Lecture 6

This lecture discusses the dependencies between visual cues. How these can be modeled by graphical models, often with causal structure.
We also briefly discuss how the models in these lectures can fit with theories of high-level vision.

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 shadows 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 [18]. 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 pre-processing step which 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 [18]. 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 [79] suggest that visual perception does seek to find causal relations underlying the visual cues.
In these experiments an observer perceives the ball to rise 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, 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.

Causal Structure: Ball-in-a-Box Figure

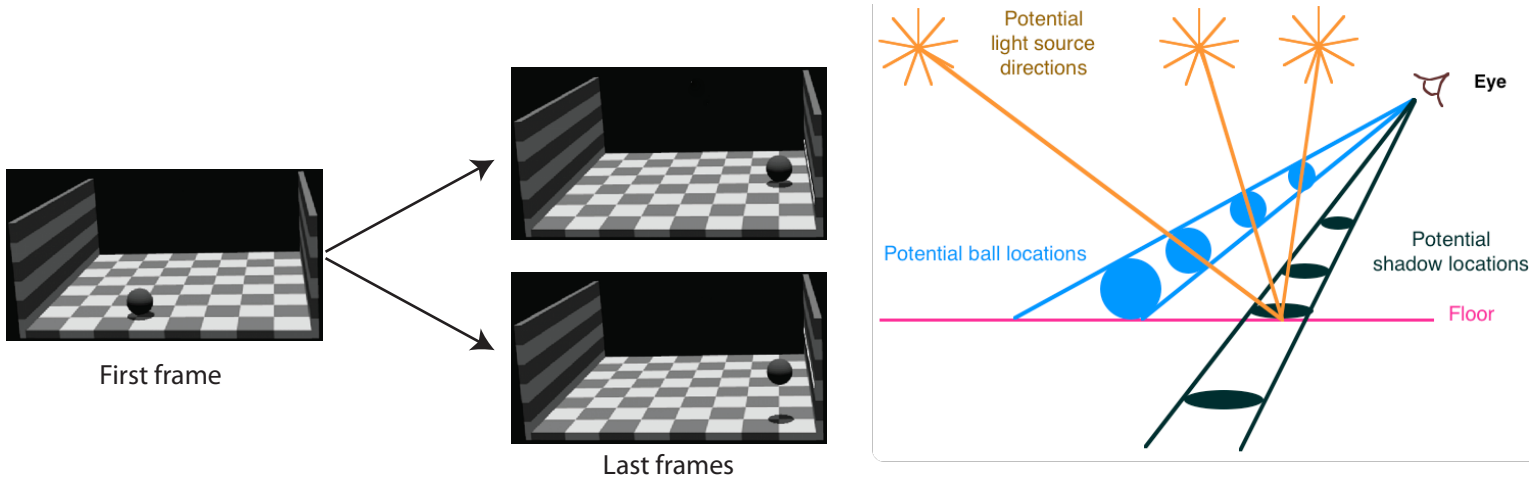

Figure 35: In the "ball-in-a-box" experiments the motion of the shadow affects the perceived motion of the ball. The ball is perceived to rise from the ground if the shadow follows a horizontal trajectory in the image; but is perceived to move towards the back of the box if the shadow follows a diagonal 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 used earlier, 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 the perspective of cognitive science.

Formal Directed Graphical Models

Directed graphical models are formally 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 $\vec{X}_{pa(\mu)}$.
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 only directly influenced 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:

$$P(\{X_\mu : \mu \in \mathcal{V}\}) = \prod_{\mu \in \mathcal{V}} P(X_\mu|\vec{X}_{pa(\mu)}). \tag{42}$$

Directed Graphical Models: Divisive Normalization and Bayes-Kalman

We have already seen two examples of directed graphical models in this chapter. Firstly, when we studied divisive normalization they were used 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].

Causal Structure: Taxonomy of Cue Interactions

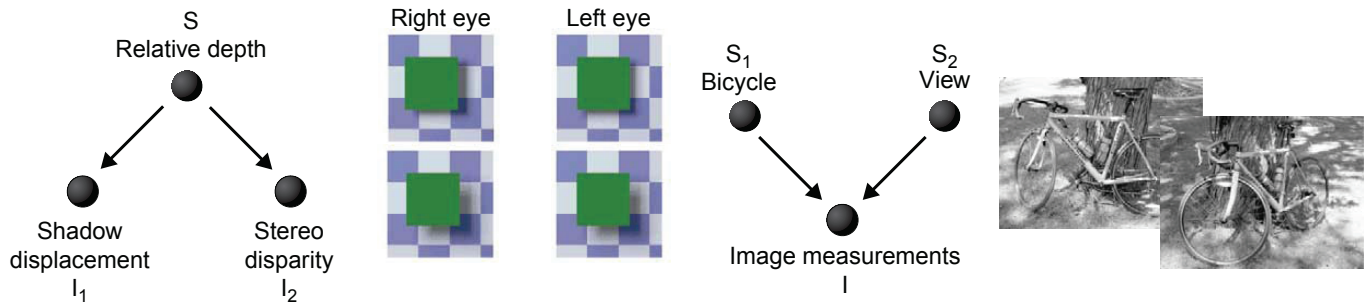

Figure 36: Graphical Models give a taxonomy between different ways that visual cues can be combined. Left Panel: An example of common cause. The shadow and binocular stereo cues are caused by the same event – two surfaces with one partially occluding the other. Right Panel: The image of the bicycle is caused by the pose of the bicycle, the viewpoint of the camera, and the lighting conditions.

Graphical Models and Explaining Away (I)

Graphical models can be used [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.

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 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. The Kanisza triangle can also be thought of in these terms. The perception is of three circles which are 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 which are being performed.
For example, consider a shaded surface. In most cases we want to perform 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 $P(L)$.
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 $P(L)$.
"Integrating out" nuisance, or generic, variables relates to the *generic viewpoint assumption* [35] 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. E.g., the lighting and geometry in most images is 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, is the cues are valid.

Model Selection Illustration

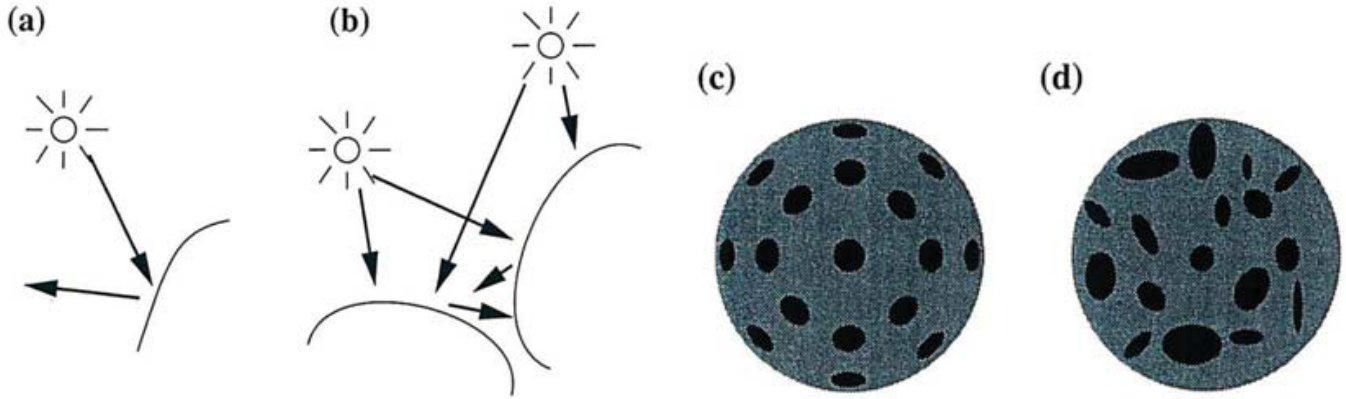

Figure 37: Model selection may need to be applied in order to decide is a cue can be used. Shape from shading cues will work for case (a) because the shading pattern is simply due to a smooth convex surface illuminated by a single source. But for case (b) the shading pattern is complex – due to mutual 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.

Model Selection Examples

Model selection also arises in situations where there are several alternatives ways which could 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 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 `http://youtu.be/gSrUBpovQdU`.

Model selection: shadows and specularity

A classic experiment [10] studies human percption where a sphere has a Lambertian (diffuse) reflection function and is viewed binocularly.
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 the sphere is perceived to 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 [189].

Model Selection Examples: Illustration

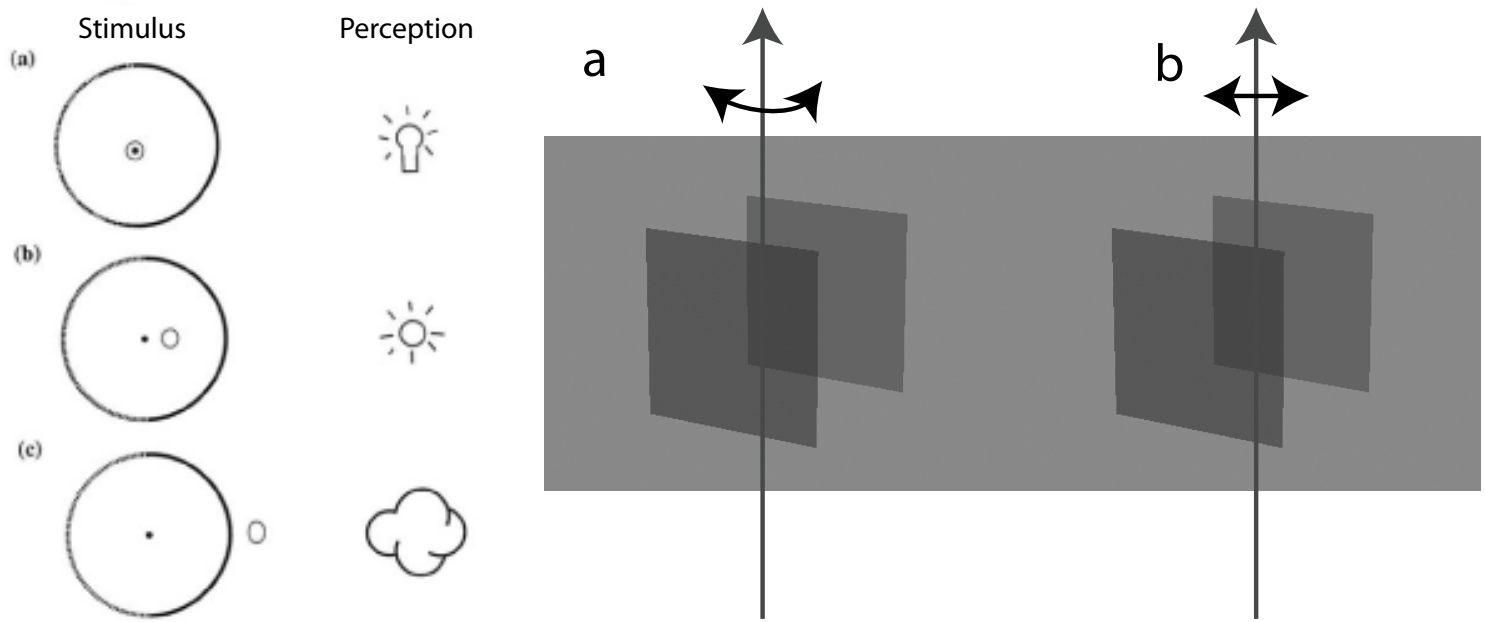

Figure 38: Examples of strong coupling with model selection. A sphere is viewed binocularly (left) and small changes in the position of the specularity lead to very different percepts (Blake and Bülthoff 1990). Similarly altering the transparency of the moving surfaces (right) can make the two surfaces appear to rotate either rigidly together or independently.

Model Selection and Explaining Away

Model selection can also give an alternative explanation for "explaining away". For example consider partially occluded T. We consider two alternative models for the 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.

A similar argument can be applied to the Kanisza triangle. One interpretation is three circles which are partly occluded by a triangle. The other is is three partial circles which are arranged so that the missing parts of the circles are aligned. The first interpretation is judged to be most probable.

Flying Carpet revisited

We revisit the flying carpet illusion.
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.

Examples of Strong Coupling

We know 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 (I)

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 taking by making it seem that visual (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 perceive 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 [89] 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 is a form of strong coupling with model selection [185].

Multisensory Cue Coupling: Illustration

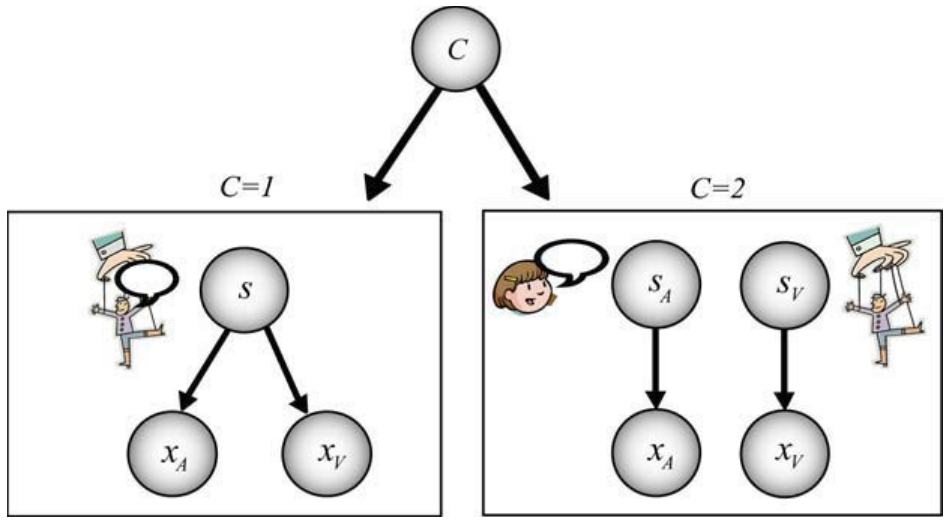

Figure 39: The subject is asked to estimate the position of the cues and to judge whether the cues are from a common cause – i.e. at the same location – or not. In Bayesian terms the task of judging whether the cause is common can be formulated as model selection – are the auditory and visual cues more likely to generated from a single cause (left) or by two independent causes (right). Figure adapted from [89].

Multisensory Cue Coupling: Comparison with Experiments (I)

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*.

Multisensory Cue Coupling: Comparison with Experiments (II)

Körding *et al.* [89] 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 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.

Multisensory Cue Coupling: Results and Figure

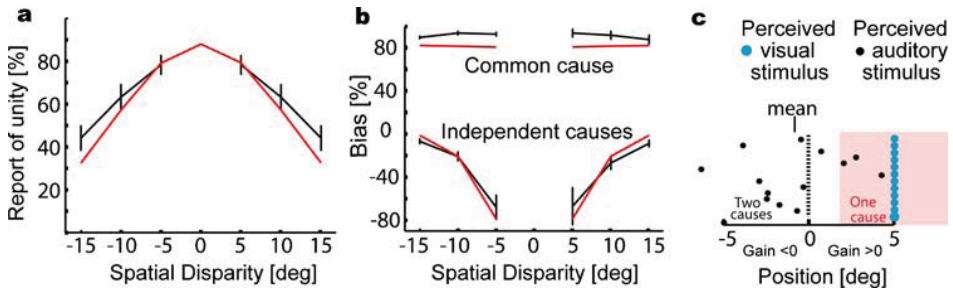

Figure 40: Reports of causal inference.a) The relative frequency of subjects reporting one cause (black) is shown (reprinted with permission from [89]) with the prediction of the causal inference model (red). b) The bias, i.e. the influence of vision on the perceived auditory position is shown (gray 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.

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.* [124] showed a variant of the model could fit the experiments even better.
They replaced the Gaussian distributions by alternative distributions which 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 [124] 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 [82].
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 is available the system turns off the isotropy assumption and interprets texture using the homogeneity assumption only.

Homogeneous and Isotropic Texture. Illustration

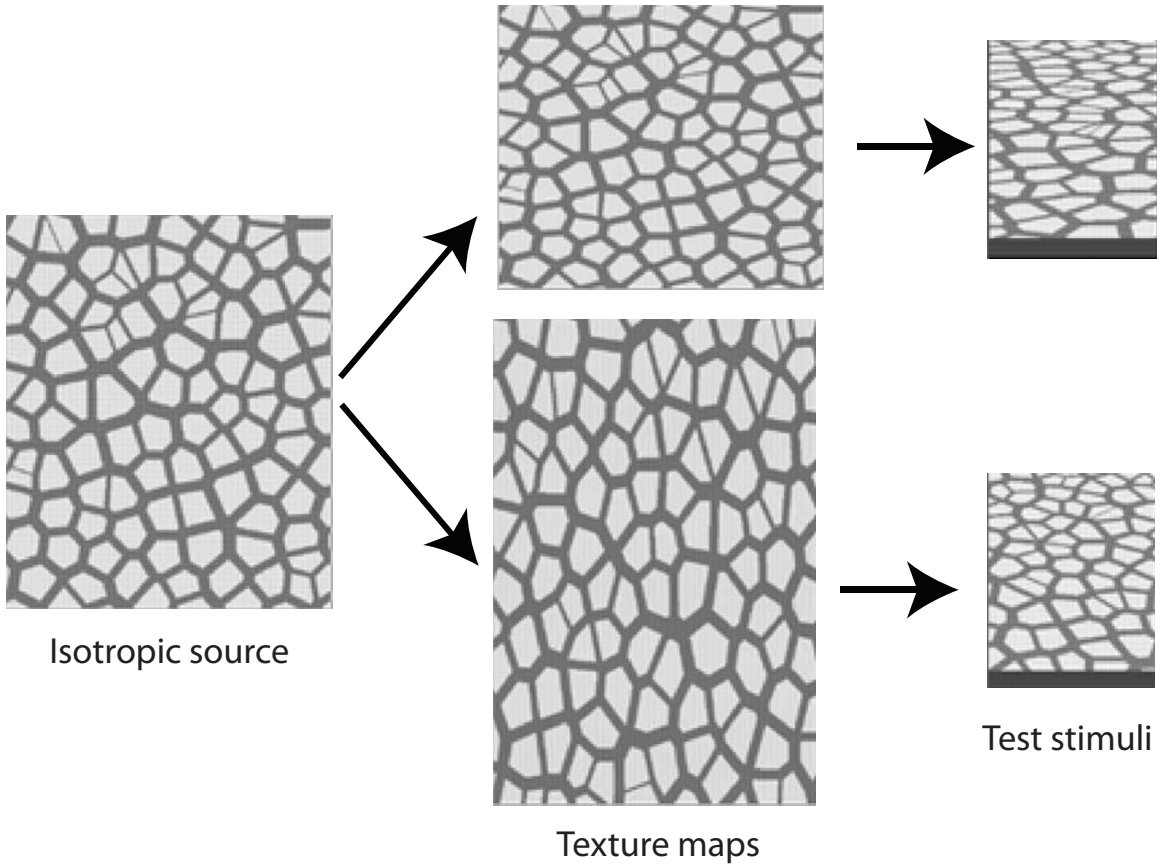

Figure 41: Generating textures that violate isotropy [82]. An isotropic source image is either stretched (top middle) or compressed (bottom middle) producing texture maps that get applied to slanted surfaces shown on the right. A person that assumes surface textures are isotropic would overestimate the slant of the top stimulus and underestimate the slant of the bottom one. Figure adapted from from [82].

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 [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:

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

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

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 (III)

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

$$P(T|(\sigma,\tau)) = p_h P_h(T|(\sigma,\tau)) + p_i P_i(T|(\sigma,\tau)), \tag{43}$$

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:

$$P(\sigma,\tau|I) = \frac{P(I|(\sigma,\tau))P(\sigma,\tau)}{P(I)}. \tag{44}$$

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].