sampler*, because it samples from the conditional distribution $P(s_i|\boldsymbol{I}, \boldsymbol{S}_{/i})$. These samples enable us to estimate the most probable state of the system $\hat{\boldsymbol{S}} = \arg\max P(\boldsymbol{S}|\boldsymbol{I})$, hence they can estimate the MAP estimator of $\boldsymbol{S}$ and make optimal decisions for visual tasks.

To apply these models to visual tasks, we need to specify the weights. One strategy is purely data driven and consists of learning the weights from training examples, this is the *Boltzmann Machine* [2] which is out of scope for this chapter. Another strategy is to specify distributions for specific visual tasks, and we will give examples in the next few sections.

Dynamical System Models of Neurons. There is an alternative way to model sets of neurons using *dynamical systems* based on simplified models of their biophysics [139],[26]. Pioneering work on this topic was done by Wilson and Cowan [180], Grossberg and Mingolla [53,56], Hopfield and Tank [63], Abbott and Kepler [1], and others. There is no space to cover the richness of these models and, in any case, our chapter concentrates on the probabilistic formulation. But we will discuss an important subclass of dynamical models [63] which, as we will show, have very close relations to the probabilistic approach.

Following Hopfield and Tank, these dynamical systems are described as follows. A neuron is described by two (related) variables: (i) a continuous valued variable $u_i \in \{-\infty, \infty\}$, and (i) a continuous variable $q_i \in \{0, 1\}$. Roughly speaking, u_i represents the input to the cell body (soma), due to the direct input and the input from other neurons, and q_i describe the probability that the cell will fire an action potential. These variables are related by the equations $u_i = \log(q_i/(1-q_i))$ or, equivalently, by $q_i = \sigma(u_i)$ (where $\sigma(.)$ is the sigmoid function).

The dynamics of the neuron is given by:

$$\frac{du_i}{dt} = -u_i + \sum_j w_{ij} I_j + \sum_k \theta_{ik} q_k. \tag{34}$$

Here, as before, $\sum_j w_{ij} I_j + \sum_k \theta_{ik} q_k$ represent the direct input and the input from the other neurons.

It can be shown, next section, that this dynamic system continually decreases a function $F(\boldsymbol{q})$, so that $(dF)/dt \leq 0$. The function F acts as a *Lyaponov function* for the system in the sense that it decreases monotonically as time t increases and is bounded below. The existence of a Lyaponov function for the dynamics guarantees that the system converges to a state which minimizes $F(\boldsymbol{q})$ (note that $F(\boldsymbol{q})$ will typically have many minimum, and the system may converge to any one of them). This dynamical system is illustrated in interactive demo (4b).

Relations between probabilistic models and dynamical system models. Perhaps surprisingly, there is a very close relationship between the dynamic systems in equation (34) and the stochastic update in equation (28). More specifically, the dynamic system is a *mean field approximation* to the stochastic dynamics. *Mean field theory* (MFT) was developed by physicists as a way to approximate stochastic systems.

To explain this relationship we first define *the mean field free energy* $F(\boldsymbol{q})$:

$$F(\boldsymbol{q}) = -\sum_{ij} W_{ij} I_j q_i - (1/2) \sum_{ij} \theta_{ij} q_i q_j + \sum_i \{q_i \log q_i + (1-q_i) \log(1-q_i)\}. \tag{35}$$

Next we specify dynamics by performing steepest descent on the free energy (multiplies by a positive factor):

$$\frac{dq_i}{dt} = -q_i(1-q_i) \frac{\partial F(\boldsymbol{q})}{\partial q_i}. \tag{36}$$

Interestingly these are identical to the dynamical system in equation (34). This can be seen by introducing a new variable $u_i = \log q_i/(1-q_i)$, which implies that $q_i = \sigma(u_i)$. Note that $\partial F/\partial q_i = -\sum_j W_{ij}I_j - \sum_j \theta_{ij}q_j + \log q_i/(1-q_i)$, $u_i = \log q_i/(1-q_i)$, and $dq_i/q_i(1-q_i) = du_i$.

Equation (36) implies that the dynamical system decreases the free energy $F(\boldsymbol{q})$ monotonically with time t. This is because $dF/dt = -\sum_i(\partial F/\partial q_i)(\partial q_i/\partial t) = -\sum_i q_i(1-q_i)(\partial F/\partial q_i)^2$. Hence $F(\boldsymbol{q})$ is a Lyaponov function for equations (34,36) and so the dynamics converges to a fixed point.

This shows that there is a close connection between the neural dynamical system and minimizing the mean field free energy. In turn, the mean field free energy is related to deterministic approximations to stochastic update methods like Gibbs sampling [4] [60]. This connection is technically advanced and is not needed to understand the rest of this chapter. Briefly, the mean field free energy $F(\boldsymbol{q})$ is the *Kullback-Leibler divergence* $F(Q) = \sum_{\boldsymbol{S}} Q(\boldsymbol{S}) \log \frac{Q(\boldsymbol{S})}{P(\boldsymbol{S}|\boldsymbol{I})}$ between the distribution $P(\boldsymbol{S}|\boldsymbol{I})$ and a factorized distribution $Q(\boldsymbol{S}) = \prod_i q_i^{S_i}(1-q_i)^{1-S_i}$ (plus an additive constant, this can be verified by substitution). Hence the dynamical system seeks to find the factorized distribution $\hat{Q}(\boldsymbol{S})$ which best approximates $P(\boldsymbol{S}|\boldsymbol{I})$ by minimizing the Kullback-Leibler divergence. In this approximation the response q_i is an approximation to the expected response $\sum_{S_1} S_i P(\boldsymbol{S}|\boldsymbol{I})$. The connections between mean field theory and neural models was described in [182]). For technical discussions about mean field theory and Gibbs sampling see [184].

4.1 The Line Process Model

Our first example is the classic *line process* model [40][12][122] which was developed as a way to segment images. It has explicit *line process* variables which "break" images into regions where the intensity is piecewise smooth. Our presentation will follow the work of ([85]) who translated it into neural circuits.

The model takes intensity values $\boldsymbol{I}$ as input and outputs smoothed intensity values. But this smoothness is broken at places where the intensity changes are too high, see figure (28). The model has continuous variables $\boldsymbol{J}$ representing the intensity and binary-valued variables $\boldsymbol{l}$ for the line processes (or edges). The model is formulated as performing *maximum a posteriori* (MAP) estimation. The algorithm for estimating MAP is a neural network model which can be derived from the original Markov Model [40] by mean field theory [36]. Note that in this model the variables do not have to represent intensity. Instead they can represent texture, depth, or any other property which is spatially smooth except at sharp discontinuities.

For simplicity we present the weak membrane model in one-dimension. The input is $\boldsymbol{I} = \{I(x) : x \in \mathcal{D}\}$, the estimated, or smoothed, image is $\boldsymbol{J} = \{J(x) : x \in \mathcal{D}\}$, and the line processes are denoted by $\boldsymbol{l} = \{l(x) : x \in \mathcal{D}\}$, where $l(x) \in \{0, 1\}$.

The model is specified by a posterior probability distribution:

$$P(\boldsymbol{J}, \boldsymbol{l}|\boldsymbol{I}) = \frac{1}{Z}\exp\{-E[\boldsymbol{J}, \boldsymbol{l} : \boldsymbol{I}]/T\},$$

where

$$E[\boldsymbol{J},\boldsymbol{l}:\boldsymbol{I}] = \sum_x (I(x)-J(x))^2 + A\sum_x (J(x+1)-J(x))^2(1-l(x)) + B\sum_x l(x).$$

The first term ensures that the estimated intensity $J(x)$ is close to the input intensity $I(x)$. The second encourages the estimated intensity $J(x)$ to be spatially smooth (e.g., $J(x) \approx J(x+1)$), unless a line process is activated by setting $l(x) = 1$. The third pays a penalty for activating a line process. The result encourages the estimated intensity to be piecewise smooth unless the input $I(x)$ changes significantly, in which case a line process is switched on and the smoothness is broken. The parameter T is the variance of the probability distribution and has a default value $T = 1$.

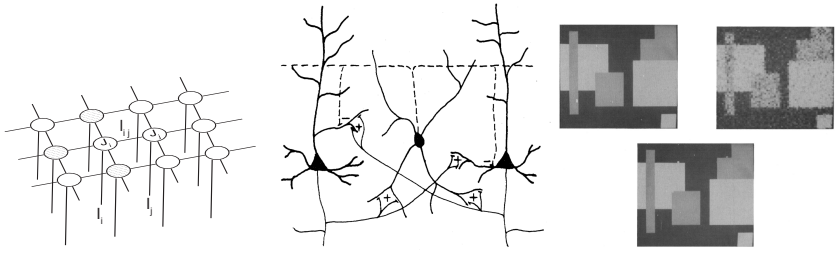

Fig. 28. A representation of the Line-process model (far left) compared to a real neural network (left). On the right we show the original image (upper left), the image corrupted with noise (upper right) and the image estimated using the line-process model (bottom).

This model can be implemented by a neural circuit [85]. The connections between these neurons is shown in figure (28). To implement this model [85] proposed a neural net model which is equivalent to doing mean field theory on the weak membrane MRF (as discussed earlier) by replacing the binary-valued line process variables $l(x)$ by continuous variables $q(x) \in [0,1]$ (corresponding roughly to the probability that the line process is switched on).

This gives an algorithm which updates the regional variables $\boldsymbol{J}$ and the line variables $\boldsymbol{q}$ in a coupled manner. It is helpful, as before, to introduce a new variable $\boldsymbol{u}$ which relates by $q(x) = \frac{1}{1+\exp\{-u(x)/T\}}$ and $u(x) = T\log\frac{q(x)}{1-q(x)}$.

$$\frac{dJ(x)}{dt} = -2(J(x)-I(x))$$

$$= -2A\{(1-q(x))(J(x)-J(x+1)) + (1-q(x-1))(J(x)-J(x-1))\}, \quad (37)$$

$$\frac{dq(x)}{dt} = \frac{1}{T}q(x)(1-q(x))\{A(J(x+1)-J(x))^2 - B - T\log\frac{q(x)}{1-q(x)}\}, \quad (38)$$

$$\frac{du(x)}{dt} = -u(x) + A(J(x+1)-J(x))^2 - B. \quad (39)$$

The update rule for the estimated intensity $\boldsymbol{J}$ behaves like non-linear diffusion which smooths the intensity while keeping it similar to input $\boldsymbol{I}$. The diffusion is modulated by the strength of the edges $\boldsymbol{q}$. The update for the lines $\boldsymbol{q}$ is driven by the differences between the estimated intensity, if this is small then the lines are not activated.

This algorithm has a Lyaponov function $L(\boldsymbol{J}, \boldsymbol{q})$ (derived using mean field theory methods) and so will converge to a fixed point, with

$$L(\boldsymbol{J}, \boldsymbol{q}) = \sum_x (I(x) - J(x))^2 + A \sum_x (J(x+1) - J(x))^2 (1 - q(x)) + B \sum_x q(x) + T \sum_x \{q(x) \log q(x) + (1 - q(x)) \log(1 - q(x))\} \quad (40)$$

There is some evidence that a generalization of this models roughly matches the electrophysiological findings for those types of stimuli shown in figure (33). The generalization is performed by replacing the intensity variables $I(x), J(x)$ by a filterbank of Gabor filters so that the weak membrane model enforces edges at places where the texture properties change [95]. The experiments, and their relation to the weak membrane models are reviewed in [96]. The initial responses of the neurons, for the first 80 msec, are consistent with the linear filter models described in section (2). But after 80 msec the activity of the neurons change and appear to take spatial context into account. The findings of the electrophysiological experiments are summarized as follows: (1) There are two sets of neurons where one set encodes regional properties (such as average brightness) and the other set codes boundary location (in agreement with J and l variable in the model respectively). (2) The processes for computing the region and the boundary representations are tightly coupled, with both processes interacting with and constraining each other (as in the dynamical equations above). (3) During the iterative process, the regional properties diffuse within each region and tend to become constant, but these regional properties do not cross the region (in agreement with the model). (4) The interruption of the spreading of regional information by boundaries results in sharp discontinuities in the responses across two different regions (in agreement with the model). The development of abrupt changes in regional responses also results in a gradual sharpening of the boundary response, reflecting increased confidence in the precise location of the boundary. These findings are roughly consistent with neural network implementations of the weak membrane model. But other explanations are possible. For example, the weak membrane model requires lateral (sideways) interaction and it is possible that the computations are done hierarchically using feedback from V2 to V1.

While the weak membrane model is broadly consistent with the perceptual phenomena of segmentation and "filling-in", the types of filling-in, their dynamics, and the neural representations of contours and surface is complicated [167,86]. Exactly how contour and surface information is represented and processed in cortex is an active topic of research [54,142].

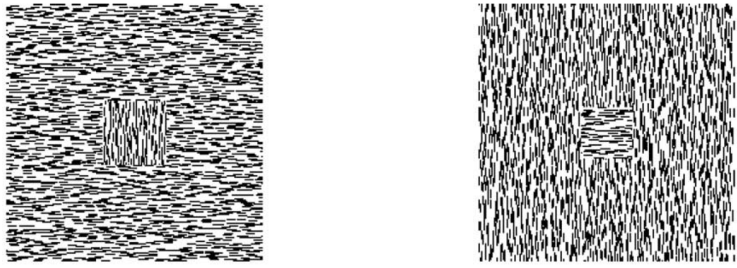

Fig. 29. The stimuli for the experiments in the experiments by TS Lee and his collaborators [96].

4.2 Edge Detection with Spatial Context

Our second example is to develop a model for detecting edges using spatial context. This relates to the phenomena known as association fields, see figure (26)(left panel), where Gabor filters which are spatially aligned (in orientation and direction) get grouped into a coherent form.

For this model, we have a set of neurons at every spatial position x, each tuned to a different angle $\theta_i : i = 1, ..., 8$, and a default cell at angle θ_0. The first cells are designed to detect edges at each orientation – i.e. they can be driven by the log-likelihood ratio of an edge detector at orientation θ_i at this position. The default cell is a dummy that is intended to fire if there is no edge present at this position. This organization forms a population of cells arrayed according to orientation (similar to a hypercolumn in V1). See figure (27)(right panel).

We define a Gibbs distribution for the activity s_{x,θ_i} of the cells. The energy function $E(\boldsymbol{s})$ contains four types of terms: (I) A term $\sum_x \sum_{i=0}^8 s_{x,i}\phi(f_1, ..., f_M)$. This term represents the local evidence for an edge at each point and for its orientation. It is essentially the same term for local edge detection as discussed in the previous section where, if we ignore orientation so $s_x \in \{\pm 1\}$, then $\phi(f_1, ..., f_M) = \log \frac{P(f(I(x))|s=1)}{P(f(I(x))|s=0)}$. (II) A term $\sum_x (\sum_{i=0}^8 s_{x,i} - 1)^2$. This term is intended to ensure that only one cell is active at any spatial position. This corresponds to an inhibitory interaction between cells in the same hypercolumn. The cells in the hypercolumn give alternative, and inconsistent, interpretations of the input – hence only one of them can be correct. (III) A term that encourages edges to be continuous and for their directions to change smoothly. To define this term, we let $\boldsymbol{\theta}_i = (\cos\theta_i, \sin\theta_i)$ and $\boldsymbol{\theta}_i^T = (-\sin\theta_i, \cos\theta_i)$ denote the tangent to the edge and the normal. This term encourages there to be edges in the tangent direction, while the next term discourages them in the normal direction, see figure (27)(far right panel). This term is motivated by the intuition that curves are spatially smooth and can be justified by the statistics of natural images [38],[30]. We write it as $\sum_{x,y} \sum_{i,j=1}^8 W^T_{(x,\theta_i),(y,\theta_j)} s_{x,i} s_{y,j}$, where

$$W^T_{(x,\theta_i),(y,\theta_j)} = -\exp\{-|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j|/K_1\} \exp\{-|x-y|/K_2\} \exp\{-|\hat{xy} - \boldsymbol{\theta}_i|/K_3\}. \quad (41)$$

and ($\hat{xy}$ is the unit vector in direction $x - y$). This term encourages edges which are in similar directions (first term), nearby in position (second term), and where the edge orientation is similar to the difference $x - y$ between the two points. This term is excitatory. (IV) The final terms is inhibitory and discourages edges to be parallel to each other (if they are nearby). It is written as $\sum_{x,y} \sum_{i,j=1}^{8} W^N_{(x,\theta_i),(y,\theta_j)} s_{x,i} s_{y,j}$. Here

$$W^N_{(x,\theta_i),(y,\theta_j)} = \exp\{-|x-y|/K_4\} \exp\{-|\hat{xy} - \boldsymbol{\theta}_i^T|\} \tag{42}$$

The first term says this interaction decreases with distance. The second term discourages edges which are parallel to each other.

This gives an overall energy:

$$E(\boldsymbol{s}) = \sum_x \sum_{i=0}^{8} s_{x,i} \phi(f_1, ..., f_M) + \hat{K}_0 \sum_x (\sum_{i=0}^{8} s_{x,i} - 1)^2$$
$$+\hat{K}_1 \sum_{x,y} \sum_{i,j=1}^{8} W^T_{(x,\theta_i),(y,\theta_j)} s_{x,i} s_{y,j} + \hat{K}_2 + \sum_{x,y} \sum_{i,j=1}^{8} W^N_{(x,\theta_i),(y,\theta_j)} s_{x,i} s_{y,j}. \tag{43}$$

This yields a probability:

$$P(\boldsymbol{s}|\boldsymbol{f}) = \frac{1}{Z} \exp\{-E(\boldsymbol{s})\}.$$

This model can be implemented in neural networks by defining either stochastic or deterministic neural dynamics (i.e. either Gibbs sampling or mean field theory). The resulting update equations are more complex that those defined for our earlier examples but have the same basic ingredients. Models of this type can qualitatively account for associative field phenomena.

4.3 Stereo Models

This section introduces computational models for estimating depth by binocular stereo. The key problem is to solve the *correspondence problem* between the inputs in the two eyes to determine the *disparity*. Then the depth of the points in space can be estimated by trigonometry. (This pre-supposes that the eyes are *calibrated*, meaning that the distance between the eyes and the direction of gaze are known, which is beyond the scope of this chapter). As discussed in section (1), Julesz [71] showed that humans could perceive depth from stereo if the images consisted of random dot stereograms which minimize the effect of feature similarity cues, suggesting that human vision can solve this task by relying mainly on geometric regularities (assumed about the structure of the world). Other researchers [18] have studied human estimation of surface shape quantitatively and showed, among other things, bias towards fronto-parallel surfaces.

Most stereo algorithm address the correspondence problem by assuming that: (i) image features in the two eyes are more likely to correspond if they have similar appearance, (ii) the surface being viewed obeys prior knowledge such as being

piecewise smooth (e.g., like the weak membrane model). The first assumption depends on local properties of the images while the second assumption uses non-local context. In an earlier section we discussed how a population of Gabor filters could be used to match local image features. In this section we describe how context can be used to impose prior knowledge about the geometry of the scene. We will study classic models which assume that the surface is piecewise smooth. This leads to a markov field model which includes excitatory connections, imposing the geometric constraints, with inhibitory connections which prevent points from one eye having more than one match in the second eye. This yields an algorithm which involves cooperation, to implement the excitatory constraints, and competition to deal with the inhibitory constraints. This is consistent with findings from recent electrophysiological experiments [146],[145]. These complement earlier experiments [125] which tested the local stereo models described in section (2).

A Cooperative Stereo Model. We now specify a computational model for stereo which, for simplicity, we formulate in one-dimension. There is a long history of this type of model starting with the cooperative stereo algorithm [28,110] and current computer vision stereo algorithms are mostly designed on similar principles.

We specify the left and right images by $\boldsymbol{I}_L, \boldsymbol{I}_R$ and denote features extracted from them by $\boldsymbol{f}(\boldsymbol{I}_L) = \{f(x_L) : x_L \in \mathcal{D}_L\}$, $\boldsymbol{f}(\boldsymbol{I}_R) = \{f(x_R) : x_R \in \mathcal{D}_R\}$. We define a discrete-valued correspondence variable $V(x_L, x_R)$, so that $V(x_L, x_R) = 1$ means that the features at x_L, x_R in the two images correspond, and hence the disparity is $x_L - x_R$. If the features do not match then we set $V(x_L, x_R) = 0$. We encourage all data-points to match one, but allow some datapoints to be unmatched and others to match more than once (by paying a penalty).

We specify a distribution $P(\boldsymbol{V}|\boldsymbol{f}(\boldsymbol{I}_L), \boldsymbol{f}(\boldsymbol{I}_R)) = \frac{1}{Z}\exp\{-E(\boldsymbol{V}; \boldsymbol{f}(\boldsymbol{I}_L), \boldsymbol{f}(\boldsymbol{I}_R))/T\}$, where the energy $E(\boldsymbol{V}; \boldsymbol{f}(\boldsymbol{I}_L), \boldsymbol{f}(\boldsymbol{I}_R))$ is given by:

$$E(\boldsymbol{V}; \boldsymbol{f}(\boldsymbol{I}_L), \boldsymbol{f}(\boldsymbol{I}_R)) = \sum_{x_L, x_R} V(x_L, x_R) M(f(x_L), f(x_R))$$
$$+A\sum_{x_L}(\sum_{x_R} V(x_L, x_R) - 1)^2 + A\sum_{x_R}(\sum_{x_L} V(x_L, x_R) - 1)^2$$
$$+C \sum_{x_L, x_R} \sum_{y_L \in N(x_L)} \sum_{y_R \in N(x_R)} V(x_L, x_R) V(y_L, y_R)\{(x_R - x_L) - (y_R - y_L)\}^2 \quad (44)$$

The first term imposes that there are matches between image points with similar features, here $M(.,.)$ is a measure which takes small values if $f(x_L), f(x_R)$ are similar and large values if they are different. We will discuss at the end of this section how $M(f(x_L), f(x_R))$ relates to model for local stereo discussed earlier. The second two terms penalize image points which are either unmatched, or are matched more than once. The third term encourages the disparities, $x_L - x_R$, to be similar for neighboring points (here $N(.)$ defines a spatial neighborhood as before). These models can be applied to two-dimensional images by solving

the correspondence problem for each epipolar line separately (by maximizing $P(\boldsymbol{V}|\boldsymbol{f}(\boldsymbol{I}_L), \boldsymbol{f}(\boldsymbol{I}_R))$). This is shown in figure (30)(right panel). The parameter T is the variance of the model, as for the line process model, and has default value $T = 1$.

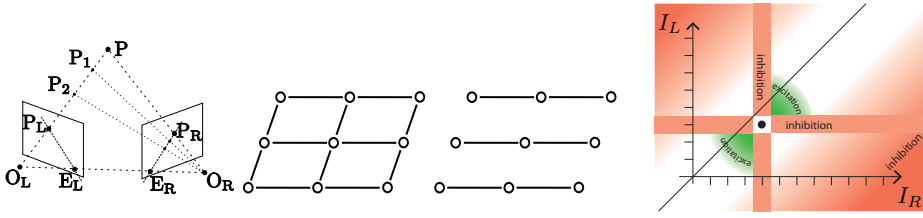

Fig. 30. Stereo. The geometry of stereo (left). A point P in 3-D space is projected onto points PL; PR in the left and right images. The projection is specified by the focal points OL,OR and the directions of gaze of the cameras (the camera geometry). The geometry of stereo enforces that points in the plane specified by P,OR OL, must be projected onto corresponding lines EL;ER in the two images (the epipolar line constraint). If we can find the correspondence between the points on epipolar lines then we can use trigonometry to estimate their depth, which is (roughly) inversely proportional to the disparity, which is the relative displacement of the two images. Right Panel: binocular stereo requires solving the correspondence problem which involves excitation (to encourage matches with similar depths/disparities) and inhibition (to prevent points from having multiple matches).

As for previous models, we can obtain a neural circuit model by performing mean field theory on $P(\boldsymbol{V}|\boldsymbol{f}(\boldsymbol{I}_L), \boldsymbol{f}(\boldsymbol{I}_R))$. (We restrict each point to have one or zero matches for this algorithm). This replaces $V(x_L, x_R) \in \{0, 1\}$ by continuous-valued $q(x_L, x_R) \in [0, 1]$ and an associated variable $u(x_L.x_R) = T \log \frac{q(x_L,x_R)}{1-q(x_L,x_R)}$ with $q(x_l, x_R) = \frac{1}{1+\exp\{-u(x_L,x_R)\}}$.

The update equation is:

$$\frac{du(x_L, x_R)}{dt} = -u(x_L, x_R) - M(f(x_L), f(x_R))$$
$$-2A(\sum_{y_R \neq x_R} q(x_L, y_R) - 1) - 2A(\sum_{y_L \neq x_L} q(y_L, x_R) - 1),$$
$$-2C \sum_{y_L \in N(x_L)} \sum_{y_R \in N(x_R)} q(y_L, y_R)\{(x_R - x_L) - (y_R - y_L)\}^2. \quad (45)$$

This update includes the standard integration term (first term) and the second term encourages matches where the features agree. There is also inhibition between competing matches (the third and fourth term), and excitation for matches which are consistent with a smooth surface (last term).

There is a variant of this algorithm which is used in interactive demo (4c). This algorithm is a discrete Hopfield network which attempts to minimize the

energy $E(\boldsymbol{V}; \boldsymbol{f}(\boldsymbol{I}_L), \boldsymbol{f}(\boldsymbol{I}_R))$ in equation (44). The algorithm starts by assigning initial values, 0 or 1, to each state variable $V(x_L, x_R)$. The algorithm proceeds by selecting a state variable, changing its value (e.g., changing $V(x_L, x_R) = 1$ to $V(x_L, x_R) = 0$), calculating if this change reduces the energy $E(\boldsymbol{V}; \boldsymbol{f}(\boldsymbol{I}_L), \boldsymbol{f}(\boldsymbol{I}_R))$, and keeping the change if it does. This process repeats until the algorithm converges (i.e., all possible changes raise the value of the energy).

How does the cooperative stereo algorithm relate to our earlier algorithm for computing stereo disparity locally? Recall that the algorithm estimated the disparity at a single point by having a set of neurons which were tuned to different disparities $\{D_i : i = 1, ..., N\}$, summing the votes $v(D_i)$ for each disparity by equation (20),and selecting the disparity with most votes. Using the cyclopean coordinate system [71], we express the disparity by $D(x) = \frac{1}{2}(x_R - x_L)$ where $x = \frac{1}{2}(x_R + x_L)$. At each point x we specify a population of neurons which encode the votes $v(D(x))$ for the different disparities. Then, instead of using winner-take-all to make a local decision, we feed the responses $v(D(x))$ back into cooperative stereo algorithm by defining $M(f(x_L), f(x_R)) = \exp\{-v(\frac{1}{2}(x_R - x_L))\}$ (the negative exponential $\exp\{-\}$ is required so the $M(f(x_L), f(x_R))$ is small if the vote for disparity $D(x) = \frac{1}{2}(x_R - x_L)$ is large).

Analysis of electrophysiological studies [146],[145] were in general agreement with the predictions of this type of stereo algorithm. In particular, studies showed that neural populations responses included excitation between cells tuned to similar disparities at neighboring spatial positions and inhibition between cells tuned to different disparities at the same position, see figure (31). In addition, Samonds *et al.* [147] implemented a variant of the stereo algorithm described above and showed that it could account for additional phenomena such as sharper tuning to the disparity for larger stimuli and performance on anti-correlated stimuli (where the left and right images have opposite polarity).

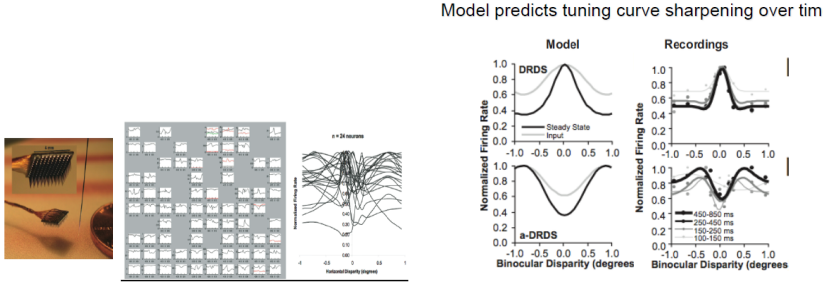

Fig. 31. Experiments for testing stereo algorithms [146],[145]. Left Panel: the experimental setup. Right Panel: the experiments give evidence for excitation between similar disparity and inhibition to prevent multiple matches.

4.4 Motion

Similar models have been applied to a range of motion phenomena. Early computational studies [163] showed that several perceptual phenomena of long-range motion could be described by a 'minimal mapping' theory that uses a slowness prior. Subsequent work showed that smoothness priors accounted for findings on short-range motion [61], including the surprising fact that an ellipse rotating in the image plane is perceived to move non-rigidly. Yuille and Grzywacz [186] qualitatively showed that a slow-and-smooth prior could account for a large range of motion perceptual phenomena – including motion capture and motion cooperation – both for short- and long-range motion. Weiss and his collaborators showed that slow [178] and slow-and-smooth priors [179]) could explain other short-range motion phenomena, such as how percepts can change dramatically as we alter the balance between the likelihood and prior terms (i.e. for some stimuli the prior dominates the likelihood and vice versa). All these models combine local estimates of the motion, such as those described in the previous section, with contextual cues implementing slow-and-smooth priors. They can be formulated using the same mathematical techniques. See `http://www.michaelbach.de/ot/mot-motionBinding/` to see how spatial context can be affected by other cues such as occlusion. It is also possible to perceive three-dimensional structure by observing a motion sequence (somewhat similar to binocular stereo) as can be seen in `http://michaelbach.de/ot/mot-ske/`.

The perception of motion can be strongly influenced by its past history and not merely by the change of image from frame to frame. For example, Anstis and Ramachandran [5] demonstrated perceptual phenomena where motion perception seems to require a temporal coherence prior in addition to the slow and smoothness priors described earlier in this section. Similarly, Watamaniuk *et al.* [176] demonstrated that humans could detect a coherently moving dot despite the presence of many incoherently moving dots. These classes of phenomena can be addressed by models which make prior assumptions about how motion changes over time. These can be performed [186] by adapting the Bayes-Kalman filter [72] [62] filter which gives an optimal way to combine information over time.

The task of the Bayes-Kalman filter is to estimate the state x_t of a system at time t dependent on a set of observations $y_t, ..., y_1$ (e.g., x_t could be the position of an airplane and y_t a noisy measurement of the airplane's position at time t). The model assumes a probability distribution $P(x_{t+1}|x_y)$ for how the state changes over time and a likelihood function $P(y_t|x_t)$ for the observation.

The task is to estimate the state x_t of a system at time t dependent on a set of observations $y_t, ..., y_1$ (e.g., x_t could be the position of an object and y_t a noisy measurement of the object position at time t). The model assumes a probability distribution $P(x_{t+1}|x_y)$ for how the state changes over time and a likelihood function $P(y_t|x_t)$ for the observation. This can be formulated by a Markov model, see figure (32)(left) where the observations $y_t, ..., y_1$ and states $x_t, ..., x_1$ are represented by the blue and red dots respectively (the lower and upper dots if viewed in black and white).

The purpose of Bayes-Kalman is to estimate the distribution $P(x_t|Y_t)$ of the state x_t conditioned on the measurements $Y_t = \{y_t, ..., y_1\}$ up to time t. It performs this by repeatedly performing the following two steps, which are called prediction and correction. The prediction uses the prior $P(x_{t+1}|x_t)$ to predict distribution $P(x_{t+1}|Y_t)$ of the state at $t+1$:

$$P(x_{t+1}|Y_t) = \int dx_t P(x_{t+1}|x_t) P(x_t|Y_t). \tag{46}$$

The correction step integrates the new observation y_{t+1} to estimate $P(x_{t+1}|Y_{t+1})$ by:

$$P(x_{t+1}|Y_{t+1}) = \frac{P(y_{t+1}|x_{t+1}) P(x_{t+1}|Y_t)}{P(y_{t+1}|Y_t)}. \tag{47}$$

Bayes-Kalman is initialized by setting $P(x_1|y_1) = P(y_1|x_1)P(x_1)/P(y_1)$ where $P(x_1)$ is the prior for the original position of the object at the start of the sequence. Then equations (46,47) are run repeatedly. The effect of prediction is to introduce uncertainly about the state x_t, while correction reduces uncertainty by providing a new measurement, see figure (32)(right).

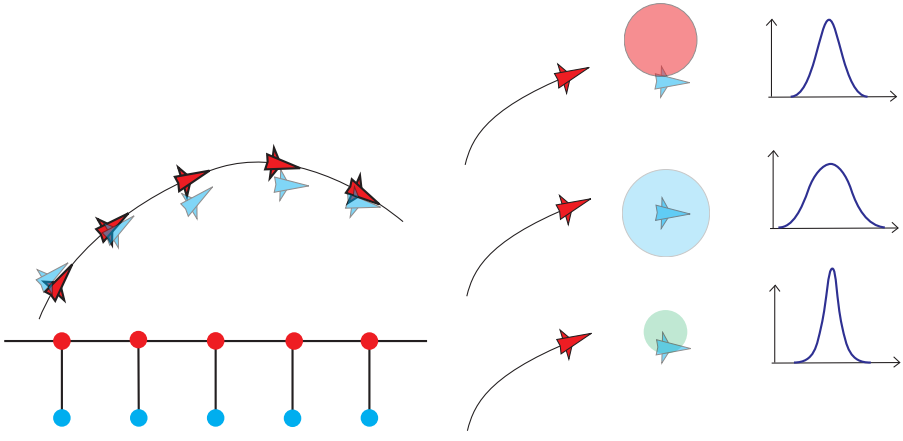

Fig. 32. Left Panel: Graph illustrating the unobserved states (red) and the observed states (blue) as a function of time. The airplanes true positions are shown in red and their observations (biased) are shown in blue. The Bayes-Kalman filter integrates observations to make estimate the true state using prior probabilities. Right Panel: Bayes-Kalman updates a probability distribution for the estimated position of the target. The variance of the distribution is illustrated by the one-dimensional figure (on the right) and the size of the circle (red, blue, or green). In the prediction stage (middle panel) the variance becomes large and after the measurement the variance becomes smaller.

4.5 Summary of Models with Context

This section illustrated how neural networks and Markov models could be used to apply context to visual tasks. We concentrated on edge detection, segmentation, and binocular stereo. We stressed how context can include excitatory and inhibitory interactions. And how inference can be performed using stochastic neurons (e.g., Gibbs sampling) or dynamic neural networks (e.g., mean field approximations). These models have some relations to psychophysics and electrophysiology. But we stress that detailed biological evidence in favor of these models remains preliminary due to the current limitations of experimental techniques. We note that current computer vision algorithms which address similar visual tasks are more complex although based on similar principles [11].

5 Cue Coupling

The ideas in this section are logical extensions of the ideas in the earlier sections. But we are now addressing more complex aspects of vision and so the techniques and the tools become more complex and more abstract as we begin to reason about surfaces, objects, and their relations. This section contains an interactive demo (5) for combining cues by weighted average.

5.1 Vision Modules and Cue Combination

At the behavioral level, psychophysicists have studied how humans combine different visual cues – such as shading, texture, binocular stereo, structure from motion – to estimate depth, surface geometry, and other surface properties. For example, quantifiable psychophysics experiments are broadly consistent with the predictions of the types of models discussed in the previous two sections– see [18,23] – but with some exceptions [160].

We now address how these cues can be combined. The most straightforward manner is to use a separate module for each cue to compute different estimates of the properties of interest, e.g., the surface geometry, and then merge these estimates into a single representation. This was proposed by Marr [109] who justified this strategy by invoking the principle of modular design. In particular, Marr proposed that surfaces should be represented by a *2 1/2D sketch* which specifies the shape of a surface by the distance of the surface points from the viewer. A related representation, *intrinsic images*, also represents surface shape together with the material properties of the surface. This strategy of making separate estimates for different cues and then combining them has also been followed by computer vision researchers.

How to combine cues computed independently? The natural way to combine these cues, from a probabilistic perspective, requires taking into account the uncertainty of the cues. This will be discussed in section (5.2) where we show that this can reduce to combining cues by taking weighted linear combinations, where the weights depend on the uncertainty. This strategy is able to account for some examples of cue combination [91,69,32].

Next we consider the problem of cue combination from a deeper probabilistic analysis, see section (5.3) [22]. Theoretical analysis suggests that we need to distinguish between situations where the cues are statistically independent of each other and the cases where they are not. We need also need to determine whether cues are using similar, and hence redundant, prior information. These considerations leads to a distinction between *weak* and *strong* coupling, where weak coupling corresponds to the traditional view of modules while strong coupling considers more complex interactions. To understand strong coupling it is helpful to consider the *causal factors* which generate the image. In addition, it involves the idea of *model selection*, called "competitive priors" in [185], which is a selection process that arises in two types of situations. Firstly, some image cues only apply to parts of the image and we need a process to select those parts (e.g., shape from texture cues are only valid within certain parts of an image). Secondly, for some images there may be several alternative ways to generate them, yielding several possible interpretations, and we need to select the most probable. Psychophysics experiments (e.g., [10,157]) investigate this by setting up situations where small changes in the image are sufficient to switch from one interpretation to another.

To describe causal models, and strong coupling, in more detail requires studying how images are formed in terms of the observer's viewpoint, the structure and material properties of objects in the world, and the lighting conditions. Then we proceed to two examples by Kording *et al.* and by Knill in section (5.4)

There is strong evidence that high-level recognition affects the estimation of three-dimensional shape (e.g., a rigidly rotating inverted face mask is perceived as non-rigidly deforming face, while most rigidly rotating objects are perceived to be rigid). But most visual cues are able to function even when images contain no objects. We briefly discuss this in section (6).

5.2 Combining Cues with Uncertainty

We first consider simple models which assume that the cues compute representations independently and then combine their outputs by taking linear weighted combinations. In some specific cases, these weights can be derived as measures of the uncertainly of the cues.

Combining cues by linear weighted sums. Suppose there are two cues for depth, or some other scene property, which separately give estimates $\boldsymbol{S}_1^*, \boldsymbol{S}_2^*$. One strategy to combine these cues is by linear weighted combination yielding a combined estimate $\boldsymbol{S}^*$:

$$\boldsymbol{S}^* = \omega_1 \boldsymbol{S}_1^* + \omega_2 \boldsymbol{S}_2^*, \tag{48}$$

where ω_1, ω_2 are positive weights such that $\omega_1 + \omega_2 = 1$.

Landy and Maloney [91] reviewed many early studies on cue combination and argued that they could be qualitatively explained by this type of model. They also discussed situations where the individual cues did not combine and "gating mechanisms" which require one cue to be switched off.

Special case where weights are derived from uncertainties. An important special case of this model is when the weights are measures of the uncertainty, or reliability, of the two cues. In this case, the linear weighted combination rule is optimal for the restricted case of Bayes estimation when there are no priors (we will discuss priors in the next section). It also yields detailed experimental predictions which have been successfully tested for some types of cue coupling, Jacobs [69], Ernst & Banks [32], although there are some interesting exceptions [21,48].

This special case associates each cue by an uncertainty σ_1^2, σ_2^2 and sets the weights to be $w_1 = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}$ and $w_2 = \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2)}$. It follows that $w_1/w_2 = \sigma_2^2/\sigma_1^2$, so the cue with lowest uncertainty will have a higher weight. For example, if $\sigma_2^2 >> \sigma_1$ then $w_1 >> w_2$. This special case give linear combination rule:

$$\boldsymbol{S}^* = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\boldsymbol{S}_1^* + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\boldsymbol{S}_2^*. \tag{49}$$

This combination rule can be shown to be optimal for the following conditions. Suppose the two cues are modeled by inputs $\{\boldsymbol{C}_i : i = 1, 2\}$ and outputs $\boldsymbol{S}$ which are related by conditional distributions $\{P(\boldsymbol{C}_i|\boldsymbol{S}) : i = 1, 2\}$. We assume that these cues are *conditionally independent* so that $P(\boldsymbol{C}_1, \boldsymbol{C}_2|S) = P(\boldsymbol{C}_1|\boldsymbol{S})P(\boldsymbol{C}_2|\boldsymbol{S})$ and both distributions are Gaussians:

$$P(\boldsymbol{C}_1|\boldsymbol{S}) = \frac{1}{Z_1}\exp\{-\frac{|\boldsymbol{C}_1-\boldsymbol{S}|^2}{2\sigma_1^2}\}, \quad P(\boldsymbol{C}_2|\boldsymbol{S}) = \frac{1}{Z_2}\exp\{-\frac{|\boldsymbol{C}_2-\boldsymbol{S}|^2}{2\sigma_2^2}\}. \tag{50}$$

In this case, the optimal estimates of the output $\boldsymbol{S}$, for each cue independently, are given by the maximum likelihood estimates:

$$\boldsymbol{S}_1^* = \arg\max_S P(\boldsymbol{C}_1|\boldsymbol{S}) = \boldsymbol{C}_1, \quad \boldsymbol{S}_2^* = \arg\max_S P(\boldsymbol{C}_2|\boldsymbol{S}) = \boldsymbol{C}_2 \tag{51}$$

But if both cues are available, then the optimal estimate is given by:

$$\boldsymbol{S}^* = \arg\max_{\boldsymbol{S}} P(\boldsymbol{C}_1, \boldsymbol{C}_2|\boldsymbol{S}) = \arg\max_{\boldsymbol{S}} P(\boldsymbol{C}_1|\boldsymbol{S})P(\boldsymbol{C}_2|\boldsymbol{S}) \tag{52}$$

$$= \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\boldsymbol{C}_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\boldsymbol{C}_2, \tag{53}$$

which reduces to the rule above using $\boldsymbol{S}_1^* = \boldsymbol{C}_1$ and $\boldsymbol{S}_2^* = \boldsymbol{C}_2$.

This shows that the optimal combination is a weighted linear sum of the two cues where the weights can be obtained from the uncertainties σ_1^2, σ_2^2 of the individual cues. This is the maximum likelihood estimate which, as discussed earlier, is the optimal Bayesian way to combine cues if there is no prior $P(\boldsymbol{S})$ (and the loss function is the default). Interactive demo (5) illustrates how the cues coupling result depends on the means and variances of each cue.

Extensions of combinations weighted by uncertainties. This section covers advanced material showing how the results in the previous section can be

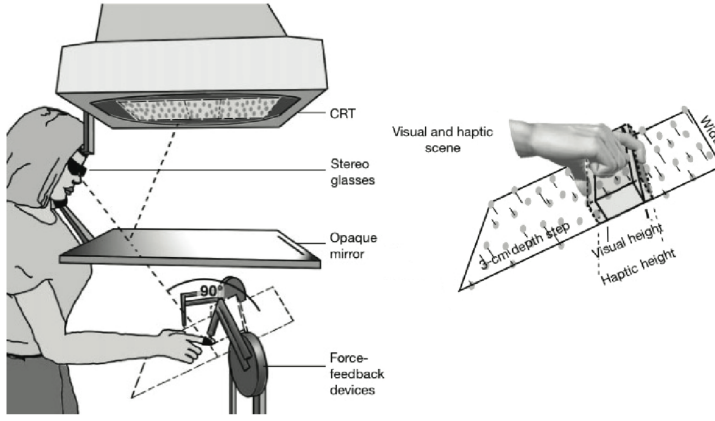

Fig. 33. The work of Ernst and Banks shows that cues are sometimes combined by weighted least squares where the weights depend on the variance of the cues. Figure reprinted with permission from [32].

extended, as approximations, to a much richer class of models. These extensions require that conditional independence holds, $P(\boldsymbol{C}_1, \boldsymbol{C}_2|\boldsymbol{S}) = P(\boldsymbol{C}_1|\boldsymbol{S})P(\boldsymbol{C}_2|\boldsymbol{S})$, and there is no prior. The best estimates for each cue individually are given $\boldsymbol{S}_1^* = \arg\max_{\boldsymbol{S}} \log P(\boldsymbol{C}_1|\boldsymbol{S})$ and $\boldsymbol{S}_2^* = \arg\max_{\boldsymbol{S}} \log P(\boldsymbol{C}_2|\boldsymbol{S})$. We can locally approximate $\log P(\boldsymbol{C}_1|\boldsymbol{S}) = \log P(\boldsymbol{C}_2|\boldsymbol{S}^*) + \frac{1}{2}(\boldsymbol{S} - \boldsymbol{S}_1^*)^T \frac{\partial^2 \log P(\boldsymbol{C}_1|\boldsymbol{S})}{\partial \boldsymbol{S} \partial \boldsymbol{S}}_{\boldsymbol{S}=\boldsymbol{S}_1^*} (\boldsymbol{S} - \boldsymbol{S}_1^*)$ (we use notation $_{\boldsymbol{S}=\boldsymbol{S}_1^*}$ to indicate where we are evaluating the derivative). This is a Taylor series approximation of $\log P(\boldsymbol{C}_1|\boldsymbol{S})$ about its maximum values $\boldsymbol{S}_1^*$. The first term in the expansion, $\frac{\partial \log P(\boldsymbol{C}_1|\boldsymbol{S})_{\boldsymbol{S}=\boldsymbol{S}_1^*}}{\partial \boldsymbol{S}}$ is zero because $\boldsymbol{S}_1^*$ is a maximum. The second term is the second order derivative of $\log P(\boldsymbol{C}_1|\boldsymbol{S})$ with respect to $\boldsymbol{S}$. This reduces to $-\frac{1}{\sigma_1^2}$ for the Gaussian case discussed above, hence the second order derivative evaluated at the maxima $\boldsymbol{S}_1^*$ gives an estimate of the "local variance" for the likelihood function . If we make this Taylor series approximation for both likelihood functions, then we can estimate an approximation solutions $\boldsymbol{S}_{app}^*$ when both cues are present:

$$\boldsymbol{S}^* = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}\{\Sigma_1^{-1}\boldsymbol{S}_1 + \Sigma_2^{-1}\boldsymbol{S}_2\}, \tag{54}$$

where Σ_1^{-1} is the matrix for $\frac{\partial^2 \log P(\boldsymbol{C}_1|\boldsymbol{S})}{\partial \boldsymbol{S} \partial \boldsymbol{S}}_{\boldsymbol{S}=\boldsymbol{S}_1^*}$ (and similarly for Σ_2^{-1}).

This Taylor series approximation is only valid if the results $\boldsymbol{S}^*$ is "close" to $\boldsymbol{S}_1^*$ and $\boldsymbol{S}_2^*$. How close depends on the form of the functions $\log P(\boldsymbol{C}_1|\boldsymbol{S}), \log P(\boldsymbol{C}_2|\boldsymbol{S})$ and, more specifically, how well they can be approximated near $\boldsymbol{S}_1^*, \boldsymbol{s}_2^*$ by quadratic functions. If these distributions are both Gaussians then their log-likelihoods are quadratic everywhere and so no approximation is required. In other situations, the nature of the cues determines how good an approximation we can obtain by this Taylor series expansion.

5.3 Bayesian Analysis: Weak and Strong Coupling

This section discusses more complex models for coupling cues. It approaches the problem from a Bayesian perspective [22][185]. This approach emphasizes that the uncertainties of the cues are taken into account and the statistical dependencies between the cue are made explicit. Simple examples of cue coupling, where the cues and independent, are called "weak coupling" in this framework. In some situations, if priors are not used, weak coupling reduces to the types of models studied in the previous section. By contrast, "strong coupling" is required if the cues are dependent on each other. Determining the dependencies between different cues requires sophisticated modeling to determine the causal factors underlying the cues. This also includes *model selection* where the cues allow several alternative interpretations of the image and small changes of the cues can dramatically change the percept.

The priors: avoiding double counting. As discussed in earlier sections, models of cues typically include prior probabilities about $\boldsymbol{S}$. For example, most cues for estimating shape or depth assume that the viewed scene is piecewise smooth, see the section on binocular stereo. More generally, vision is ambiguous in the sense that there are many different ways to generate the identical image and a visual system must determine which is the most probable way. Ambiguity is greatly reduced if the input is a sequence of images, but humans have little difficulty interpreting single images correctly, except if unusual conditions (or unless only small parts of the image are shown). But the bottom line, for cue integration, is that we typically need to consider the prior probabilities of $\boldsymbol{S}$ instead of neglecting them as we did in the previous section.

We now revisit the linear weighted sum rule from the perspective that cues usually require priors. To get intuition, suppose we have two cues for estimating the shape of a surface and suppose that both cues use the prior that the surface is spatially smooth. Simply taking a linear weighted sum of the cues will not be optimal, because the prior would be used twice. All priors introduce a bias to perception so we want to avoid doubling this bias. Experiments by Bülthoff and Mallot [18] suggest that priors are shared in this manner (i.e. no double counting) and so cues are not combined by linear weighted averaging. In this experiments, subjects were asked to estimate the orientation of surfaces using texture and shading cues. If only one cue was available, then the subjects tended to underestimate the surface orientation which is consistent with a prior bias towards smooth surfaces [185]. But if both cues were present, then subjects estimated the surface orientation more accurately, which is inconsistent with combining the cues by a linear weighted sum.

We first model the two cues separately by likelihoods $P(\boldsymbol{C}_1|\boldsymbol{S}), P(\boldsymbol{C}_2|\boldsymbol{S})$ and a prior $P(\boldsymbol{S})$. For simplicity we assume that the priors are the same for each cue, which is reasonable because they specify assumptions about the world independent of any observations. This gives posterior distributions for each visual

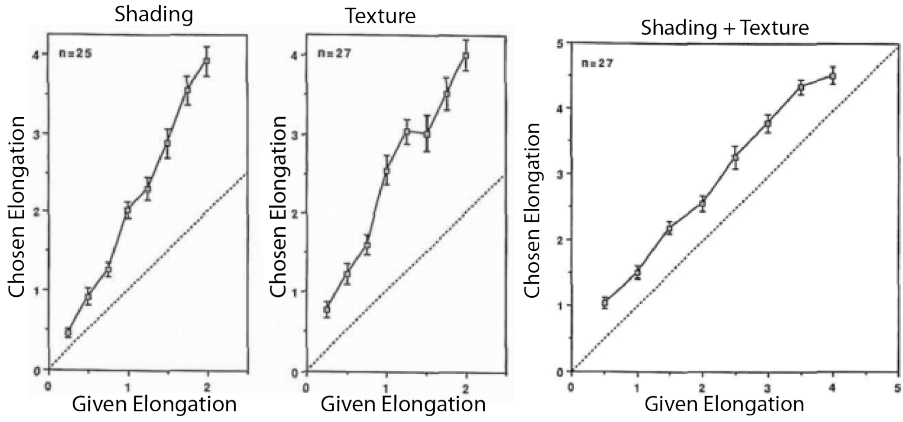

Fig. 34. Cue coupling results which are inconsistent with linear weighted average [17]. Left Panel: If depth is estimated using shading cues only then humans underestimate the perceived orientation (i.e. they see a flatter surface). Center Panel: Humans also underestimate the orientation if only texture cues are present. Right Panel: But if both shading and texture cues are available then humans perceive the orientation correctly. This is inconsistent with taking the linear weighted average of the results for each cue separately. Figure reprinted with permission from [17].

cue:

$$P(\boldsymbol{S}|\boldsymbol{C}_1) = \frac{P(\boldsymbol{C}_1|\boldsymbol{S})P(\boldsymbol{S})}{P(\boldsymbol{C}_1)}, \quad P(\boldsymbol{S}|\boldsymbol{C}_2) = \frac{P(\boldsymbol{C}_2|\boldsymbol{S})P(\boldsymbol{S})}{P(\boldsymbol{C}_2)}. \tag{55}$$

This yields estimates of surface shape to be $\boldsymbol{S}_1^* = \arg\max_{\boldsymbol{S}_1} P(\boldsymbol{S}|\boldsymbol{C}_1)$ and $\boldsymbol{S}_2^* = \arg\max_{\boldsymbol{S}_2} P(\boldsymbol{S}|\boldsymbol{C}_2)$. The optimal way to combine the cues is to estimate $\boldsymbol{S}$ from the posterior probability $P(\boldsymbol{S}|\boldsymbol{C}_1, \boldsymbol{C}_2)$:

$$P(\boldsymbol{S}|\boldsymbol{C}_1, \boldsymbol{C}_2) = \frac{P(\boldsymbol{C}_1, \boldsymbol{C}_2|\boldsymbol{S})P(\boldsymbol{S})}{P(\boldsymbol{C}_1, \boldsymbol{C}_2)}. \tag{56}$$

If the cues are *conditionally independent*, $P(\boldsymbol{C}|\boldsymbol{S}) = P(\boldsymbol{C}_1|\boldsymbol{S})P(\boldsymbol{C}_2)|\boldsymbol{S})$, then this simplifies to:

$$P(\boldsymbol{S}|\boldsymbol{C}_1, \boldsymbol{C}_2) = \frac{P(\boldsymbol{C}_1|\boldsymbol{S})P(\boldsymbol{C}_2|\boldsymbol{S})P(\boldsymbol{S})}{P(\boldsymbol{C}_1, \boldsymbol{C}_2)}. \tag{57}$$

Coupling the cues in equation (57) cannot correspond to a linear weighted sum (as in the previous section). This is because a linear weighted sum would essentially be using the prior twice, once for each cue. In general, additional cues provide extra information and interact nonlinearly. To understand this, suppose the prior is $P(\boldsymbol{S}) = \frac{1}{Z_p}\exp\{-\frac{|\boldsymbol{S}-\boldsymbol{S}_p|^2}{2\sigma_p^2}\}$. Then, setting $t_1 = 1/\sigma_1^2, t_2 = 1/\sigma_2^2, t_p = 1/\sigma_p^2$, the optimal combination is $\boldsymbol{S}^* = \frac{t_1\boldsymbol{C}_1 + t_2\boldsymbol{C}_2 + t_p\boldsymbol{S}_p}{t_1+t_2+t_p}$, hence the best estimate is a linear weighted combination of the two cues $\boldsymbol{C}_1, \boldsymbol{C}_2$ and the mean $\boldsymbol{S}_p$ of the prior. By contrast, the estimate using each cue individually are given by

$\boldsymbol{S}_1^* = \frac{t_1\boldsymbol{C}_1+t_p\boldsymbol{S}_p}{t_1+t_2+t_p}$ and $\boldsymbol{S}_2^* = \frac{t_2\boldsymbol{C}_2+t_p\boldsymbol{S}_p}{t_1+t_2+t_p}$. Hence any way to linearly combine the individual estimates $\boldsymbol{S}_1^*, \boldsymbol{S}_2^*$ will result in double counting the bias towards the mean $\boldsymbol{S}_p$ of the prior. (Note $t = 1/\sigma^2$ to make the connection to the previous notation).

To summarize, visual perception for some cues can be predicted by assuming that the cues involve a prior and this biases human perception. But if more cues are available, then perception will generally become less biased. Theories of optimal cue coupling should take this into account and prevent "double counting" the priors.

Cue Dependence and Causal Structure. The previous models in this chapter assumed that cues were conditionally independent, see equation (57). But, in practice cues are rarely independent. For many situations, such as the flying carpet example, the perception of depth is due to perspective, segmentation and shadows cues interacting in a complex way. The perspective and segmentation cues must determine that the beach is a flat ground plane. Segmentation cues must isolate the person, the towel, and the shadow. Then the visual system must decide that the shadow is cast by the towel and hence presumably must lie above the ground plane. These complex interactions are impossible to model using the simple conditional independent model described above.

The conditional independent model is also problematic even for the simple stimuli used by Bülthoff and Mallot [18] for studying shape from shading and texture. The model presupposes that it is possible to extract cues $\boldsymbol{C}_1, \boldsymbol{C}_2$ directly from the image $\mathbf{I}$ by a pre-processing step which computes $\boldsymbol{C}_1(\mathbf{I})$ and $\boldsymbol{C}_2(\mathbf{I})$. This requires decomposing the image $\mathbf{I}$ into texture and shading components. This may be possible, see figure (35), by exploiting the property that shading tends to be spatially smooth and texture is more jagged. In general, however, determining these cues from image is not straightforward and detailed modeling of it lies beyond the scope of this chapter.

Fig. 35. The input image (left) is decomposed into the base, or shading component, (center), and the detail (right).

The following set of experiments [79] suggest that visual perception does seek to find causal relations underlying the visual cues. In Kersten's "ball-in-a-box" experiments an observer perceives the ball to rise off the floor of the box only if this is consistent with a cast shadow, see figure (36). To solve this task, the

visual system must detect the surface and the orientation of the floor of the box (and decide it is flat), detect the ball, estimate the light source direction, and the motion of the shadow. It seems plausible that in this case, the visual system is unconsciously doing inverse inverse graphics to determine the most likely three-dimensional scene that generated the image sequence. There seems to be no way to explain how these geometric (perspective projection) and shading cues combine to give this percept using only the linear weighted sum rule. Another example, but for color perception, is given by [13].

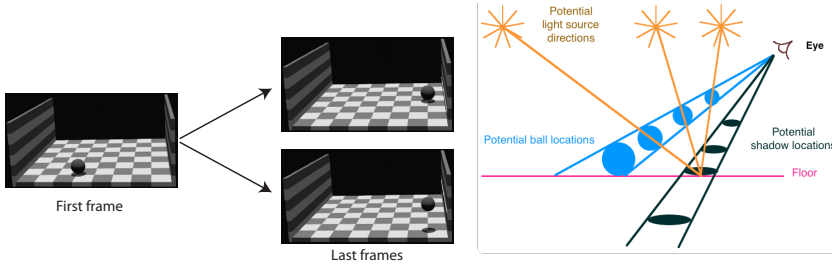

Fig. 36. In the "ball-in-a-box" experiments the motion of the shadow affects the perceived motion of the ball. The ball is perceived to rise from the ground if the shadow follows a horizontal trajectory in the image; but is perceived to move towards the back of the box if the shadow follows a diagonal trajectory. See `http://youtu.be/hdFCJepvJXU`. Left panel shows the first frame and the last frames for the two movies. Right panel. The explanation is that the observer resolves the ambiguities in the projection of a three-dimensional scene to perceive the 3D trajectory of the ball ([79].

Directed Graphical Models. Directed, or causal, graphical models [130] offer a mathematical language to describe these phenomena. These are similar to the "undirected" graphical models in section (4), because the graphical structure makes the conditional dependencies between variables explicit, but differ because the edges between nodes are directed. See [52] for an introduction to undirected and directed graphical models from a cognitive science perspective.

Formally *directed graphical models* are specified by follows. The random variables X_μ are defined at the nodes $\mu \in \mathcal{V}$ of a graph. The edges $\mathcal{E}$ specify which variables directly influence each other. For any node $\mu \in \mathcal{V}$, the set of parent nodes $pa(\mu)$ are the set of all nodes $\nu \in \mathcal{V}$ such that $(\mu, \nu) \in \mathcal{E}$, where (μ, ν) means that there is an edge between nodes μ and ν pointing to node μ. We denote the state of the parent node by $\boldsymbol{X}_{pa(\mu)}$. This gives a local *markov property* – the conditional distribution $P(X_\mu|\boldsymbol{X}_{/\mu}) = P(X_\mu|\boldsymbol{X}_{pa(\mu)})$, so the state of X_μ is only directly influenced by the state of its parents (note $\boldsymbol{X}_{/\mu}$ denotes the states of all nodes except for node μ). Then the full distribution for all the variables

can be expressed as:

$$P(\{X_\mu : \mu \in \mathcal{V}\}) = \prod_{\mu \in \mathcal{V}} P(X_\mu | \boldsymbol{X}_{pa(\mu)}). \quad (58)$$

We have already seen two examples of directed graphical models in this chapter. Firstly when we studied divisive normalization and used them in figure (25) to represent the dependencies between the stimuli, the filter responses, and the common factor. Secondly when described the Bayes-Kalman filter where the hidden state x_t at time t "causes" the hidden state x_{t+1} at time t and the observation y_t. Note that in some situations, the directions of the edges indicates physical causation between variables but in others the arrows merely represent statistical dependence. The relationship between graphical models and causality is complex and is clarified in [129].

These types of diagrams are helpful because they can be used to give a taxonomy between different ways that visual cues can be combined. For example, figure (37) shows two cases: (i) where two factors combine to cause an image, and (ii) where there is *common cause* of two cues.

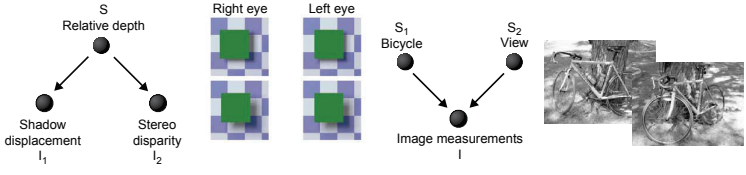

Fig. 37. Left Panel: An example of common cause. The shadow and binocular stereo cues are caused by the same event – two surfaces with one partially occluding the other. Right Panel: The image of the bicycle is caused by the pose of the bicycle, the viewpoint of the camera, and the lighting conditions. Figure reprinted with permission from [80].

Graphical models were used by Pearl [130] to illustrate the phenomena of *explaining away*. This describes how our interpretations of events can change suddenly as new information becomes available. For example, suppose a house alarm A can be activated by either a burglary B or by an earthquake E. This can be modeled by $P(A|B,E)$ and priors $P(B), P(E)$ for a burglary and an earthquake. In general, the prior probability of a burglary is much higher than the prior probability of an earthquake. So if an alarm goes off then it is much more probable to be caused by a burglary, formally $P(B|A) >> P(E|A)$. But suppose, after the alarm has sounded, you are worried about your house and check the internet only to discover that there has been an earthquake. In this case, this new information "explains away" the alarm and so you stop worrying about a burglary.

Variants of this phenomena arise in vision. In one version the "new information" is a more detailed analysis of the image. Suppose you see the "partly

occluded T" where a large part of the letter T is missing. In this case there is no obviously reason why part of the T is missing, so the perception may be only of two isolated segments. On the other hand, if there is a grey smudge over the missing part of the T then most observers perceive the T directly. The presence of the smudge "explains away" why part of the T is missing. This is a common phenomena which will return to in the next section from a different perspective.

Directed Graphical Models and Visual Tasks. Recall also that the human visual system performs a range of visual tasks and that the way cues are combined can depend on the tasks which are being performed. The language of directed graphical models is also useful for addressing this issue. For example, consider shape from shading whose goal is to determine the shape of a surface from its shading. But the shading depends both on the shape of the surface and on the light source direction. So an alternative task is to determine the light source direction. This can be formulated by a model $P(I|S,L)P(S), P(L)$ where I is the observed image, S is the surface shape, and L is the light source direction. $P(I|S,L)$ is the probability of generating an image I from shape S with lighting L, and $P(S), P(L)$ are prior probabilities on the surface shape and the lighting.

Suppose we only want to estimate the surface shape S. Then we do not care about the lighting L and the optimal procedure is to integrate it out to obtain a likelihood $P(I|S) = \int dL P(I|S,L)P(L)$ which is combined with a prior $P(S)$ to estimate S (If the loss function depends only on S then Bayesian decision theory says it should be integrated out, if it is a continuous variable, or summed out of it is discrete). Conversely, if we only want to estimate the lighting then we should integrate out the surface shape to obtain a likelihood $P(I|L) = \int dS P(I|S,L)P(S)$ and combine it with a prior $P(L)$. On the other hand, if we want to estimate both the surface shape and the lighting then should estimate them using the full model $P(I|S,L)$ with priors $P(S)$ and $P(L)$.

This issue of "integrating out" nuisance, or generic, variables relates to the *generic viewpoint assumption* [35]. This assumption states that the estimation of one variable, such as the surface shape, should be insensitive to small changes in another variable (e.g., the lighting). The intuition is that interpretations of the image that depend on something which is very unlikely, such as observing it from a special viewpoint, should be discouraged. It can be shown [35] that this can assumption can be formulated in terms of integrating out nuisance variables, but this analysis is beyond the scope of this chapter.

Model Selection. Certain types of cue coupling require *model selection*. While some cues, such as binocular stereo and motion, are usually valid in most places of the image other cues only apply to some images and often only for subparts of each image. For example, shape from shading is a well-studied cue which relates the intensity values to the changing geometry of a surface. But the lighting conditions in real world image are extremely complex and it is extremely difficult, for example, to distinguish between the image shading on a planar wall which

is due to a nearby light source from the shading pattern of a curved object illuminated from a light source that is further away. Another example, is shape from texture which assumes that there are homogeneous, or isotropic, texture painted on a three dimensional surface so that the projection of the texture onto the image plane gives a cue for the shape of the surface. But the texture on most surfaces rarely has this type of regular pattern, and hence shape from texture can only be applied to a limited class of surfaces, see figure 38. Similarly perspective cues give very strong perceptions of shape but they rely on the present straight lines in the image (and other Manhattan structures) so they valid for images of Manhattan but will not work in the jungle. These considerations show that cue combination often requires *model selection*, in order to determine in what parts of the image, if any, is the cue applicable.

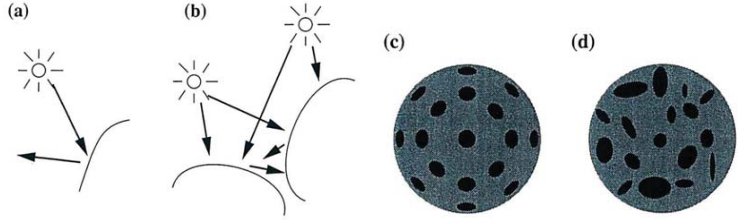

Fig. 38. Model selection may need to be applied in order to decide is a cue can be used. Shape from shading cues will work for case (a) because the shading pattern is simply due to a smooth convex surface illuminated by a single source. But for case (b) the shading pattern is complex – due to mutual reflection between the two surfaces – and so shape from shading cues will be almost impossible to use. Similarly, shape from texture is possible for case (c) because the surface contains a regular texture pattern but is much harder for case (d) because the texture is irregular.
Figure reprinted with permission from [189].

Model selection also arises in situations where there are several alternatives ways which could generate the image. In other words, there are several possible interpretations. This requires doing model selection to determine the model which best describes the data. This was called "competitive priors" by Yuille and Bülthoff [189]. By careful experimental design it is possible to adjust the image so that small changes shift the balance between one interpretation and another. Examples include the experiments where there are two rotating planes, which can be arranged to have two competing explanations [78]. By making slight variations to the transparency cues there are two surfaces which can be seen to either move rigidly together or to move independently, see figure (39)(right) and `http://youtu.be/gSrUBpovQdU`.

The work by Blake and Bülthoff [10] is a classic example where a sphere has a Lambertian (diffuse) reflection function and is viewed binocularly, see figure (38)(left). A specular component is adjusted so that it can lie in front of, between the center and the sphere, or at the center of the sphere. If the

specularity lies at the center then it is perceived as a light bulb and the sphere is perceived to be transparent. If the specularity is placed at the right position between the center and the sphere, then the sphere is perceived to be glossy. If the specularity lies in front of the sphere then it is seen as a cloud floating in front of a matte (diffuse, Lambertian sphere). Yuille and Bülthoff [189] interpreted both these phenomena as strong coupling with competitive priors.

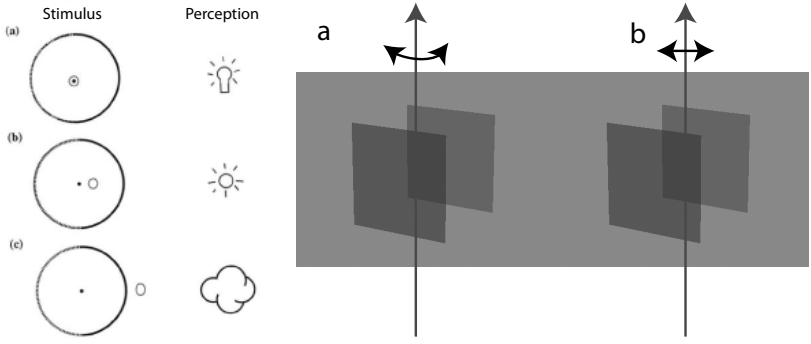

Fig. 39. Examples of strong coupling with competitive priors. A sphere is viewed binocularly (left) and small changes in the position of the specularity lead to very different percepts (Blake and Bülthoff 1990). Similarly altering the transparency of the moving surfaces (right) can make the two surfaces appear to rotate either rigidly together or independently.

We can also account for the "explaining away" examples in the previous section. We can consider two alternative models for the partial T. The first model is of two individual segments plus a smudge region. The second is a T which is partially hidden by a smudge. In this case the second model is more plausible since it would be very unlikely, an accidental viewpoint (or alignment), that the smudge happened to cover the missing part of the T, unless it really did occlude it. We can extend this argument to the Kanisza triangle, see figure (8). Humans perceive an illusory contour linking the three pac-men. But close inspection shows that there is no local edge evidence for the existence of the contour (except within the pac-men themselves). There are two possible interpretations of this figure. The first is that there are three partial circles which are arranged in a very special manner so that their edges are aligned (i.e. a straight line from one edge extend to another edge). The second is that there are three circles which are partially hidden by a triangular surface. In this second interpretation, the missing parts of the circles are *explained away* by the hypothesis that they are hidden by a triangle. Thus we have two possible models and we can assign probabilities to each then the phenomena can be explained provided the explanation in terms of the triangular surface is more probable.

Finally, we can revisit the flying carpet illusion shown in figure (3). Like Kersten's ball-in-a-box experiments it requires estimating the depth and orientation of the ground plane (i.e. the beach), segmenting and recognizing the woman, the towel she is standing on, and detecting the shadow. Then using the shadow cues, which requires making some assumption about the lighting, to estimate that the towel is hovering above the ground. This is a very complex way to combine all the cues in this image. Observe that it relies on the generic viewpoint assumption, in the sense that it is unlikely for there to be a shadow of that shape in that particular part of the image unless if was cast by some object. The real object that cast the shadow (the flag) is outside the image and so the visual system "attaches" the shadow to the towel which then implies that the towel must hover off the ground.

5.4 Examples of Strong Coupling

This section gives two examples of strong coupling. The first example deals with coupling different modalities while the second example concerns the perception of texture.

Multisensory Cue Coupling. The second example involves multisensory integration with structural uncertainty. Human observers are sensitive to both visual and auditory cues. Sometimes these cues have a common cause – e.g., you see a dog moving and hear it barking. In other situations the auditory and visual cues are due to different causes – e.g., a cat moves and a nearby dog barks (we ignore the possibility that the dog's barking is caused by the cat moving, or vice versa). Ventriloquists are able to fake these interactions by making the audience think that a puppet is speaking by associating the sound (produced by the ventriloquist) with the movement of the puppet. The Ventriloquism effect occurs when visual and auditory cues have different causes – and so are in conflict – but the audience perceive them as having the same cause.

Körding and his collaborators [89] developed an ideal observer model which determines whether two cues have a common cause or not. They formulated this using a meta-variable C, see figure (40). The common cause condition $C = 1$ means that the positions of the cues x_A, x_V are generated by the same process S, see figure (40)(left), by a distribution $P(x_A, x_V|S) = P(x_A|S)P(x_V|S)$. Here $P(x_A|S)$ and $P(x_V|S)$ are normal distributions $N(x_A|S, \sigma_A^2)$, $N(x_V|S, \sigma_V^2)$ – with the same mean S and variances σ_A^2, σ_V^2. It is assume that the visual cues are more precise than the auditory cues so that $\sigma_A^2 > \sigma_V^2$. The true position S is drawn from a probability distribution $P(S)$ which is assumed to be a normal distribution $N(0, \sigma_p^2)$. By contrast, $C = 2$ means that the cues are generated by two different processes S_A and S_B, in which case we have $P(x_A|S_A)$ and $P(x_V|S_V)$ which are both Gaussian $N(S_A, \sigma_A^2)$ and $N(S_V, \sigma_V^2)$, see figure (40)(right). We assume that S_A and S_V are independent samples from the normal distribution $N(0, \sigma_p^2)$. Note that this model involves model selection, between $C = 1$ and $C = 2$, and so in vision terminology is a form of strong coupling and competitive priors [185].

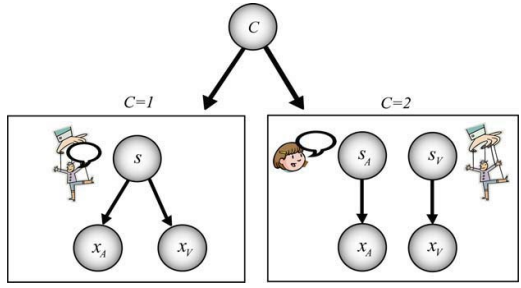

Fig. 40. The subject is asked to estimate the position of the cues and to judge whether the cues are from a common cause – i.e. at the same location – or not. In Bayesian terms the task of judging whether the cause is common can be formulated as model selection – are the auditory and visual cues more likely to generated from a single cause (left) or by two independent causes (right). Figure reprinted with permission from [89].

This model was compared to experiments where brief auditory and visual stimuli were presented simultaneously with varying amount of spatial disparity. Subjects were asked to identify the spatial location of the cue and/or whether they perceive a common cause [171]. The closer the visual stimulus was to the audio stimulus the more likely subjects perceived a common cause. In this case subjects' estimate of its position is strongly biased by the visual stimulus (because it is considered more precise with $\sigma_V^2 > \sigma_A^2$). But if subjects perceive distinct causes then their estimate is pushed away from the visual stimulus and exhibits *negative bias*. Körding *et al.* [89] argue that this bias is a selection bias stemming from restricting to trials in which causes are perceived as being distinct. For example, if the auditory stimulus is at the center and the visual stimulus at 5 degrees to right of center – then sometimes the (very noisy) auditory cue will be close to the visual cue and hence judged to have a common cause while on other cases the auditory cause will be further away (more than 5 degrees). Hence the auditory cue will have a truncated Gaussian (if judged to be distinct) and will yield negative bias.

More formally, the beliefs $P(C|x_A, x_V)$ in these two hypotheses $C = 1, 2$ are obtained by summing out the estimated positions s_A, s_B of the two cues as follows:

$$\begin{aligned} P(C|x_A, x_V) &= \frac{P(x_A, x_V|C)P(C)}{P(x_A, x_V)} \\ &= \frac{\int dS P(x_A|S)P(x_V|S)P(S)}{P(x_A, x_V)}, \quad \text{if } C = 1, \\ &= \frac{\int\int dS_A dS_V P(x_A|S_A)P(x_V|S_V)P(S_A)P(S_V)}{P(x_A, x_V)}, \quad \text{if } C = 2. \end{aligned} \tag{59}$$

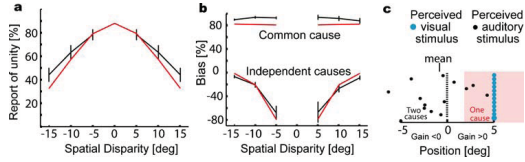

Fig. 41. Reports of causal inference.a) The relative frequency of subjects reporting one cause (black) is shown (reprinted with permission from [89]) with the prediction of the causal inference model (red). b) The bias, i.e. the influence of vision on the perceived auditory position is shown (gray and black). The predictions of the model are shown in red. c) A schematic illustration explaining the finding of negative biases. Blue and black dots represent the perceived visual and auditory stimuli, respectively. In the pink area people perceive a common cause.

There are two ways to combine the cues. The first is model selection. This estimates the most probable model $C^* = \arg\max P(C|x_V, x_A)$ from the input x_A, x_V and then uses this model to estimate the most likely positions s_A, s_V of the cues from the posterior distribution:

$$P(s_V, s_A) \approx P(s_V, s_A|x_V, x_A, C^*) = \frac{P(x_V, x_A|s_V, s_A, C^*)P(s_V, s_A|C^*)}{P(x_V, x_A|C^*)} \quad (60)$$

The second way to combine the cues is by *model averaging*. This does not commit itself to choosing C^* but instead averages over both models:

$$\begin{aligned} P(s_V, s_A|x_V, x_A) &= \sum_C P(s_V, s_A|x_V, x_A, C)P(C|x_V, x_A) \\ &= \sum_C \frac{P(x_V, x_A|s_V, s_A, C)P(s_V, s_A|C)P(C|x_V, x_A)}{P(x_V, x_A|C)}, \end{aligned} \quad (61)$$

where $P(C = 1|x_V, x_A) = \pi_C$ (the posterior mixing proportion).

Natarajan *et al.* [124] investigated these issues further. In particular, they showed that human performance on these types of experiments could be better modeled by replacing the Gaussian distributions by alternative distributions which are less sensitive to rare events. As people who modeled the stock market learnt to their cost in 2008, Gaussian distributions are non-robust because the tails of their distributions fall off rapidly which gives very low probability to rare events. Hence in many real-world applications distributions with longer tails are preferred. Following this reasoning Natarajan *et al.* assumed that the observations x_A, Xx_V were generated by distributions with longer tails. More precisely, they assumed that the data is distributed by a mixture of a Gaussian distribution (as in the models above) and a uniform distribution which yields longer tails. More formally, they assume x_A $\pi N(x_A : s_A, \sigma_A^2) + \frac{(1-\pi)}{r_l}$ and

x_V $\pi N(x_V : s_V, \sigma_V^2) + \frac{(1-\pi)}{r_1}$ where π is a mixing proportion and $U(x) = 1/r_1$ is a uniform distribution defined over the range r_1.

Homogeneous and Isotropic Texture. The second example is by Knill and concerns the estimating of orientation in depth (slant) from texture cues [82]. This relates to competitive priors because there are several alternative models for generating the image and the human observer must infer which is most likely, see figure 42. More formally, the data is generated by a mixture of models which enables non-linear cooperative interaction interactions between cues. In this example the data could be generated by isotropic homogeneous texture or by homogeneous texture only. Knill's finding is that human vision is biased to interpret image texture as isotropic but if enough data is available the system turns off the isotropy assumption and interprets texture using the homogeneity assumption only.

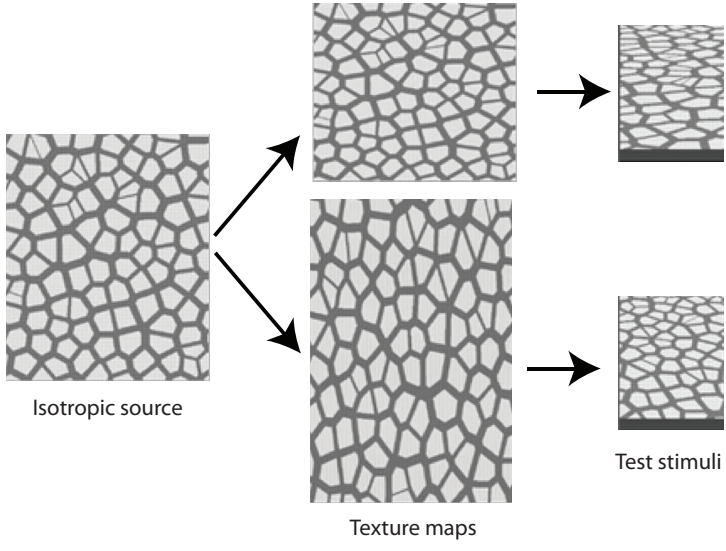

Fig. 42. Generating textures that violate isotropy [82]. An isotropic source image is either stretched (top middle) or compressed (bottom middle) producing texture maps that get applied to slanted surfaces shown on the right. A person that assumes surface textures are isotropic would overestimate the slant of the top stimulus and underestimate the slant of the bottom one. Figure adapted with permission from [82].

The posterior probability distribution for S is given by:

$$P(S|I) = \frac{P(I|S)P(S)}{P(I)}, \quad P(I|S) = \sum_{i=1}^{n} \phi_i P_i(I|S), \tag{62}$$

where ϕ_i is prior probability of model i, and $p_i(I|S)$ is corresponding likelihood function.

More specifically, texture features T can be generated by either an isotropic surface or a homogeneous surface. The surface is parameterized by tilt and slant σ, τ. Homogenous texture is described by two parameters α, θ and isotropic texture is a special case where $\alpha = 1$. This gives two likelihood models for generating the data:

$$P_h(T|(\sigma,\tau),\alpha,\theta), \quad P_i(T|(\sigma,\tau),\theta) \tag{63}$$

Here $P_i(T|(\sigma,\tau),\theta) = P_h(T|(\sigma,\tau),\alpha = 1,\theta)$.

Isotropic textures are a special case of homogenous textures (also rigid motion is a special class of non-rigid motion). The homogeneous model has more free parameters and hence has more flexibility to fit the data which suggests that human observers should always prefer it. But the Occam factor [108] means that this advantage will disappear if we put priors $P(\alpha)P(\theta)$ on the model parameters and integrate them out. This gives:

$$P_h(T|(\sigma,\tau)) = \int\int d\alpha d\theta P_h(T|(\sigma,\tau),\alpha,\theta), \quad P_i(T|(\sigma,\tau)) = \int d\theta P_h(T|(\sigma,\tau),\theta) \tag{64}$$

Integrating over the model priors smooths out the models. The more flexible model, P_h, has only a fixed amount of probability to cover a large range of data (e.g. all homogeneous textures) and hence has lower probability for any specific data (e.g. isotropic textures).

Knill describes how to combine these models using model averaging. The combined likelihood function is obtained by taking a weighted average:

$$P(T|(\sigma,\tau)) = p_h P_h(T|(\sigma,\tau)) + p_i P_i(T|(\sigma,\tau)), \tag{65}$$

Where (p_h, p_i) are prior probabilities that the texture is homogeneous or isotropic. We use a prior $P(\sigma,\tau)$ on the surface and finally achieve a posterior:

$$P(\sigma,\tau|I) = \frac{P(I|(\sigma,\tau))P(\sigma,\tau)}{P(I)}. \tag{66}$$

This model has a rich interpretation. If the data is consistent with an isotropic texture then this model dominates the likelihood and strongly influences the perception. Alternatively, if the data is consistent only with homogeneous texture then this model dominates. This gives a good fit to human performance [82].

6 Summary and the relations of early and high-level vision

This chapter has given a rapid tour of early vision. In particular, we have provided a modern perspective and conceptualization of early vision in terms of

probabilistic graphical models. In this final section we briefly mention how early vision relates to high level vision.

Marr's theory of vision [109] proposed that vision is done in a feedforward manner broken down into early vision, performed largely in visual areas V1 and V2 and high level vision performed in the Inferior Temporal (IT) lobe. In Marr's theory, processing is done in a feedforward manner. A second feedforward class of theories model the ventral stream (visual areas V1, V2, IT) by a hierarchical neural network where, as we ascend the hierarchy, the receptive fields of neurons are tuned to increasingly complex visual structures but are increasingly less sensitive to the precise positions of the input features. Models of this type, such as Hmax, have been developed in detail by Poggio and his collaborators (cf. [140]) and shown to correspond to many of the known aspects of the neuroscience and also shown to work well on some computer vision datasets. They concentrate on object detection and recognition and do not address tasks such as depth estimation or image parsing. They have limited representations and so it is unclear how they can address a large range of visual tasks [39].

By contrast, Mumford [120] argued for the importance of top-down processing citing the large number of backprojections in the cortex and pointing out that this matched the "analysis by synthesis" pattern theoretic approach proposed by Grenander [51],[156]. This class of theories has been developed by Mumford and Lee [92]. Related ideas are also discussed by Rao and Ballard [138] who suggest that top-down processing can be used to implement predictive processing somewhat similar to the Bayes-Kalman models briefly discussed in this chapter. Ullman and others [164],[31] have also proposed theories of vision which include bottom-up and top-down processing.

How do the models in this chapter fit into these three frameworks? There is clearly no problem in incorporating them into both classes of feedforward model. They can be used to compute the representations required by Marr's theory. They could also be used as the first stages of a feedforward network model like Hmax. The situation is more complicated for the third type of framework which combines bottom-up and top-down processing. But this can also be formulated by extending the graphical model theories in this chapter so that they are hierarchical. In these models the low-level nodes represent elementary features, such as edges, and the intermediate-level nodes represent compositions of the lower-level features, such as the grouping of edges to form longer segments, or the grouping of parallel line segments. These intermediate-level structures are combined together to form larger structures such as objects and object parts. These theories are sometimes called compositional [41,192] because they build objects by composition and they are closely related to stochastic grammars ([193,121]. For these classes of theories, the early and high levels of vision are strongly coupled (similar to strong coupling of cues in this chapter). Inference can be performed either bottom-up, where it is driven directly by the input image and low-level hypotheses are combined together to make hypotheses for more complex structures, or top-down where high-level hypotheses drive the computation.

References

1. L. Abbott and T. Kepler. Model neurons: From hodgkin-huxley to hopfield. *Statistical mechanics of neural networks*, pages 5–18, 1990.
2. D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for boltzmann machines*. *Cognitive science*, 9(1):147–169, 1985.
3. E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *JOSA A*, 2(2):284–299, 1985.
4. D. J. Amit. *Modeling brain function: The world of attractor neural networks*. Cambridge University Press, 1992.
5. S. Anstis and V. Ramachandran. Visual inertia in apparent motion. *Vision research*, 27(5):755–764, 1987.
6. J. Atick and A. Redlich. What does the retina know about natural scenes? *Neural Computation*, 4(2):196–210, 1992.
7. H. Barlow. Possible principles underlying the transformation of sensory messages. *Sensory Communication*, 1961.
8. H. Barlow. Measurements of the quantum efficiency of discrimination in human scotopic vision. *The Journal of Physiology*, 160(1):169–188, 1962.
9. R. Basri. Lambertian Reflectance and Linear Subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 218–233, 2003.
10. A. Blake and H. B. Bülthoff. Does the brain know the physics of specular reflection? *Nature*, 343(6254):165–168, 1990.
11. A. Blake, P. Kohli, and C. Rother. *Markov random fields for vision and image processing*. 2011.
12. A. Blake and A. Zisserman. *Visual Reconstruction (Artificial Intelligence Series)*. The MIT Press, Cambridge, MA, Mar. 2003.
13. M. G. Bloj, D. Kersten, and A. C. Hurlbert. Perception of three-dimensional shape influences colour perception through mutual illumination. *Nature*, 402(6764):877–879, Nov. 1999.
14. A. Borst and T. Euler. Seeing Things in Motion: Models, Circuits, and Mechanisms. *Neuron*, 71(6):974–994, Sept. 2011.
15. F. Briggs and W. Usrey. Corticogeniculate feedback and visual processing in the primate. *The Journal of Physiology*, 589(1):33–40, Jan. 2011.
16. G. Buckingham. Living in A Material World: How Visual Cues to Material Properties Affect the Way That We Lift Objects and Perceive Their Weight. *Journal of Neurophysiology*, 102(6):3111–3118, Dec. 2009.
17. H. Bülthoff, H. A. Mallot, B. T. Troscianko, et al. Integration of stereo, shading and texture. In *11th European Conference on Visual Perception*, pages 119–146. Wiley, 1990.
18. H. H. Bulthoff and H. A. Mallot. Integration of depth modules: stereo and shading. *Journal of the Optical Society of America A*, 5(10):1749–1758, Oct. 1988.
19. M. Carandini. Do We Know What the Early Visual System Does? *Journal of Neuroscience*, 25(46):10577–10597, Nov. 2005.
20. M. Carandini and D. J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, Nov. 2011.
21. K. Cheng, S. J. Shettleworth, J. Huttenlocher, and J. J. Rieser. Bayesian integration of spatial information. *Psychological Bulletin*, 133(4):625–637, July 2007.
22. J. J. Clark and A. L. Yuille. Data Fusion for Sensory Information Processing Systems . Springer, 1990.
23. B. Cumming, E. B. Johnston, and A. Parker. Effects of different texture cues on curved surfaces viewed stereoscopically. *Vision Research*, 33(5-6):827–838, Mar. 1993.

24. Y. Dan, J. Atick, and R. Reid. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *The Journal of Neuroscience*, 16(10):3351–3362, 1996.
25. J. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2:1160–1169, 1985.
26. P. Dayan and L. Abbott. *Theoretical Neuroscience*. The MIT Press, Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems, 2001.
27. J. B. Demb. Multiple mechanisms for contrast adaptation in the retina. *Neuron*, 36(5):781–783, Dec. 2002.
28. P. Dev. Perception of depth surfaces in random-dot stereograms: a neural model. *International Journal of Man-Machine Studies*, 7(4):511–528, 1975.
29. R. Durbin and G. Mitchison. A dimension reduction framework for understanding cortical maps. *Nature*, 343(6259):644–647, 1990.
30. J. H. Elder and R. M. Goldberg. Ecological statistics of Gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2(4):324–353, 2002.
31. B. Epshtein, I. Lifshitz, and S. Ullman. Image interpretation by a single bottom-up top-down cycle. *Proceedings of the National Academy of Sciences of the United States of America*, 105(38):14298, 2008.
32. M. O. Ernst and M. S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, Jan. 2002.
33. F. Fang, H. Boyaci, and D. Kersten. Border ownership selectivity in human early visual cortex and its modulation by attention. *Journal of Neuroscience*, 29(2):460–465, Jan. 2009.
34. D. J. Felleman and D. C. van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991.
35. W. Freeman. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368(6471):542–545, Apr. 1994.
36. D. Geiger and A. Yuille. A common framework for image segmentation. *International Journal of Computer Vision*, 6(3):227–243, 1991.
37. W. S. Geisler. Contributions of ideal observer theory to vision research. *Vision Research*, 51(7):771–781, 2011.
38. W. S. Geisler and J. S. Perry. Contour statistics in natural images: Grouping across occlusions. *Visual neuroscience*, 26(01):109–121, 2009.
39. S. Geman. Invariance and selectivity in the ventral visual pathway. *Journal of Physiology-Paris*, 2006.
40. S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.
41. S. Geman, D. Potter, and Z. Chi. Composition systems. *Quarterly of Applied Mathematics*, 60(4):707–736, 2002.
42. A. P. Georgopoulos, R. Caminiti, J. F. Kalaska, and J. T. Massey. Spatial coding of movement: a hypothesis concerning the coding of movement direction by motor cortical populations. *Exp Brain Res Suppl*, 7(32):336, 1983.
43. J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, 1986.
44. C. D. Gilbert and T. Wiesel. The influence of contextual stimuli on the orientation selectivity of cells in primary visual cortex of the cat. *Vision Research*, 30(11):1689–1701, 1990.
45. R. Glenn Northcutt and J. H. Kaas. The emergence and evolution of mammalian neocortex. *Trends in Neurosciences*, 18(9):373–379, 1995.

46. J. M. Gold, P. J. Mundy, and B. S. Tjan. The Perception of a Face Is No More Than the Sum of Its Parts. *Psychological Science*, Mar. 2012.
47. T. Gollisch and M. Meister. Eye Smarter than Scientists Believed: Neural Computations in Circuits of the Retina. *Neuron*, 65(2):150–164, Jan. 2010.
48. M. Gori, M. Del Viva, G. Sandini, and D. C. Burr. Young Children Do Not Integrate Visual and Haptic Form Information. *Current Biology*, 18(9):694–698, May 2008.
49. C. S. Green and D. Bavelier. Exercising your brain: A review of human brain plasticity and training-induced learning. *Psychology and Aging*, 23(4):692–701, 2008.
50. D. M. Green and J. A. Swets. *Signal detection theory and psychophysics*. New York, Wiley, 1966.
51. U. Grenander. *Pattern synthesis*. Appl. Math. Sci. Springer, New York, NY, 1976.
52. T. Griffiths and A. Yuille. A primer on probabilistic inference. *The probabilistic mind: Prospects for Bayesian cognitive science*, pages 33–57, 2008.
53. S. Grossberg. Some physiological and biochemical consequences of psychological postulates. *Proceedings of the National Academy of Sciences of the United States of America*, 60(3):758, 1968.
54. S. Grossberg and S. Hong. A neural model of surface perception: Lightness, anchoring, and filling-in. *Spatial Vision*, 19(2):263–321, Apr. 2006.
55. S. Grossberg and E. Mingolla. Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychological review*, 92(2):173, 1985.
56. S. Grossberg and E. Mingolla. Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentations. *Attention, Perception, & Psychophysics*, 38(2):141–171, 1985.
57. N. M. Grzywacz and A. Yuille. A model for the estimate of local image velocity by cells in the visual cortex. *Proceedings of the Royal Society of London. B. Biological Sciences*, 239(1295):129–161, 1990.
58. B. Hassenstein and W. Reichardt. 1956.
59. D. J. Heeger. Normalization of cell responses in cat striate cortex. *Visual neuroscience*, 9(02):181–197, 1992.
60. J. Hertz. *Introduction to the theory of neural computation*, volume 1. Basic Books, 1991.
61. E. C. Hildreth. The computation of the velocity field. *Proceedings of the Royal Society of London, Series B*, 221:189–220, 1984.
62. Y. Ho and R. Lee. A Bayesian approach to problems in stochastic estimation and control. *Automatic Control, IEEE Transactions on*, 9(4):333–339, 1964.
63. J. Hopfield and D. Tank. Computing with neural circuits: a model. *Science (New York, NY)*, 233(4764):625–633, 1986.
64. D. Hubel. Evolution of ideas on the primary visual cortex, 1955–1978: A biased historical account. *Bioscience reports*, 2(7):435–469, 1982.
65. D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.
66. D. Hubel and T. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, 1968.
67. A. Hyvärinen. Statistical Models of Natural Images and Cortical Visual Representation. *Topics in Cognitive Science*, 2(2):251–264, Apr. 2010.
68. L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

69. R. Jacobs. Optimal integration of texture and motion cues to depth. *Vision Research*, 39(21):3621–3629, Oct. 1999.
70. J. Jones and L. Palmer. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1187–1211, 1987.
71. B. Julesz. *Foundations of Cyclopean perception*. The University of Chicago Press, Chicago, London, 1971. Anaglyphoscope in pocket.
72. R. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.
73. G. Kanizsa. *Organization in Vision*. Praeger Press, 1979.
74. N. Kanwisher. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences of the United States of America*, 107(25):11163, 2010.
75. M. Kapadia, G. Westheimer, and C. D. Gilbert. Spatial distribution of contextual interactions in primary visual cortex and in visual perception. *Journal of Neurophysiology*, 84(4):2048–2062, 2000.
76. P. J. Kellman and M. E. Arterberry. The Cradle of Knowledge: Development of Perception in Infancy - Philip J. Kellman, Martha E. Arterberry - Google Books. The MIT Press, Cambridge, MA, 2000.
77. D. Kersten. Predictability and redundancy of natural images. *JOSA A*, 4(12):2395–2400, 1987.
78. D. Kersten, H. H. Bulthoff, B. Schwartz, and K. Kurtz. Interaction between transparency and structure from motion. *Neural Computation*, 4(4):573–589, 1992.
79. D. Kersten, P. Mamassian, and D. C. Knill. Moving cast shadows induce apparent motion in depth. *PERCEPTION-LONDON-*, 26(2):171–192, 1997.
80. D. Kersten and A. Yuille. Bayesian models of object perception. *Current Opinion in Neurobiology*, 13(2):150–158, 2003.
81. M. Kinoshita, C. D. Gilbert, and A. Das. Optical Imaging of Contextual Interactions in V1 of the Behaving Monkey. *Journal of Neurophysiology*, 102(3):1930–1944, Sept. 2009.
82. D. Knill. Mixture models and the probabilistic structure of depth cues. *Vision Research*, 43(7):831–854, Mar. 2003.
83. D. C. Knill. Contour into texture: information content of surface contours and texture flow. *Journal of the Optical Society of America A*, 18(1):12–35, 2001.
84. D. C. Knill and D. J. Kersten. Apparent surface curvature affects lightness perception. *Nature*, 351(228-230), 1991.
85. C. Koch, J. Marroquin, and A. Yuille. Analog "neuronal" networks in early vision. *Proceedings of the National Academy of Sciences of the United States of America*, 83(12):4263–4267, June 1986.
86. H. Komatsu. The neural mechanisms of perceptual filling-in. *Nature Reviews Neuroscience*, 7(3):220–231, Mar. 2006.
87. S. Konishi, A. Yuille, J. Coughlan, and S. Zhu. Statistical edge detection: Learning and evaluating edge cues. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(1):57–74, 2003.
88. S. Konishi and A. L. Yuille. Statistical cues for domain specific image segmentation with performance analysis. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 125–132. IEEE, 2000.
89. K. P. Körding, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams. Causal Inference in Multisensory Perception. *PLoS ONE*, 2(9):e943, Sept. 2007.
90. V. A. Lamme. The neurophysiology of figure-ground segregation in primary visual cortex. *The Journal of Neuroscience*, 15(2):1605–1615, 1995.

91. M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young. Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, 35:389–412, 1995.
92. T. Lee and D. Mumford. Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A, Optics, Image Science, and Vision*, 20(7):1434–1448, July 2003.
93. T. Lee and M. Nguyen. Dynamics of subjective contour formation in the early visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 98(4):1907, 2001.
94. T. S. Lee. Image representation using 2d gabor wavelets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(10):959–971, 1996.
95. T. S. Lee, D. Mumford, and A. Yuille. Texture segmentation by minimizing vector-valued energy functionals: The coupled-membrane model. In *Computer VisionECCV'92*, pages 165–173. Springer, 1992.
96. T. S. Lee and A. L. Yuille. Efficient coding of visual scenes by grouping and segmentation: theoretical predictions and biological evidence. In K. Doya, S. Ishii, A. Pouget, and R. P. Rao, editors, *Bayesian Brain: Probabilistic Approaches to Neural Coding*, pages 1–29. Dec. 2006.
97. P. Lennie. Single units and visual cortical organization. *Perception*, 27:889–935, 1998.
98. P. Lennie. Single units and visual cortical organization. *Perception*, 27:889–936, 1998.
99. J. Y. Lettvin, H. R. Maturana, W. S. McCulloch, and W. H. Pitts. What the frog's eye tells the frog's brain. *Proceedings of the IRE*, 47(11):1940–1951, 1959.
100. R. Linsker. From basic network principles to neural architecture: Emergence of orientation-selective cells. *Proceedings of the National Academy of Sciences of the United States of America*, 83(21):8390, 1986.
101. R. Linsker. From basic network principles to neural architecture: Emergence of spatial-opponent cells. *Proceedings of the National Academy of Sciences of the United States of America*, 83(19):7508, 1986.
102. J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.
103. M. Livingstone and D. Hubel. Anatomy and physiology of a color system in the primate visual cortex. *The Journal of Neuroscience*, 4(1):309–356, 1984.
104. N. Logothetis and D. L. Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 19(1):577–621, 1996.
105. H. Lu and A. Roe. Optical Imaging of Contrast Response in Macaque Monkey V1 and V2. *Cerebral Cortex*, 17(11):2675–2695, Jan. 2007.
106. W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, Nov. 2006.
107. W. Maass. On the computational power of winner-take-all. *Neural computation*, 12(11):2519–2535, 2000.
108. D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, first edition edition, Oct. 2003.
109. D. Marr. Vision. *W.H. Freeman*, 1982.
110. D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194(4262):283–287, 1976.
111. R. H. Masland and P. R. Martin. The unsolved mystery of vision. *Current biology : CB*, 17(15):R577–82, Aug. 2007.
112. J. T. McIlwain. Distributed spatial coding in the superior colliculus: a review. *Visual Neuroscience*, 6(01):3–13, 1991.

113. F. Mechler and D. L. Ringach. On the classification of simple and complex cells. *Vision Research*, 42(8):1017–1033, 2002.
114. M. Meister and M. Berry. The neural code of the retina. *Neuron*, 22:435–450, 1999.
115. B. W. Mel, D. L. Ruderman, and K. A. Archie. Translation-invariant orientation tuning in visual complex cells could derive from intradendritic computations. *The Journal of Neuroscience*, 18(11):4325–4334, 1998.
116. W. Merigan and J. Maunsell. How parallel are the primate visual pathways? *Annual Review of Neuroscience*, 16(1):369–402, 1993.
117. D. Milner and M. Goodale. *The Visual Brain in Action (Oxford Psychology Series)*. Oxford University Press, USA, 2 edition, Dec. 2006.
118. M. C. Morrone and D. Burr. Feature detection in human vision: A phase-dependent energy model. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, pages 221–245, 1988.
119. R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 891–898. IEEE, 2014.
120. D. Mumford. On the computational architecture of the neocortex. *Biological Cybernetics*, 66(3):241–251, 1992.
121. D. Mumford and A. Desolneux. *Pattern Theory: The Stochastic Analysis of Real-World Signals (Applying Mathematics)*. A K Peters Ltd, Aug. 2010.
122. D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 42(5):577–685, 1989.
123. J. J. Nassi, D. C. Lyon, and E. M. Callaway. The Parvocellular LGN Provides a Robust Disynaptic Input to the Visual Motion Area MT. *Neuron*, 50(2):319–327, Apr. 2006.
124. R. Natarajan, I. Murray, L. Shams, and R. Zemel. Characterizing response behavior in multisensory perception with conflicting cues. In *Advances in neural information processing systems NIPS*, 2008.
125. I. Ohzawa, G. C. Deangelis, and R. D. Freeman. Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science*, 249(4972):1037–1041, 1990.
126. E. Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, Aug. 1982.
127. B. A. Olshausen. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.
128. G. Papandreou, L.-C. Chen, and A. L. Yuille. Modeling image patches with a generic dictionary of mini-epitomes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2059–2066. IEEE, 2014.
129. J. Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press.
130. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1 edition, Sept. 1988.
131. D. Pelli. The quantum efficiency of vision. In C. Blakemore, editor, *Vision: Coding and Efficiency*, pages 3–24. Vision: Coding and efficiency, Cambride, 1990.
132. E. Peterhans and R. von der Heydt. Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps. *The Journal of Neuroscience*, 9(5):1749–1763, 1989.
133. D. A. Pollen and S. F. Ronner. Phase relationships between adjacent simple cells in the visual cortex. *Science*, 212(4501):1409–1411, 1981.

134. A. Pouget, P. Dayan, and R. S. Zemel. Inference and computation with population codes. *Annual Review of Neuroscience*, 26(1):381–410, 2003.
135. N. Qian. Computing stereo disparity and motion with known binocular cell properties. *Neural Computation*, 6(3):390–404, 1994.
136. C. Qiu, D. Kersten, and C. A. Olman. Segmentation decreases the magnitude of the tilt illusion. *Journal of Vision*, 13(13):19–19, Nov. 2013.
137. C. A. Ramachandra and B. W. Mel. Computing local edge probability in natural scenes from a population of oriented simple cells. *Journal of vision*, 13(14):19, 2013.
138. R. P. N. Rao and D. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87, 1999.
139. F. Rieke, D. Warland, R. R. de Ruytervan Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code*. MIT Press, 1997. Paper back edition 1999.
140. M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999.
141. A. W. Roe, L. Chelazzi, C. E. Connor, B. R. Conway, I. Fujita, A. L. Gallant, H. Lu, and W. Vanduffel. Towards a unified theory of visual area v4. *Neuron*, 74, April 2012.
142. A. W. Roe, L. Chelazzi, C. E. Connor, B. R. Conway, I. Fujita, J. L. Gallant, H. Lu, and W. Vanduffel. Toward a Unified Theory of Visual Area V4. *Neuron*, 74(1):12–29, Apr. 2012.
143. A. W. Roe, G. Chen, and H. D. Lu. Functional architecture of area v2. *Encyclyopedia of Neuroscience*, 10:331–349, 2009.
144. N. C. Rust and J. J. DiCarlo. Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *Journal of Neuroscience*, 30(39):12978–12995, Sept. 2010.
145. J. M. Samonds, B. Potetz, and T. S. Lee. Relative luminance and binocular disparity preferences are correlated in macaque v1, matching natural scene statistics. *Proc Nat Acad Sci USA (PNAS)*, 109(16):6313–6318, April 2012.
146. J. M. Samonds, B. R. Potetz, and T. Lee. Cooperative and Competitive Interactions Facilitate Stereo Computations in Macaque Primary Visual Cortex. *Journal of Neuroscience*, 29(50):15780–15795, Dec. 2009.
147. J. M. Samonds, B. R. Potetz, C. W. Tyler, and T. S. Lee. Recurrent connectivity can account for the dynamics of disparity processing in v1. *The Journal of Neuroscience*, 33(7):2934–2946, 2013.
148. Y. Sasaki, T. Watanabe, and D. Purves. The Primary Visual Cortex Fills in Color. *Proceedings of the National Academy of Sciences of the United States of America*, 101(52):18251–18256, Dec. 2004.
149. T. Schenk and R. D. McIntosh. Do we have independent visual streams for perception and action? *Cognitive Neuroscience*, 1(1):52–62, Feb. 2010.
150. P. R. Schrater, D. C. Knill, and E. P. Simoncelli. Mechanisms of visual motion detection. *Nature neuroscience*, 3(1):64–68, 2000.
151. E. L. Schwartz. Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding. *Vision Research*, 20(8):645–669, 1980.
152. O. Schwartz, T. J. Sejnowski, and P. Dayan. Perceptual organization in the tilt illusion. *Journal of Vision*, 9(4):19.1–20, 2009.
153. S. M. Sherman and R. W. Guillery. The role of the thalamus in the flow of information to the cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 357(1428):1695–1708, Dec. 2002.

154. S. Shushruth, L. Nurminen, M. Bijanzadeh, J. M. Ichida, S. Vanni, and A. Angelucci. Different Orientation Tuning of Near- and Far-Surround Suppression in Macaque Primary Visual Cortex Mirrors Their Tuning in Human Perception. *Journal of Neuroscience*, 33(1):106–119, Jan. 2013.
155. E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.
156. Springer. *Pattern analysis: lectures in pattern theory, volume II*, New York, 1978. Springer.
157. R. Sundareswara and P. Schrater. Perceptual multistability predicted by search model for Bayesian decisions. *Journal of Vision*, 8(5):12–12, May 2008.
158. V. Talebi and C. L. Baker. Natural versus synthetic stimuli for estimating receptive field models: a comparison of predictive robustness. *The Journal of Neuroscience*, 32(5):1560–1576, 2012.
159. B. S. Tjan, W. L. Braje, G. E. Legge, and D. Kersten. Human efficiency for recognizing 3-D objects in luminance noise. *Vision Research*, 35(21):3053–3069, 1995.
160. J. T. Todd, A. H. Oomes, J. J. Koenderink, and A. M. Kappers. On the affine structure of perceptual space. *Psychological Science*, 12(3):191–196, 2001.
161. E. J. Trenti, J. F. Barraza, and M. P. Eckstein. Learning motion: Human vs. optimal Bayesian learner. *Vision Research*, 50(4):460–472, Feb. 2010.
162. D. Y. Tsao, S. Moeller, and W. A. Freiwald. Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences*, 105(49):19514–19519, Dec. 2008.
163. S. Ullman. *The interpretation of Visual Motion*. MIT Press, Cambridge, MA, 1979.
164. S. Ullman. Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex*, 5(1):1–11, 1995.
165. D. Van Essen, C. H. Anderson, and D. Felleman. Information processing in the primate visual system: an integrated systems perspective. *Science*, 255(5043):419–423, Jan. 1992.
166. J. H. van Hateren and D. Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1412):2315–2320, 1998.
167. R. von der Heydt. Image parsing mechanisms of the visual cortex. *The visual neurosciences*, pages 1139–1150, 2003.
168. R. von der Heydt and E. Peterhans. Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *The Journal of Neuroscience*, 9(5):1731–1748, 1989.
169. R. von der Heydt, E. Peterhans, and G. Baumgartner. Illusory contours and cortical neuron responses. *Science*, 224(4654):1260, 1984.
170. M. J. Wainwright and E. Simoncelli. Scale mixtures of Gaussians and the statistics of natural images. *Advances in neural information processing systems*, 12(1):855–861, 2000.
171. M. T. Wallace, G. E. Roberson, W. D. Hairston, B. E. Stein, J. W. Vaughan, and J. A. Schirillo. Unifying multisensory signals across time and space. *Experimental Brain Research*, 158(2), Apr. 2004.
172. P. Wallisch and J. A. Movshon. Structure and Function Come Unglued in the Visual Cortex. *Neuron*, 60(2):194–197, Oct. 2008.
173. B. A. Wandell, S. O. Dumoulin, and A. A. Brewer. Visual Field Maps in Human Cortex. *Neuron*, 56(2):366–383, Oct. 2007.

174. B. A. Wandell and J. Winawer. Imaging retinotopic maps in the human brain. *Vision Research*, 51(7):718–737, Apr. 2011.
175. D. K. Warland, P. Reinagel, and M. Meister. Decoding visual information from a population of retinal ganglion cells. *Journal of Neurophysiology*, 78(5):2336–2350, 1997.
176. S. N. Watamaniuk, S. P. McKee, and N. M. Grzywacz. Detecting a trajectory embedded in random-direction motion noise. *Vision research*, 35(1):65–77, 1995.
177. A. B. Watson, H. Barlow, and J. G. Robson. What does the eye see best? *Nature*, 302(5907):419–422, Apr. 1983.
178. Y. Weiss and E. H. Adelson. Slow and smooth: A bayesian theory for the combination of local motion signals in human vision. Technical Report 1624, Massachusetts Institute of Technology, 1998.
179. Y. Weiss, E. P. Simoncelli, and E. H. Adelson. Motion illusions as optimal percepts. *Nature Neuroscience*, 5:598–604, 2002.
180. H. R. Wilson and J. D. Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24, 1972.
181. R. Young, R. Lesperance, and W. Meyer. The Gaussian derivative model for spatial-temporal vision: I. Cortical model. *Spatial Vision*, 3(4):261–319, 2001.
182. A. Yuille. Energy functions for early vision and analog networks. *Biological Cybernetics*, 61:115–123, 1987.
183. A. Yuille. Efficient coding of visual scenes by grouping and segmentation. *Bayesian Brain: Probabilistic Approaches to Neural Coding, MIT Press, Cambridge, MA*, pages 145–188, 2006.
184. A. Yuille. Loopy belief propagation, mean-field, and Bethe approximations. *Markov Random Fields for Vision and Image Processing*, 2011.
185. A. Yuille and H. H. Bulthoff. Bayesian decision theory and psychophysics. In D. Knill and W. Richards, editors, *Perception as Bayesian Inference*, page 123. Cambridge University Press, 1996.
186. A. Yuille, P. Y. Burgi, and N. M. Grzywacz. Visual motion estimation and prediction: A probabilistic network model for temporal coherence. *Computer Vision, 1998. Sixth International Conference on*, pages 973–978, 1998.
187. A. Yuille, D. Kammen, and D. Cohen. Quadrature and the development of orientation selective cortical cells by Hebb rules. *Biological Cybernetics*, 61(3):183–194, 1989.
188. A. Yuille and D. Kersten. Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308, July 2006.
189. A. L. Yuille and H. H. Bülthoff. Bayesian decision theory and psychophysics. In D. Knill and W. Richards., editors, *Bayesian Approaches to Perception*. Cambridge University Press, 1996.
190. L. Zhaoping. *Understanding vision: theory, models, and data*. Oxford University Press, 2014.
191. H. Zhou, H. Friedman, and R. von der Heydt. Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience*, 20(17):6594–6611, 2000.
192. L. Zhu, Y. Chen, and A. Yuille. Recursive Compositional Models for Vision: Description and Review of Recent Work. *Journal of Mathematical Imaging and Vision*, 41(1-2):122–146, Apr. 2011.
193. S.-c. Zhu and D. Mumford. A Stochastic Grammar of Images. *Foundations and Trends® in Computer Graphics and Vision*, 2(4):259–362, 2006.