Probabilities on Graphs: Directed and Undirected

Alan Yuille and Dan Kersten

May 6, 2015

Introduction

- The previous lectures introduced Bayes Decision Theory and probability models. The next series of lectures will discuss probabilities and their use for modeling visual phenomena.
- We will start by simple probability models which are conceptual. Then we proceed to directed graphs (related to causal graphs) and the concept of probability distributions defined over graphs,
- Next we proceed to undirected graphical models which capture spatial context and which are sufficient to give methods for some visual modules. This will introduce concepts like Gibbs distribution and Mean Field Theory.
- This type of approach is conceptually and historically important. But were limited by the difficulty of specifying realistic probability distributions able to capture the complexity of real images. But this is changing,

Cue coupling

- This section describes models for coupling different visual cues.
- Modeling visual cues requires complex models taking into account spatial and temporal context, The models in this section are simplified so that we can address the dependencies between different cues and how they can be coupled. Later lectures will discuss individual cues in more detail.

What are Visual Cues?

- A definition of a visual cue is a "statistic or signal that can be extracted from the sensory input by a perceiver, that indicates the state of some property of the world that the perceiver is interested in perceiving". This is rather vague. In reality, visual cues rely on underlying assumptions (which are often unstated) and only yield useful information in restricted situations.
- Here are examples of visual cues for depth. They include binocular stereo, shape from shading, shape from texture, structure from motion, and depth from perspective. A key property is that these depth cues are capable of estimating depth/shape by themselves if the other cues are unavailable. But often only for simplified stimuli which obey very specific assumptions.
- In practice, visual cues are often tightly coupled and require Bayesian modeling to tease out their dependencies and to capture their hidden assumptions. They can sometimes, but not always, be overridden by high level visual knowledge.

Vision modules and cue combination

- Quantifiable psychophysics experiments for individual cues are roughly consistent with the predictions of Bayesian models, see (Bulthoff & Mallot, 1988; Cumming et al., 1993) – but with some exceptions (Todd et al., 2001).These estimate the viewed shape/depth S using a generative model $P(I|S)$ for the image I and a prior $P(S)$ for the shape/depth. We will introduce these types of probability distributions in later lectures. We should stress that they are used to model realistic, but highly simplified situations where only simple families of shapes are considered (e.g., spheres and cylinders).
- But how are different visual cues 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 (Marr, 1982) who justified this strategy by invoking the principle of modular design.
- Marr proposed that surfaces should be represented by a *2 1/2D sketch* that 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.

Cue coupling from a probabilistic perspective

- We consider the problem of cue combination from a probabilistic perspective (Clark & Yuille, 1990).
- This suggests that we need to distinguish between situations when the cues are statistically independent of each other and situations when they are not. We also need to determine whether cues are using similar, and hence redundant, prior information.
- These considerations lead 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* that generate the image.
- Note that there is strong evidence that high-level recognition can affect the estimation of three-dimensional shape, e.g., a rigidly rotating inverted face mask is perceived as nonrigidly deforming face, while most rigidly rotating objects are perceived to be rigid.

Combining cues with uncertainty

- We first consider simple models that assume the cues compute representations independently, and then we combine their outputs by taking linear weighted combinations.
- Suppose there are two cues for depth that separately give estimates $\vec{S}_1^*, \vec{S}_2^*$. One strategy to combine these cues is by linear weighted combination yielding a combined estimate $\vec{S}^*$:

$$\vec{S}^* = \omega_1 \vec{S}_1^* + \omega_2 \vec{S}_2^*,$$

  where ω_1, ω_2 are positive weights such that $\omega_1 + \omega_2 = 1$.
- Landy et al. (1995) reviewed many early studies on cue combination and argued that they could be qualitatively explained by this type of model. They also discussed situations when the individual cues did not combine as well as "gating mechanisms" that require one cue to be switched off.

Case where weights are derived from uncertainties

- An important special case of this model is when the weights are measures of the uncertainty of the two cues. This approach is optimal under certain conditions and yields detailed experimental predictions, which have been successfully tested for some types of cue coupling (Jacobs, 1999; Ernst & Banks, 2002), see (Cheng et al., 2007; Gori et al., 2008) for exceptions.
- If the cues have uncertainties σ_1^2, σ_2^2, we set the weights to be $w_1 = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}$ and $w_2 = \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}$.
- The cue with lowest uncertainly has highest weight.
- This gives the linear combination rule:

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

Optimality of the linear combination rule (I)

The linear combination is optimal for the following conditions:
1. The two cues have inputs $\{\vec{C}_i : i = 1, 2\}$ and outputs $\vec{S}$ related by conditional distributions $\{P(\vec{C}_i|\vec{S}) : i = 1, 2\}$.
2. These cues are *conditionally independent* so that $P(\vec{C}_1, \vec{C}_2|S) = P(\vec{C}_1|\vec{S})P(\vec{C}_2|\vec{S})$ and both distributions are Gaussians:

$$P(\vec{C}_1|\vec{S}) = \frac{1}{Z_1} \exp\{-\frac{|\vec{C}_1 - \vec{S}|^2}{2\sigma_1^2}\},$$

$$P(\vec{C}_2|\vec{S}) = \frac{1}{Z_2} \exp\{-\frac{|\vec{C}_2 - \vec{S}|^2}{2\sigma_2^2}\}.$$

3. The prior distribution for the outputs is uniform.

Optimality of the linear combination rule (II)

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

$$\vec{S}_1^* = \arg\max_S P(\vec{C}_1|\vec{S}) = \vec{C}_1, \quad \vec{S}_2^* = \arg\max_S P(\vec{C}_2|\vec{S}) = \vec{C}_2.$$

- If both cues are available, then the optimal estimate is given by:

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

$$= \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\vec{C}_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\vec{C}_2,$$

which is the linear combination rule by setting $\vec{S}_1^* = \vec{C}_1$ and $\vec{S}_2^* = \vec{C}_2$.

Optimality of the linear combination rule: Illustration

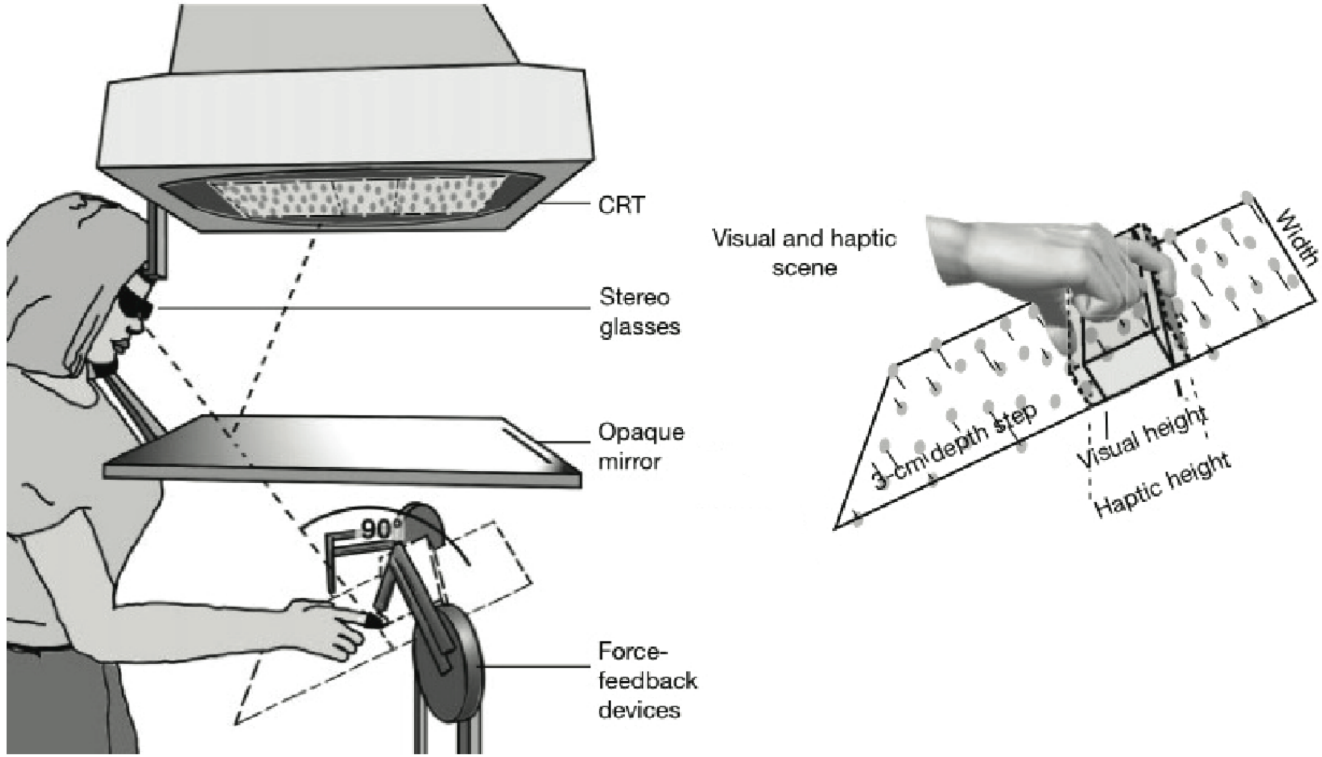

Figure 1: The work of Ernst and Banks shows that cues are sometimes combined by weighted least squares, where the weights depend on the variance of the cues. Figure adapted from Ernst & Banks (2002).

Bayesian analysis: Weak and strong coupling

- We now describe more complex models for coupling cues from a Bayesian perspective (Clark & Yuille, 1990; Yuille & Bulthoff, 1996), which emphasizes that the uncertainties of the cues are taken into account and the statistical dependencies between the cues are made explicit.
- Examples of cue coupling, where the cues are independent, are called "weak coupling" in this framework. In the likelihood functions are independent Gaussians, and if the priors are uniform, then this reduces to the linear combination rule.
- By contrast, "strong coupling" is required if the cues are dependent on each other.

The priors: Avoiding double counting

- Models of individual cues typically include prior probabilities about $\vec{S}$. For example, cues for estimating shape or depth assume that the viewed scene is piecewise smooth. Hence it is typically unrealistic to assume that the priors $P(\vec{S})$ are uniform.
- Suppose we have two cues for estimating the shape of a surface, and both use the prior that the surface is spatially smooth. Taking a linear weighted sum of the cues would not be optimal, because the prior would be used twice. Priors introduce a bias to perception, so we want to avoid doubling this bias.
- This is supported by experimental findings (Bulthoff & Mallot, 1988) in which subjects were asked to estimate the orientation of surfaces using shading cues, texture cues, or both. If only one cue, shading or texture, was available, subjects underestimated the surface orientation. But human estimates were much more accurate if both cues were present, which is inconsistent with double counting priors (Yuille & Bulthoff, 1996).

Avoiding double counting: Experiments

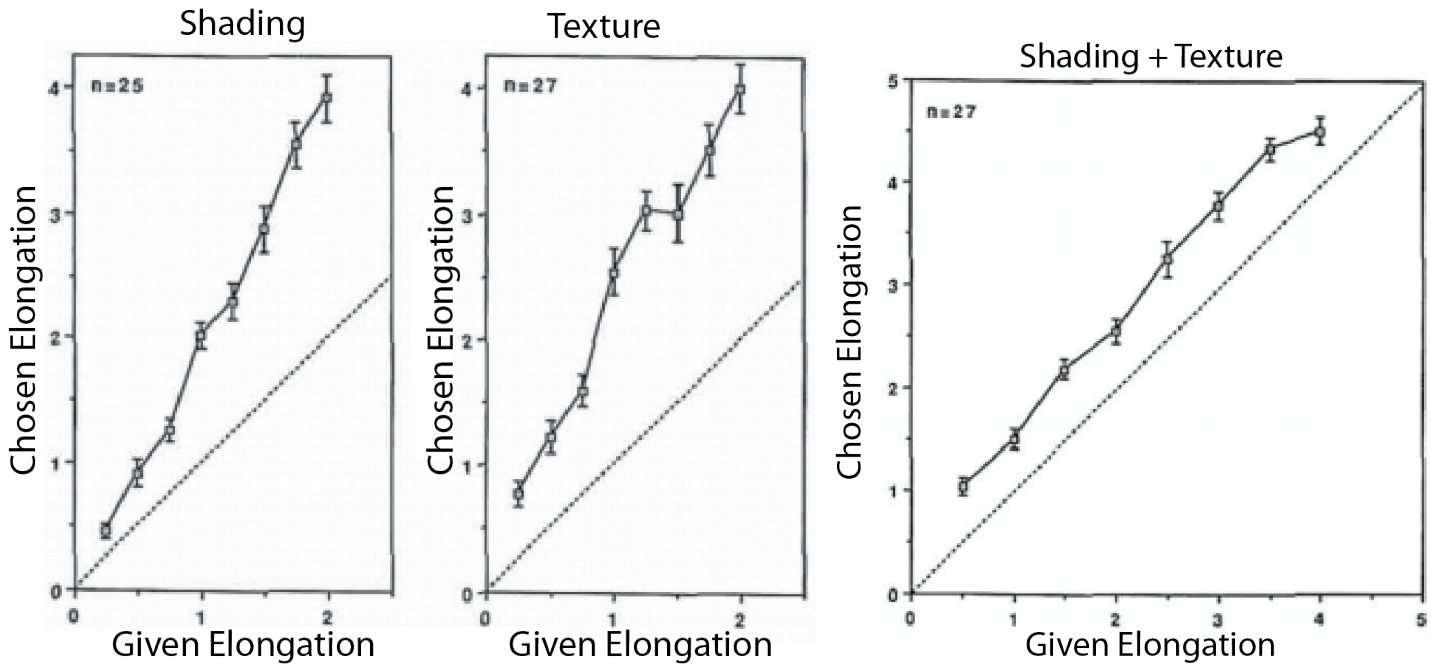

Figure 2: Cue coupling results that are inconsistent with linear weighted average (Bulthoff et al., 1990). Left: If depth is estimated using shading cues only, then humans underestimate the perceived orientation (i.e., they see a flatter surface). Center: Humans also underestimate the orientation if only texture cues are present. Right: 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 adapted from Bulthoff et al. (1990).

Avoiding double counting: Probabilistic analysis (I)

- We model the two cues separately by likelihoods $P(\vec{C}_1|\vec{S}), P(\vec{C}_2|\vec{S})$ and a prior $P(\vec{S})$. For simplicity we assume that the priors are the same for each cue.
- This gives posterior distributions for each visual cue:

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

- This yields estimates of surface shape to be $\vec{S}_1^* = \arg\max_{\vec{S}_1} P(\vec{S}|\vec{C}_1)$ and $\vec{S}_2^* = \arg\max_{\vec{S}_2} P(\vec{S}|\vec{C}_2)$.

Avoiding double counting: Probabilistic analysis (II)

- The optimal way to combine the cues is to estimate $\vec{S}$ from the posterior probability $P(\vec{S}|\vec{C}_1, \vec{C}_2)$:

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

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

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

Avoiding double counting: Probabilistic analysis (III)

- Coupling the cues, using the model in the previous slide, cannot correspond to a linear weighted sum, which would essentially be using the prior twice (once for each cue).
- To understand this, suppose the prior is $P(\vec{S}) = \frac{1}{Z_p} \exp\{-\frac{|\vec{S}-\vec{S}_p|^2}{2\sigma_p^2}\}$. Then, setting $t_1 = 1/\sigma_1^2, t_2 = 1/\sigma_2^2, t_p = 1/\sigma_p^2$, the optimal combination is $\vec{S}^* = \frac{t_1\vec{C}_1 + t_2\vec{C}_2 + t_p\vec{S}_p}{t_1+t_2+t_p}$, hence the best estimate is a linear weighted combination of the two cues $\vec{C}_1, \vec{C}_2$ and the mean $\vec{S}_p$ of the prior.
- By contrast, the estimate using each cue individually is given by $\vec{S}_1^* = \frac{t_1\vec{C}_1 + t_p\vec{S}_p}{t_1+t_2+t_p}$ and $\vec{S}_2^* = \frac{t_2\vec{C}_2 + t_p\vec{S}_p}{t_1+t_2+t_p}$.

Cue dependence and causal structure (I)

- Visual cues are rarely independent.
- In the flying carpet example, the perception of depth is due to perspective, segmentation, and shadow cues interacting in a complex way. The perspective and segmentation cues 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.

Cue dependence and causal structure (II)

- The conditional independent model is also problematic when coupling shading and texture cues (Bulthoff & Mallot, 1988). This model for describing these experiments presupposes that it is possible to extract cues $\vec{C}_1, \vec{C}_2$ directly from the image $\mathbf{I}$ by a preprocessing step that computes $\vec{C}_1(\mathbf{I})$ and $\vec{C}_2(\mathbf{I})$.
- This requires decomposing the image $\mathbf{I}$ into texture and shading components. This decomposition is practical for the simple stimuli used in (Bulthoff & Mallot, 1988). But in most natural images, it is extremely difficult, and detailed modeling of it lies beyond the scope of this chapter.

Causal structure: Ball-in-a-box

- The "ball-in-a-box" experiments (Kersten et al., 1997) suggest that visual perception does seek to find causal relations underlying the visual cues.
- In these experiments, an observer perceives the ball as rising off the floor of the box only if this is consistent with a cast shadow.
- 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, and 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 graphics to determine the most likely three-dimensional scene that generated the image sequence.

Causal structure: Ball-in-a-box figure

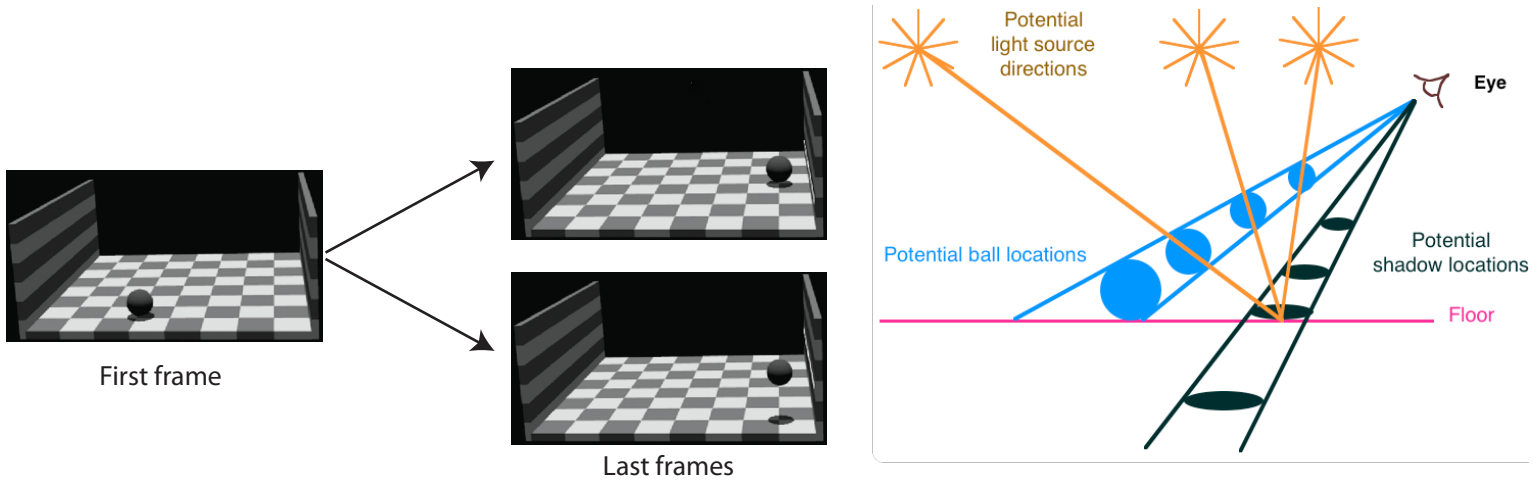

Figure 3: In the "ball-in-a-box" experiments, the motion of the shadow affects the perceived motion of the ball. The ball is perceived to rise from the ground if the shadow follows a horizontal trajectory in the image; but it is perceived to move towards the back of the box if the shadow follows a diagonal trajectory. See `http://youtu.be/hdFCJepvJXU`. Left: The first frame and the last frames for the two movies. Right: 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 (Kersten et al., 1997).

Directed graphical models

- Directed, or causal, graphical models (Pearl, 1988) offer a mathematical language to describe these phenomena. These are similar to the "undirected" graphical models used earlier, because the graphical structure makes the conditional dependencies between variables explicit, but the causal models differ in that the edges between nodes are directed.
- See chapters by Griffiths & Yuille (T. Griffiths, N. Chater, J.B. Tenenbaum. Bayesian Cognitive Science. 2024). for an introduction to undirected and directed graphical models from the perspective of cognitive science.

Formal directed graphical models

- *Directed graphical models* are formally specified as 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 $\vec{X}_{pa(\mu)}$.
- This gives a local *Markov property* – the conditional distribution $P(X_\mu|\vec{X}_{/\mu}) = P(X_\mu|\vec{X}_{pa(\mu)})$, so the state of X_μ is directly influenced only by the state of its parents (note $\vec{X}_{/\mu}$ denotes the states of all nodes except for node μ). Then the full distribution for all the variables can be expressed as:

$$P(\{X_\mu : \mu \in \mathcal{V}\}) = \prod_{\mu \in \mathcal{V}} P(X_\mu|\vec{X}_{pa(\mu)}). \tag{1}$$

Directed graphical models: Divisive normalization and Bayes-Kalman

- Two examples of directed graphical models are described in the Yuille-Kersten book chapter.
- First, when we studied divisive normalization used to represent the dependencies between the stimuli, the filter responses, and the common factor.
- Second, when exploring 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. The Bayes-Kalman filter will be discussed later in this lecture.
- Note that in some situations, the directions of the edges indicate 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 (Pearl, 2000).

Causal structure: Taxonomy of cue interactions

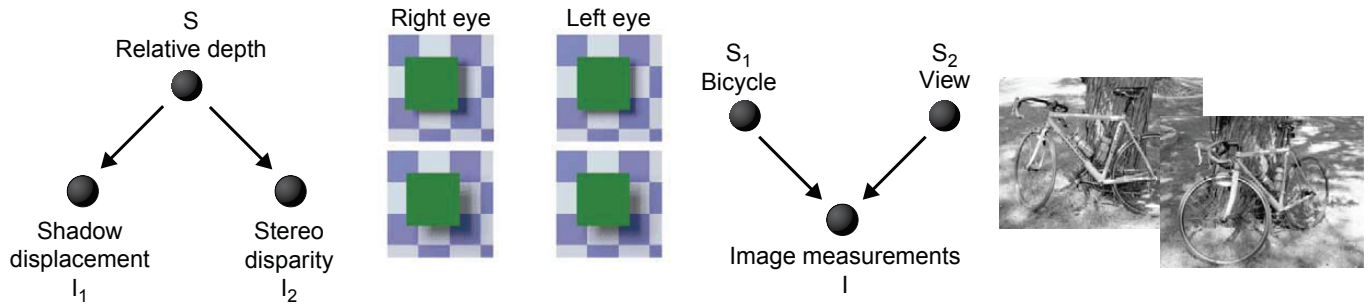

Figure 4: Graphical models give a taxonomy of different ways in which visual cues can be combined. Left: An example of common cause. The shadow and binocular stereo cues are caused by the same event – two surfaces with one partially occluding the other. Right: The image of the bicycle is caused by the pose of the bicycle, the viewpoint of the camera, and the lighting conditions.

Graphical models and explaining away (I)

- Graphical models can be used (Pearl, 1988) 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 you have a friend who claims they are psychic and predict the toss of a coin. You are sceptical and suspect they are cheating by using a double-headed coin. You challenge your friend by saying that if he is psychic then he should also be able to levitate a pencil. 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 Radio only to discover that there has been an earthquake. In this case, this new information "explains away" the alarm, so you stop worrying about a burglary.

Causal structure: Taxonomy of cue interactions (IA)

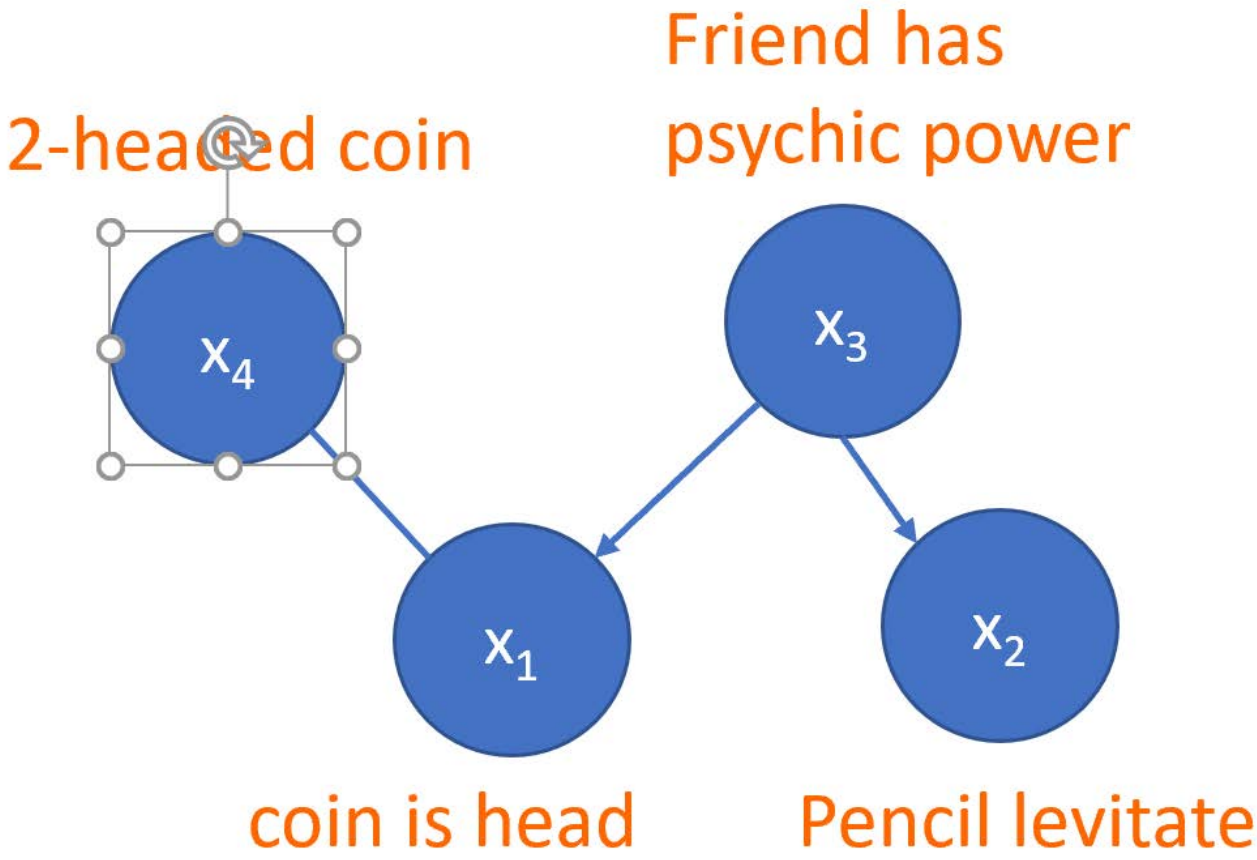

Figure 5: Graphical models showing the relationship between the variables. To confuse the reader, we replace "burglary" with "two-headed coin", "alarm goes off" by "coin is heads", "earthquake" by "friend has psychic powers", and "radio reports earthquake" by "pencil levitates". The story here is that your friend claims to have psychic powers and says he will prove it by predicting that a coin toss will be heads. You think he may be cheating by using a two-headed coin and challenge him to levitate a pencil (which should be easy if he has psychic powers).

Graphical models and explaining away (I)(A)

- The burglary example can be represented by a graph with nodes A, B, E, R. Each node has a binary-valued state variable X_A, X_B, X_E, X_R. If $X_A = 1$ the alarm activated ($X_A = 0$ if it did not). Similarly the variable X_B, X_E indicate whether there is a burglary, an earthquake or not. $X_R = 1$ if there is radio news about an earthquake ($X_R = 0$ is there is not).
- The joint distribution $P(X_A, X_B, X_E, X_R)$ can be decomposed as $P(X_A|X_B, X_E)P(X_B)P(X_E)P(X_R|X_E)$. This decomposition isolates the causal factors (but see Pearl 2000, true causality requires the ability to intervene and perform "graph surgery"). It also reduces the amount of data required to model the problem, both for learning the distributions and for inferring the best interpretation X_B, X_E of the observed data X_A, X_R.
- If there is no radio, then the graph is simplified $P(X_A, X_B, X_E) = P(X_A|X_B, X_E)P(X_B)P(X_E)$. Without the evidence from the radio we will probably interpret the alarm firing as a burglary. The new evidence (from the radio) changes our interpretation. Pearl invented this example to argue for the importance of using probabilities to model reasoning, because non-probabilistic approaches like conventional logic would find it difficult to deal with this situation.

Graphical models and explaining away (I)(B)

- Learning and Computation. How many parameters does the simple model $P(X_A, X_B, X_E)$ need? How do we compute the posteriors $P(X_B, X_E|X_A)$?
- A general distribution $P(X_A, X_B, X_E)$ has $2^3 - 1 = 7$ parameters (each variable X takes two possible values, yielding 2^3 but we subtract 1 because of the constraint $\sum_{X_A,X_B,X_E} P(X_A, X_B, X_E) = 1$). But if $P(X_A, X_B, X_E) = P(X_A|X_B, X_E)P(X_B)P(X_E)$ then it takes only $4 + 1 = 1 = 6$ parameters (note $P(X_A|X_B, X_E)$ has 4 parameters, because $\sum_{X_A} P(X_A|X_b, X_E) = 1$ for each possible value of X_B, X_E, and there are four possible values for X_B, X_E. This means that a (little) less data is required to learn it.
- To estimate the probability of an earthquake or a burglary we must compute the posterior distribution $P(X_B, X_E|X_A)$. This is $P(X_B, X_E|X_A) = \frac{P(X_A,X_B,X_E)}{P(X_A)} = \frac{P(X_A,X_B,X_E)}{\sum_{X_B,X_E} P(X_A,X_B,X_E)} = \frac{P(X_A|X_B,X_E)P(X_B)P(X_E)}{\sum_{X_B,X_E} P(X_A|X_B,X_E)P(X_B)P(X_E)}$.
- To compute the posteriors for earthquakes and burglaries separately, we compute $P(X_E|X_A) = \sum_{X_B} P(X_B, X_E|X_A)$ and $P(X_B|X_A) = \sum_{X_E} P(X_B, X_E|X_A)$. (The general rule is "sum out variables you are not interested in).

Graphical models and explaining away (I)(C)

- Similarly for the full model $P(X_A, X_B, X_E, X_R) = P(X_A|X_B, X_E)P(X_R|X_E)P(X_B)P(X_E)$, we find that there are $4 + 2 + 1 + 1 = 8$ parameters (instead of $2^4 - 1 = 15$).
- The posteriors $P(X_B, X_E|X_A, X_R)$ are computed to be $\frac{P(X_A|X_B,X_E)P(X_R|X_E)P(X_B)P(X_E)}{\sum_{X_B,X_E} P(X_A|X_B,X_E)P(X_R|X_E)P(X_B)P(X_E)}$.

Graphical models and explaining away (II)

- Variants of this phenomena arise in vision. Suppose you see the "partly occluded T" where a large part of the letter T is missing. In this case there is no obvious reason that 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.
- The Kanizsa triangle can also be thought of in these terms. The perception is of three circles partly occluded by the triangle. Hence the triangle explains why the circles are not complete. We will give a closely related explanation when we discuss model selection.

Directed graphical models and visual tasks (I)

- The human visual system performs a range of visual tasks, and the way cues are combined can depend on the tasks being performed.
- For example, consider determining the shape of a shaded surface. In most cases we need only shape from shading to estimate the shape of the surface. But occasionally we may want to estimate 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.

Directed graphical models and visual tasks (II)

- If we only want to estimate the surface shape S, then we do not care about the lighting L. The optimal Bayesian 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.
- 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 $P(L)$.
- If we want to estimate both the surface shape and the lighting, then we should estimate them using the full model $P(I|S,L)$ with priors $P(S)$ and $P(L)$.
- "Integrating out" nuisance, or generic, variables relates to the *generic viewpoint assumption* (Freeman, 1994) which 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).

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 are only valid for subparts of each image. For example, the lighting and geometry in most images are too complex to make shape from shading a reliable cue. Also shape from texture is only valid in restricted situations.
- Similarly, the visual system can use *perspective cues* to exploit the regular geometrical structure in the ball-in-a-box experiments. But such cues are only present in restricted classes of scenes, which obey the "Manhattan world" assumption. These cues 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, the cues are valid.

Model selection illustration

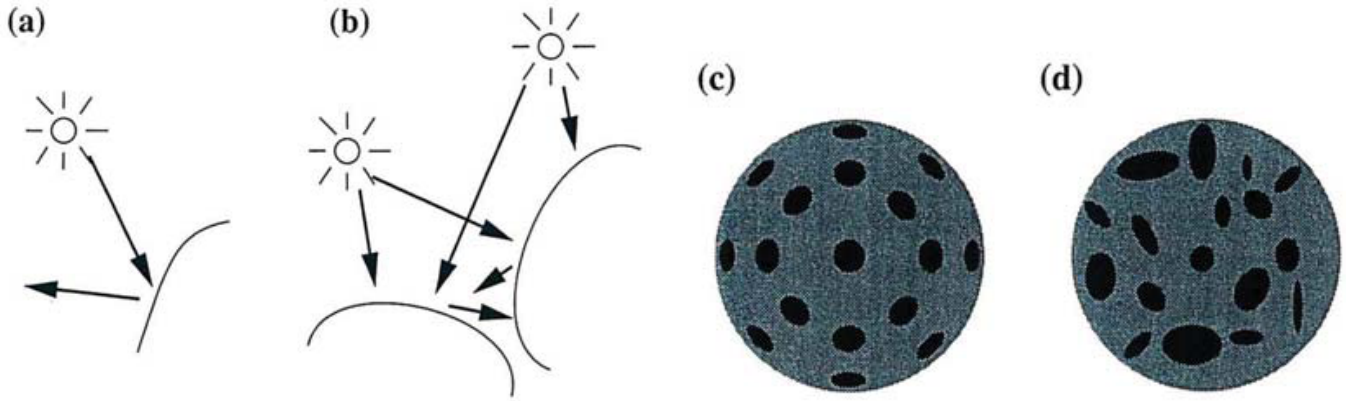

Figure 6: Model selection may need to be applied to decide if a cue can be used. Shape from shading cues will work for case (a) because the shading pattern is simply due to a smooth convex surface illuminated by a single source. But for case (b) the shading pattern is complex – due to mutual reflection 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.

Model selection examples

- Model selection also arises when there are several alternatives ways to generate the image.
- 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 with two rotating planes that can be arranged to have two competing explanations (Kersten et al., 1992). With slight variations to the transparency cues, the two surfaces can be seen to move rigidly together or to move independently (see `http://youtu.be/gSrUBpovQdU`).

Model selection: shadows and specularity

- A classic experiment (Blake & Bulthoff, 1990) studies human percption using a sphere with a Lambertian (diffuse) reflection function, which is viewed binocularly.
- A specular component is adjusted so that it can lie in front of the sphere, between the center and the sphere, or at the center of the sphere.
- If the specularity lies at the center, then it is perceived to be a transparent light bulb.
- If the specularity is placed between the center and the sphere, then the sphere is perceived to be shiny and specular.
- If the specularity lies in front of the sphere, then it is perceived as a cloud floating in front of a matte (Lambertian sphere).
- This is interpreted as strong coupling using model selection (Yuille & Bulthoff, 1996).

Model selection examples: Illustration

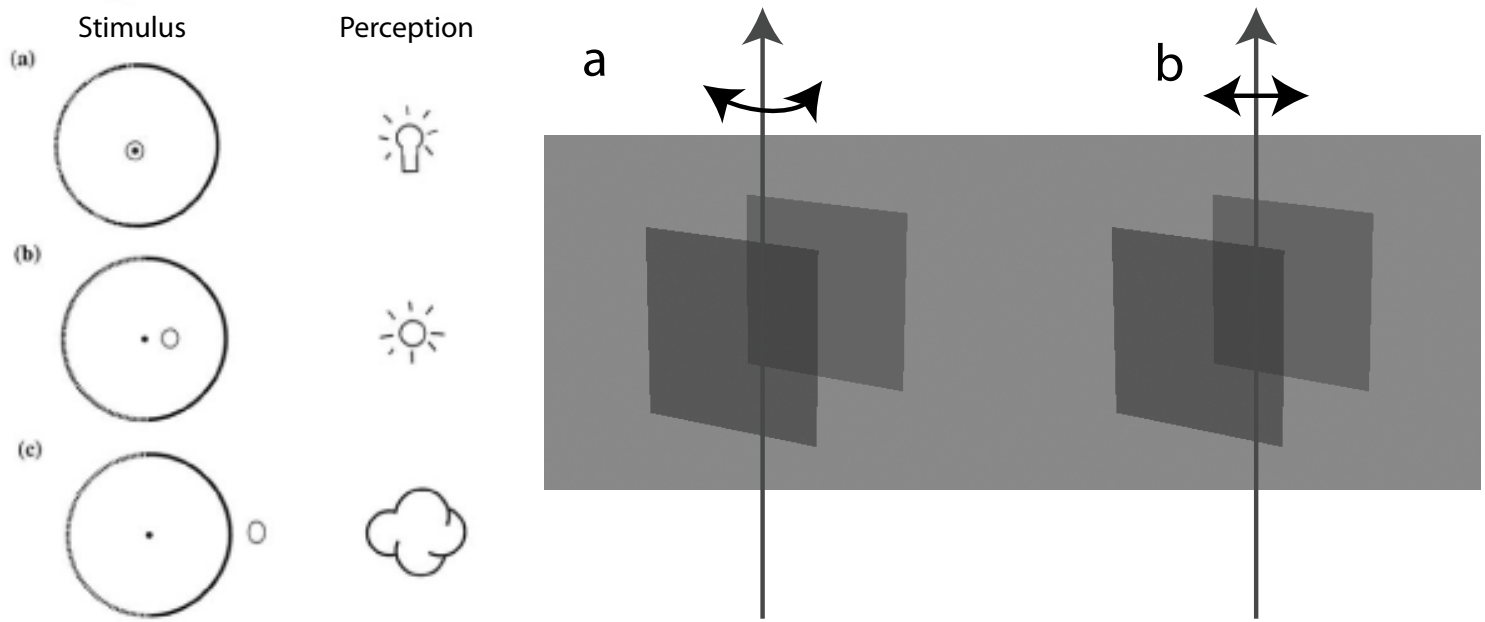

Figure 7: Examples of strong coupling with model selection. Left: A sphere is viewed binocularly, and small changes in the position of the specularity lead to very different percepts (Blake and Bülthoff, 1990). Right: Similarly altering, the transparency of the moving surfaces can make the two surfaces appear to rotate either rigidly together or independently.

Model selection and explaining away

- Model selection can also give an alternative explanation for "explaining away"
- For example, consider two alternative models for partially occluded T
- The first model is of two individual segments plus a smudge region. The second is a T that is partially hidden by a smudge. The second model is more plausible since it would be very unlikely, an accidental viewpoint (or alignment), for the smudge to happen to cover the missing part of the T, unless it really did occlude it.
- A similar argument can be applied to the Kanizsa triangle. One interpretation is three circles partly occluded by a triangle. The other is three partial circles arranged so that the missing parts of the circles are aligned. The first interpretation is judged to be most probable.

Flying carpet revisited

- Like Kersten's ball-in-a-box experiments, the flying carpet illusion requires estimating the depth and orientation of the ground plane (i.e., the beach), segmenting and recognizing the woman and the towel she is standing on, detecting the shadow, and then using the shadow cues, which requires making some assumptions 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, so the visual system "attaches" the shadow to the towel, which then implies that the towel must be hovering off the ground.

Examples of strong coupling

We now give 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

- Human observers are sensitive to both visual and auditory cues.
- Sometimes these cues have a common cause, e.g., you see a barking dog. But in other situations, the auditory and visual cues have different causes, e.g., a nearby cat moves and a dog barks in the distance.
- Ventriloquists are able to make the audience think that a puppet is talking by making it seem that visual cues (the movement of the puppet's head) and auditory cues (words spoken by the ventriloquist) are related. The ventriloquism effect occurs when visual and auditory cues have different causes – and so are in conflict – but the audience perceives them as having the same cause.

Multisensory cue coupling: The model (I)

- We describe an ideal observer for determining whether two cues have a common cause or not (Kording et al., 2007), which gives a good fit to experimental findings.
- The model is formulated using a meta-variable C, where $C = 1$ means that the cues x_A, x_V are coupled.
- More precisely, they are generated by the same process S by a distribution $P(x_A, x_V|S) = P(x_A|S)P(x_V|S)$.
  $P(x_A|S)$ 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 assumed 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 $N(0, \sigma_p^2)$.

Multisensory cue coupling: The model (II)

- $C = 2$ means that the cues are generated by two different processes S_A and S_B.
- In this case, the cues x_A and x_V are generated respectively by $P(x_A|S_A)$ and $P(x_V|S_V)$, which are both Gaussian $N(S_A, \sigma_A^2)$ and $N(S_V, \sigma_V^2)$. 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, it is a form of strong coupling with model selection (Yuille & Bulthoff, 1996).

Multisensory cue coupling: Illustration

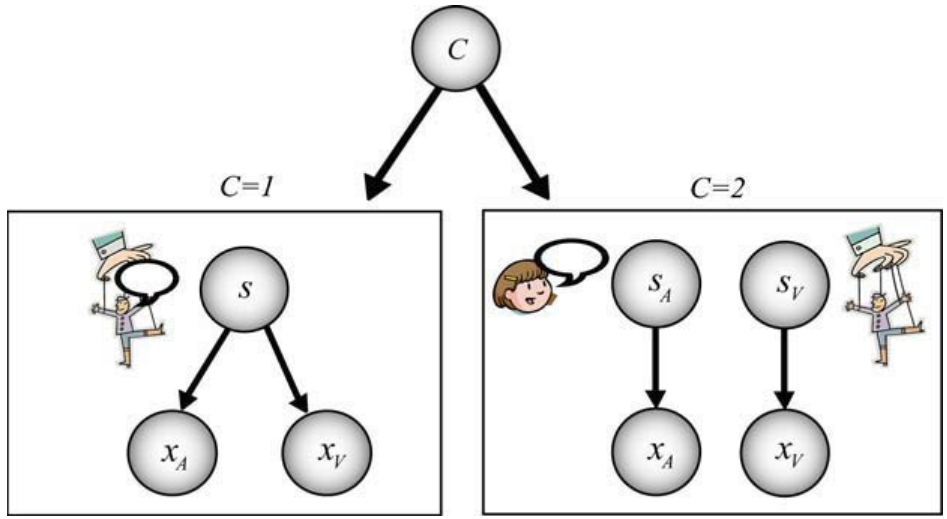

Figure 8: The subject is asked to estimate the position of the cues and to judge whether the cues are from a common cause – i.e., at the same location – or not. In Bayesian terms, the task of judging whether the cause is common can be formulated as model selection: are the auditory and visual cues more likely generated by a single cause (left) or by two independent causes (right)? Figure adapted from Kording et al. (2007).

Multisensory cue coupling: Comparison with experiments (I)

- This model was compared to experiments in which brief auditory and visual stimuli were presented simultaneously, with varying amounts of spatial disparity.
- Subjects were asked to identify the spatial location of the cue and/or whether they perceived a common cause (Wallace et al., 2004).
- The closer the visual stimulus was to the audio stimulus, the more likely subjects would perceive a common cause.
- In this case subjects' estimate of the stimuli's position was strongly biased by the visual stimulus (because it is considered more precise with $\sigma_V^2 > \sigma_A^2$).
- But if subjects perceived distinct causes, then their estimate was pushed away from the visual stimulus, and exhibited *negative bias*.

Multisensory cue coupling: Comparison with experiments (II)

- Körding et al. (2007) argue that this negative 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 in other cases, the auditory cause is farther away (more than 5 degrees).
- Hence the auditory cue will have a truncated Gaussian (if judged to be distinct) and will yield negative bias.

Multisensory cue coupling: Results and figure

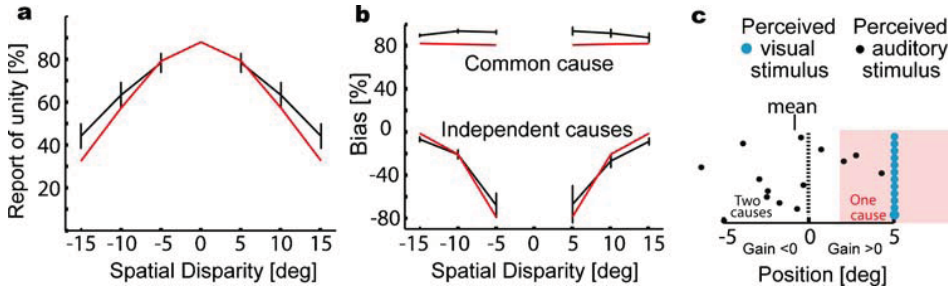

Figure 9: Reports of causal inference. (a) The relative frequency of subjects reporting one cause (black) is shown, with the prediction of the causal inference model (red). (b) The bias, i.e., the influence of vision on the perceivedauditory position, is shown (gray and black). The predictions of the 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. Reprinted with permission from Kording et al. (2007)

Multisensory cue coupling: The mathematics (I)

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:

$$P(C|x_A, x_V) = \frac{P(x_A, x_V|C)P(C)}{P(x_A, x_V)}$$

$$= \frac{\int dS P(x_A|S)P(x_V|S)P(S)}{P(x_A, x_V)}, \quad \text{if } C = 1,$$

$$= \frac{\int\int dS_A dS_V P(x_A|S_A)P(x_V|S_V)P(S_A)P(S_V)}{P(x_A, x_V)}, \quad \text{if } C = 2.$$

Multisensory cue coupling: The mathematics (II)

- 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:

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

- 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:

$$P(s_V, s_A|x_V, x_A) = \sum_C P(s_V, s_A|x_V, x_A, C)P(C|x_V, x_A)$$

$$= \sum_C \frac{P(x_V, x_A|s_V, s_A, C)P(s_V, s_A|C)P(C|x_V, x_A)}{P(x_V, x_A|C)},$$

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

Multisensory cue coupling: Extension

- Natarajan et al. (2008) showed that a variant of the model could fit the experiments even better.
- They replaced the Gaussian distributions with alternative distributions that are less sensitive to rare events. Gaussian distributions are non-robust because the tails of their distributions fall off rapidly, which gives very low probability to rare events.
- More precisely Natarajan et al. (2008) assumed that the data is distributed by a mixture of a Gaussian distribution, as above, and a uniform distribution (yielding longer tails).
- More formally, they assume $x_A \ \pi N(x_A : s_A, \sigma_A^2) + \frac{(1-\pi)}{r_l}$ and $x_V \ \pi N(x_V : s_V, \sigma_V^2) + \frac{(1-\pi)}{r_l}$, 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 (Knill, 2003).
- There are alternative models for generating the image, and the human observer must infer which is most likely. 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 are available, the system turns off the isotropy assumption and interprets texture using the homogeneity assumption only.

Homogeneous and isotropic texture: Illustration

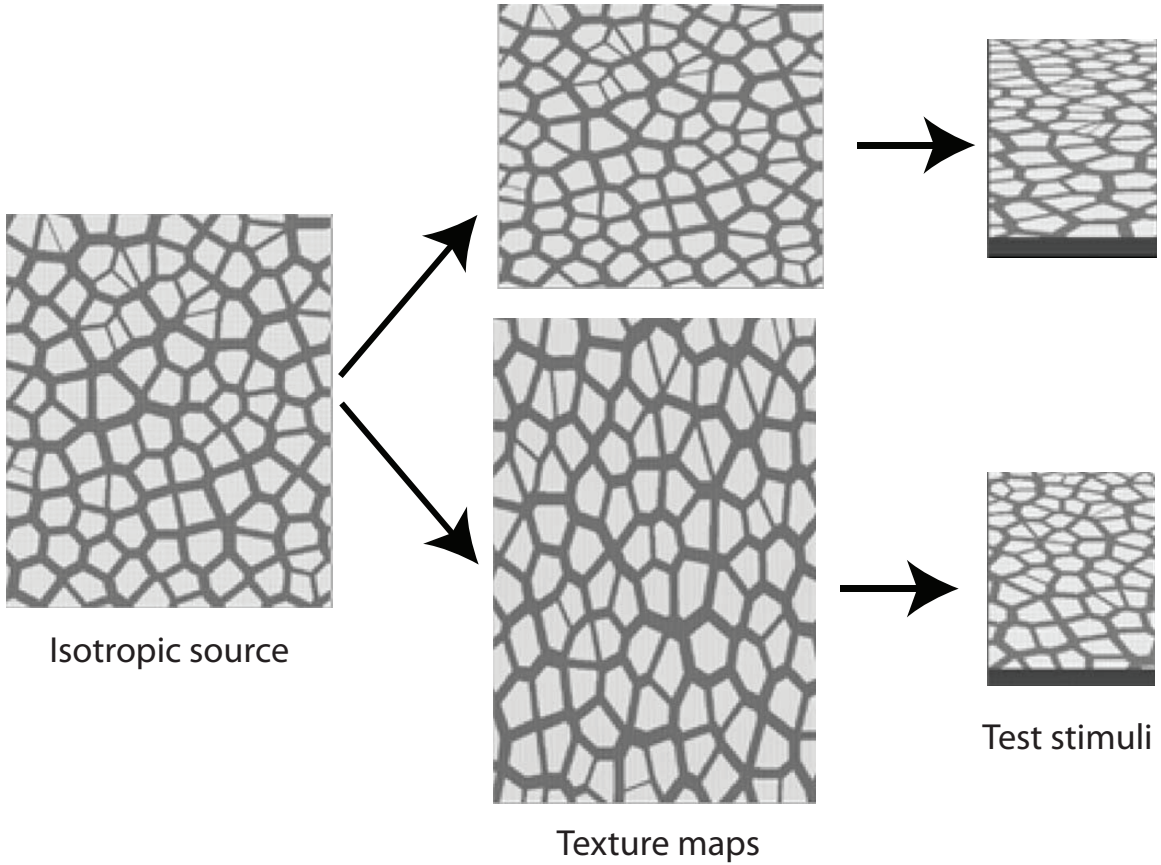

Figure 10: Generating textures that violate isotropy. An isotropic source image is either stretched (top middle) or compressed (bottom middle), producing texture maps that get applied to slanted surfaces shown on the right. A person that assumes surface textures are isotropic would overestimate the slant of the top stimulus and underestimate the slant of the bottom one. Figure adapted from from Knill (2003).

Homogeneous and isotropic texture: Theory (I)

- The posterior probability distribution for S is given by:

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

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:

$$P_h(T|(\sigma,\tau),\alpha,\theta), \quad P_i(T|(\sigma,\tau),\theta)$$

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

Homogeneous and isotropic texture: Theory (II)

- Isotropic textures are a special case of homogenous textures.
- 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 (MacKay, 2003) means that this advantage will disappear if we put priors $P(\alpha)P(\theta)$ on the model parameters and integrate them out. This gives:

$$P_h(T|(\sigma,\tau)) = \int\int d\alpha d\theta P_h(T|(\sigma,\tau),\alpha,\theta),$$

$$P_i(T|(\sigma,\tau)) = \int d\theta P_h(T|(\sigma,\tau),\theta).$$

- 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).

Homogeneous and isotropic texture: The mathematics

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

$$P(T|(\sigma,\tau)) = p_h P_h(T|(\sigma,\tau)) + p_i P_i(T|(\sigma,\tau)), \tag{2}$$

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:

$$P(\sigma,\tau|I) = \frac{P(I|(\sigma,\tau))P(\sigma,\tau)}{P(I)}. \tag{3}$$

- This model has a rich interpretation. If the data are consistent with an isotropic texture, then this model dominates the likelihood and strongly influences the perception. Alternatively, if the data are consistent only with homogeneous texture, then this model dominates. This gives a good fit to human performance (Knill, 2003).

It is straightforward to make these compositional models robust to occlusion. We simply specify a probability that the spatial patterns are corrupted by having incorrect visual concepts in some spatial positions. These corruptions correspond to occluders. Observe that it is possible to make this model robust to occluders because we are encoding parts explicitly (by visual concepts) which means that they can be automatically switched on or off (this is not possible for deep networks because they do not have explicit representations of parts, so the occluders cannot be switched off but instead will confuse the deep network).

The compositional model described so far has three types of learning; (i) the original deep network learning to learn the weights of the deep network, (ii) the clustering to learn the visual concepts, (iii) the clustering and EM algorithm to learn the spatial patterns (for different viewpoints of the objects).

This can be extended to a compositional deep network that is learnt end-to-end and minimizes a loss function that includes several different terms (classification loss for objects, loss for learning the visual concepts, and loss for learning the mixtures of spatial patterns.

Divisive normalization

- An important example is the use of probabilistic models (Wainwright & Simoncelli, 2000) to account for divisive normalization. This is a mechanism whereby cells mutually inhibit one another, effectively normalizing their responses with respect to stimulus inputs. Originally developed to explain nonlinear responses to contrast in V1 (Heeger, 1992), divisive normalization has been proposed as a basic cortical computation that underlies various effects of context, as well as higher-level processe,s such as attention (Carandini & Heeger, 2011) .
- The probabilistic approach gives 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.

Modeling divisive normalization using hidden variables

This can be modeled by assuming there are hidden variables ν that affect both responses and hence induces correlation between the responses. For example, ν could represent the local average image intensity, which could affect the response of both filters, but after the filter response, it 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 $\{I_i : i = 1, 2\}$ that are related to Gaussian models $\{g_1 : i = 1, 2\}$. Then we can model the activation of the set of filter responses:

$$P(I_1, I_2) = \int d\nu P(\nu) \prod_{i=1}^{2} P(I_i|\nu, g_i))P(g_i), \tag{4}$$

where $P(I_i|\nu, g_i) = \delta(I_i - \nu g_i)$. 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 the next figure.

Figure for divisive normalization model

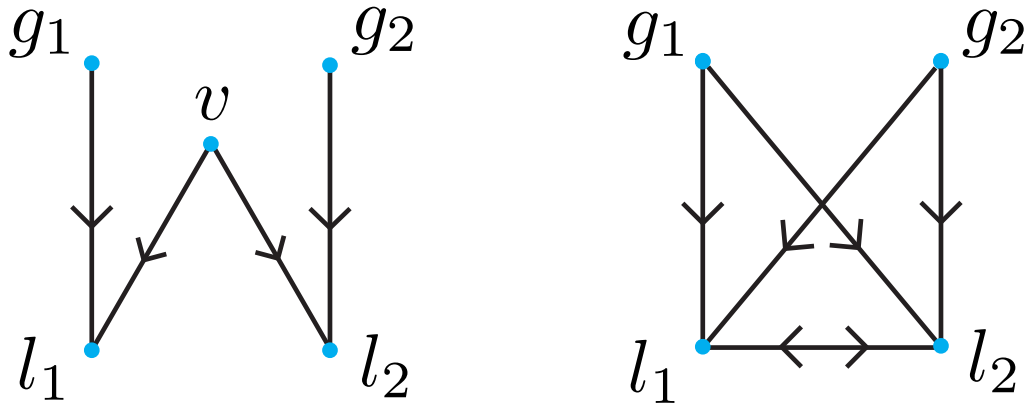

Figure 11: Left: The graphical structure of the divisive normalization model. The filter responses I_1, I_2 are generated from stimuli g_1, g_2 and by the common factor ν. The distributions of I_1, I_2 are factorized if we condition on ν. Right: But if we integrate out ν, then almost all the variables become dependent, as reflected by the complexity of the graph structure.

Divisive normalization model

- In particular, for each filter we can compute $P(g_i|I_1, I_2)$. After some algebra, this is computed to be:

$$P(g_1|I_1, I_2) = \frac{g_1^{-1} \exp\{-\frac{g_1^2 I^2}{2\sigma^2 I_1^2} - \frac{I_1^2}{2g_1^2}\}}{B(0, I/\sigma)}, \tag{5}$$

  where $I = \sqrt{I_1^2 + I_2^2}$, and $B(.,.)$ is a Bessel function. To get intuition, note that $g_1 = I_1/\nu$ and $g_1 = I_1/\nu$. So if ν is small, then $|I_1|$ and $|I_2|$ are likely to be small together, while if ν is large, then $|I_1|$ and $|I_2|$ 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 $\{I_i : i = 1, 2\}$, which gives the nonlinear response of the cell. It follows, from analysis above, that

$$E(g_1|I_1, I_2) \propto \text{sign}\{I_1\}\sqrt{|I_1|}\sqrt{\frac{|I_1|}{\sqrt{I_1^2 + I_2^2 + k}}}. \tag{6}$$

  The $\sqrt{I_1^2 + I_2^2 + k}$ term sets the gain and performs the divisive normalization.

Application to the tilt illusion

- The model has also been applied to explain the classic tilt illusion in perception (Schwartz et al., 2009; Qiu et al., 2013). 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." But for large differences between the center orientation and the surround (tilted left), the center vertical lines can appear to tilt left–an effect called "attraction." In the model, the population of neurons responding to the surround tilted lines contributes 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 local spatial context.

The tilt illusion

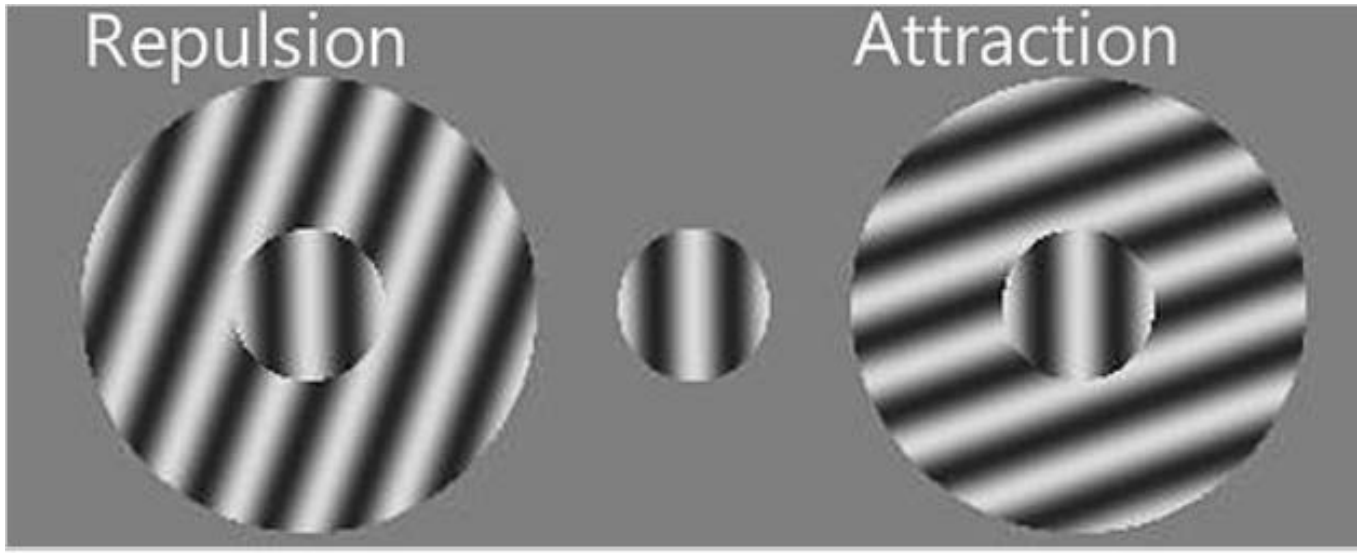

Figure 12: The perceived orientation of a grating pattern can appear to be tilted away from its true orientation due to the presence of surrounding gratings with different orientations. The central circular grating (Center Panel) appear to be tilted to the left (Left Panel) because it is *repulsed* from the orientation of the larger background grating (because the relative orientation is greater than 0 but less than 50 degrees). Conversely it is tilted slightly to the right (Right panel) when it is *attracted* to the background grating (where relative orientation is between 50 and 90 degrees).

Context and spatial interactions between neurons

- There is considerable evidence that low-level vision involves long-range spatial interaction,s 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 that differ from their neighbors tend to "pop out" and attract attention, while, conversely, similar image features that form spatially smooth structures tend to get "grouped" together to form a coherent percept, see chapter figure 12.26 (left panel). Image properties such as color tend to spread out, or fill in regions, until they hit a boundary (Grossberg & Mingolla, 1985; Sasaki et al., 2004) as shown in chapter figure 12.26 (right panel).

Context and spatial Interactions between neurons

- In general, there is a tendency for low-level vision to group 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 precursor 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 (Itti & Koch, 2001). It has been suggested that many of these processes are performed in V1 (Zhaoping, 2014), although this involves possibly feedback and interactions between V1 and V2 (Shushruth et al., 2013).

Context figures

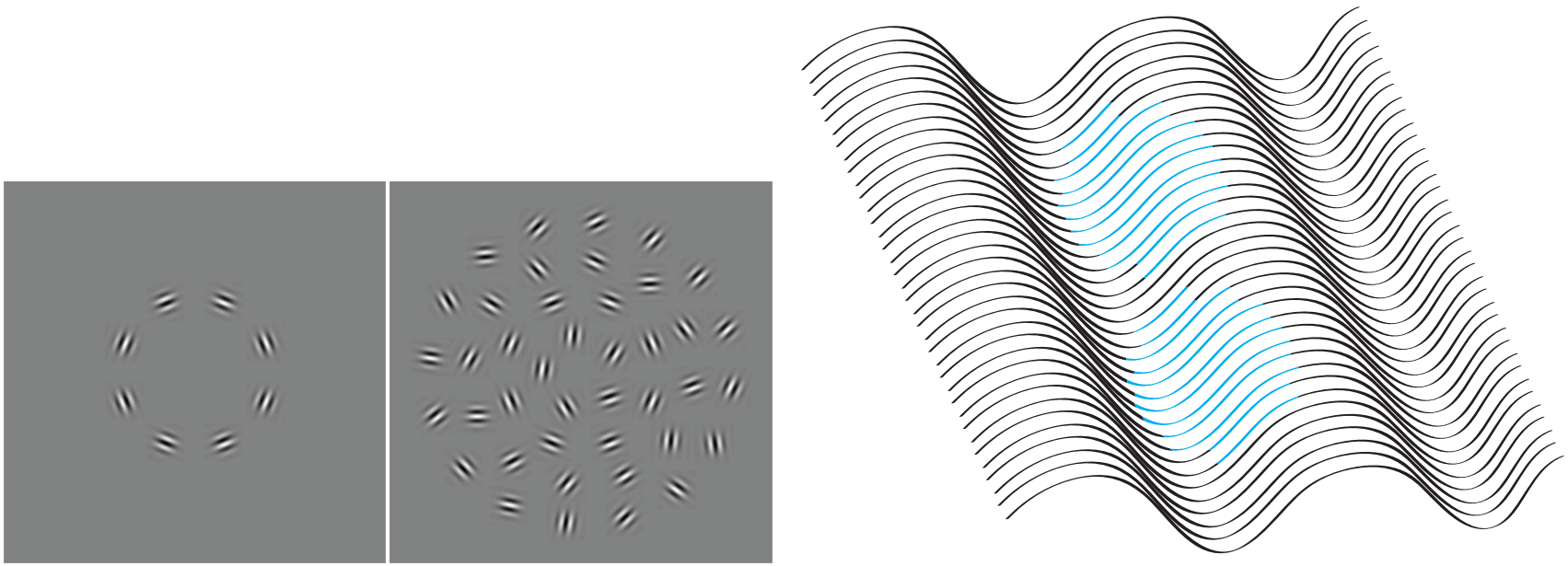

Figure 13: Left: Association fields. The circular alignment of Gabor patches (left) make it easier to see the circular form in the presence of clutter (right). Right: The neon color illusion. A bluish color appears to fill in the white regions between the blue lines, creating the appearance of blue transparent disks.

d

Context electrophysiology

The psychophysical and theoretical studies discussed so far are supported by single-electrode studies (Lamme, 1995; Lee & Yuille, 2006), which show that the activities of neurons on monkey area V1 appear to involve spatial interactions with other neurons. When monkeys are shown stimuli consisting of a textured square surrounded by a background with a different texture, 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 (Yuille, 2006). There is also a considerable literature on the related topic of *nonclassical receptive fields* (Kapadia et al., 2000).