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.

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 (out of scope of this lecture).

How to Study Visual Cues: Modules

- Visual cues – like binocular stereo, shape from shading, and shape from texture – can be formulated in Bayesian terms (see next lecture). These are often called modules, which can be studied independently.
- Some of these classic vision modules are only effective on limited types of images – this includes shape from shading and shape from texture. Others like binocular stereo and motion are still, with modifications, working today.
- Cognitive Science researchers introduced computer graphics psychophysics for studying these cues. They could create realistic stimuli and measure human performance, particularly for shape estimation. Many of the experimental findings (Bulthoff & Mallot, 1988; Cumming et al., 1993) were shown to be consistent with Bayesian theories.

Combining Visual Cues

- Marr's influential book (Marr 1982) proposed that cues for shape and depth were combined in a a *2 1/2D sketch* that represents the shape and depth of a surface by the distance of the surface points from the viewer. A related representation, *intrinsic images*, also includes the material properties of the surface. Both representations are still used.
- Marr did not give details about how cues should be combined. His book only proposed that they should be computed separately and their outputs combined.
- A simpler, more empirical theory was developed by Landy and Maloney who proposed that most cues could be combined by linear weighted sum with a gating mechanism, which we discuss later,

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. This requires determining whether cues are using similar, and hence redundant, prior information.
- This leads to a distinction between *weak*, where cues are independent, and *strong* coupling, where they have complex interactions often depending on the *causal factors* that generate the image.
- Human depth perception is certainly influenced by object recognition, e.g., a rigidly rotating inverted face mask is perceived as nonrigidly deforming face, where recognizing the image as a face overrides cues like shading, stereo, and structure from motion. This can be addressed by model selection.

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

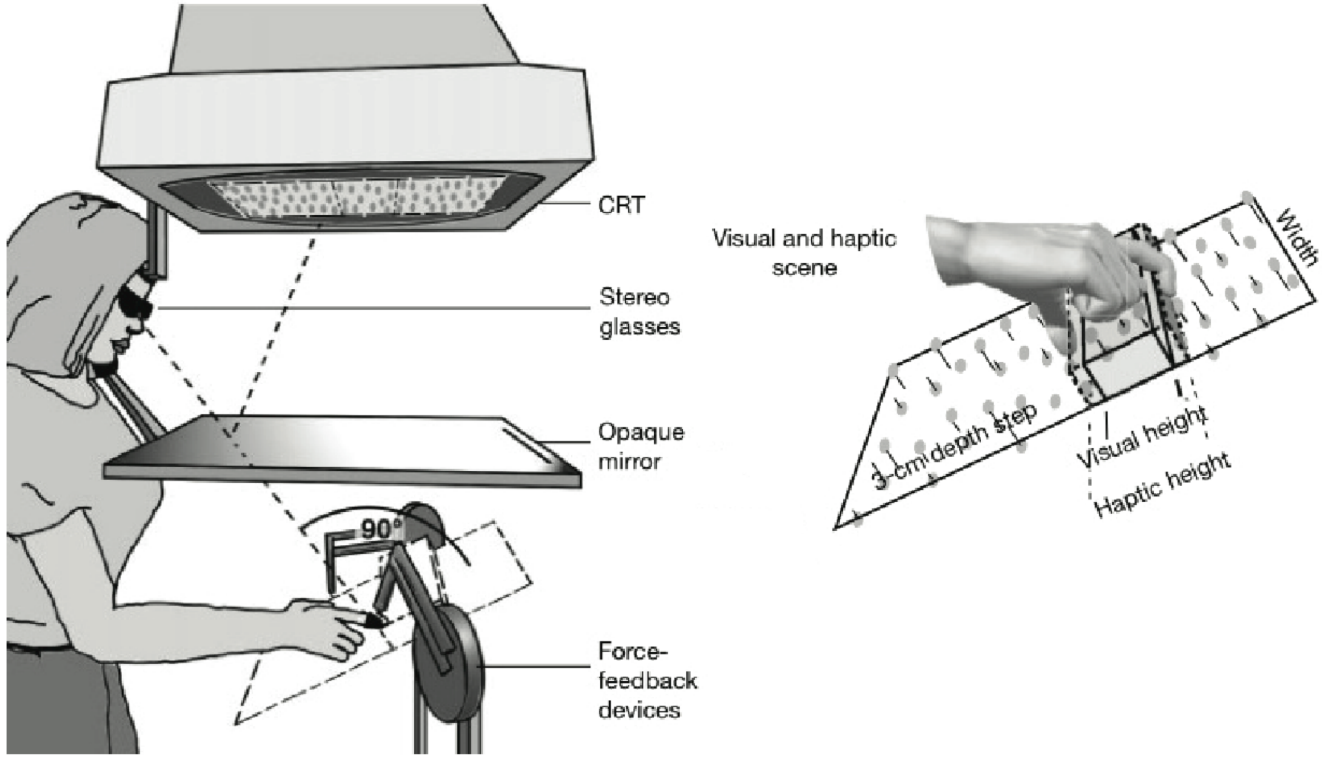

Figure 1: The work of Ernst and Banks shows that cues are sometimes combined by weighted least squares, where the weights depend on the variance of the cues. Figure adapted from Ernst & Banks (2002).

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

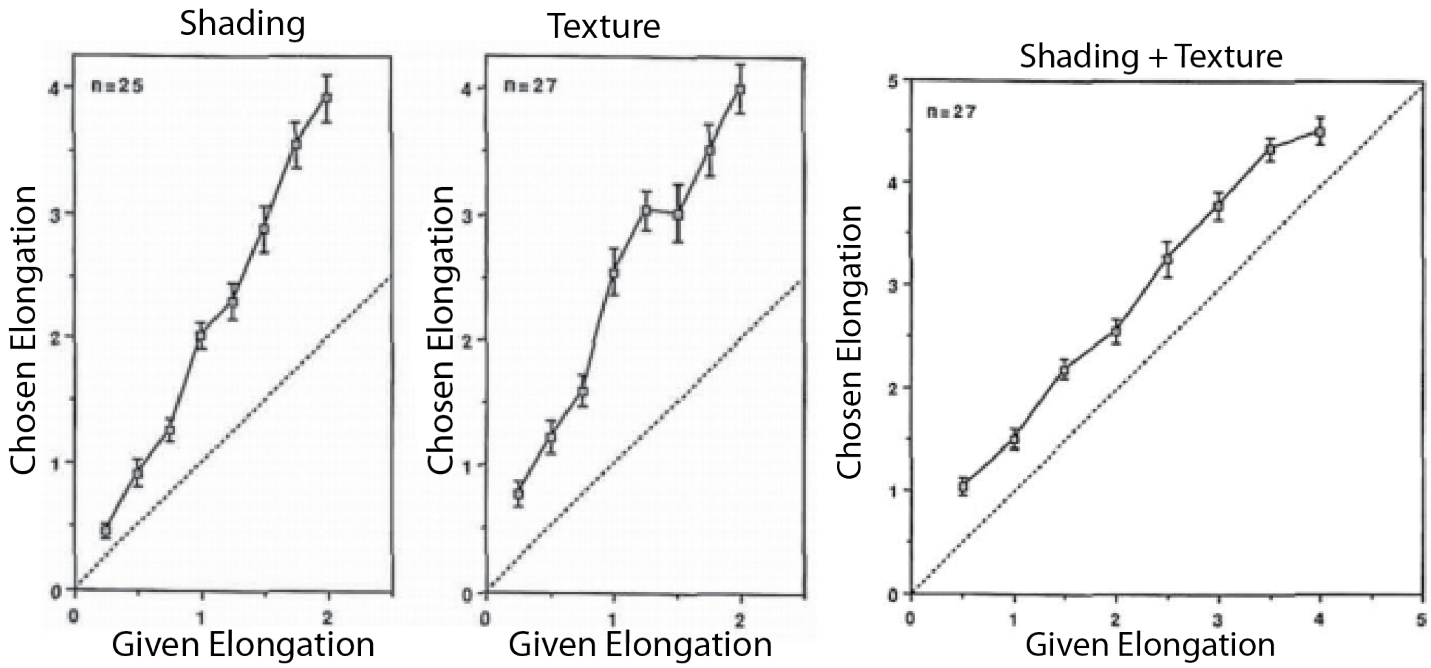

Figure 2: Cue coupling results that are inconsistent with linear weighted average (Bulthoff et al., 1990). Left: If depth is estimated using shading cues only, then humans underestimate the perceived orientation (i.e., they see a flatter surface). Center: Humans also underestimate the orientation if 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. But their dependence is hard to model. It depends on the causal factors that combine to generate the image.
- 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.
- Open question – are current Vision Language Models able to correctly interpret the flying carpet example? If so, how?

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

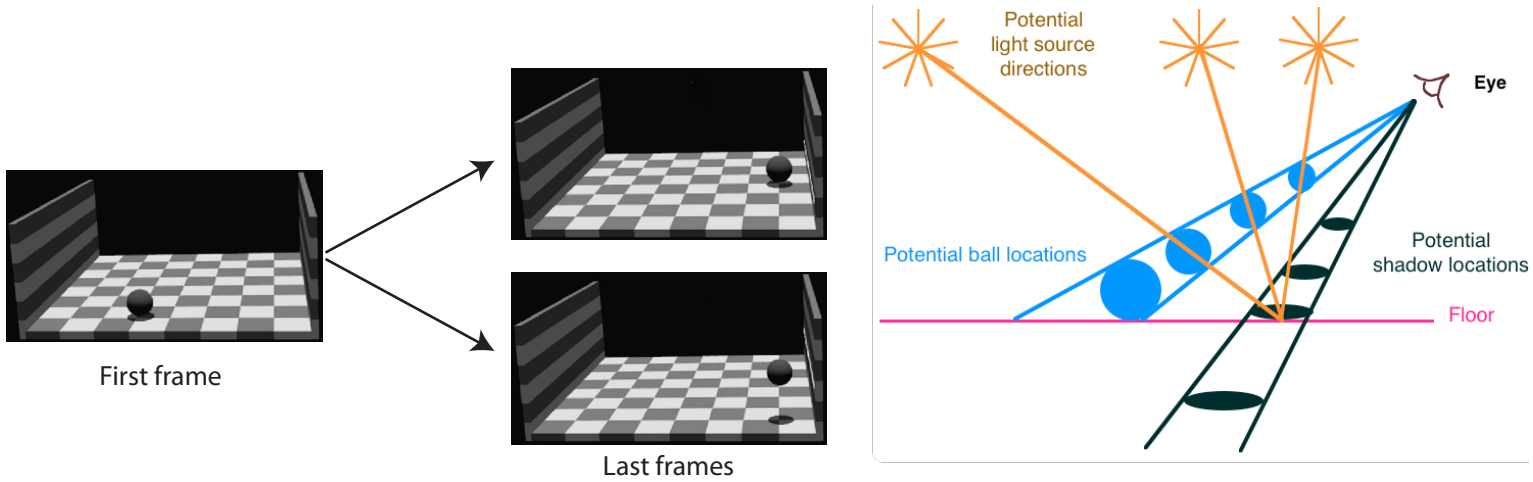

Figure 3: In the "ball-in-a-box" experiments, the motion of the shadow affects the perceived motion of the ball. The ball is perceived to rise from the ground if the shadow follows a horizontal trajectory in the image; but it is perceived to move towards the back of the box if the shadow follows a diagonal 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

- Modeling how different factors combine to generate an image can sometimes be modeled by directed graphs.
- Directed, or causal, graphical models (Pearl, 1988) offer a mathematical language to describe these phenomena. The graphical structure makes the conditional dependencies between variables explicit. 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) – causality require humans to be able to intervene and alter the graph structure.
- See chapters by Griffiths & Yuille (T. Griffiths, N. Chater, J.B. Tenenbaum. Bayesian Cognitive Science. 2024). for an introduction to directed and undirected 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

- Examples of directed graphical models are described in the Yuille-Kersten book chapter. Some will be described in the next lecture.
- Divisive normalization and explaining the tilt illusion.
- The Bayes-Kalman filter which combines information over time to estimate temporal properties of motion.
- Taxonomy of Cue Interactions, see next page.

Causal structure: Taxonomy of cue interactions

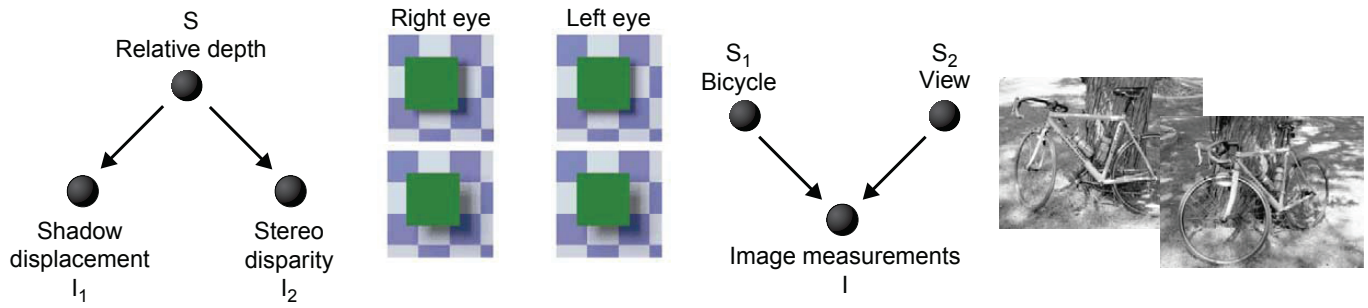

Figure 4: Graphical models give a taxonomy of different ways in which visual cues can be combined. Left: An example of common cause. The shadow and binocular stereo cues are caused by the same event – two surfaces with one partially occluding the other. Right: The image of the bicycle is caused by the 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. They correctly predict that the coin toss is heads. You are skeptical 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. The probability of the coin toss x_1 depending on a 2-headed coin x_4 or you friend having psychic powers x_3 can be modeled $P(x_1|x_4, x_3)$ and priors $P(x_4), P(x_3)$ for a 2-headed coin and your friend having psychic powers. In general, the prior probability of a 2-headed coin is much higher than the prior probability of your friend having psychic powers. So the coin toss is heads, it is more likely that the coin is 2-sided. But suppose, after the coin toss, you challenge your friend to levitate a pencil. and they succeed. In this case, you change your interpretation and believe your friend has psychic powers.

Causal structure: Taxonomy of cue interactions (IA)

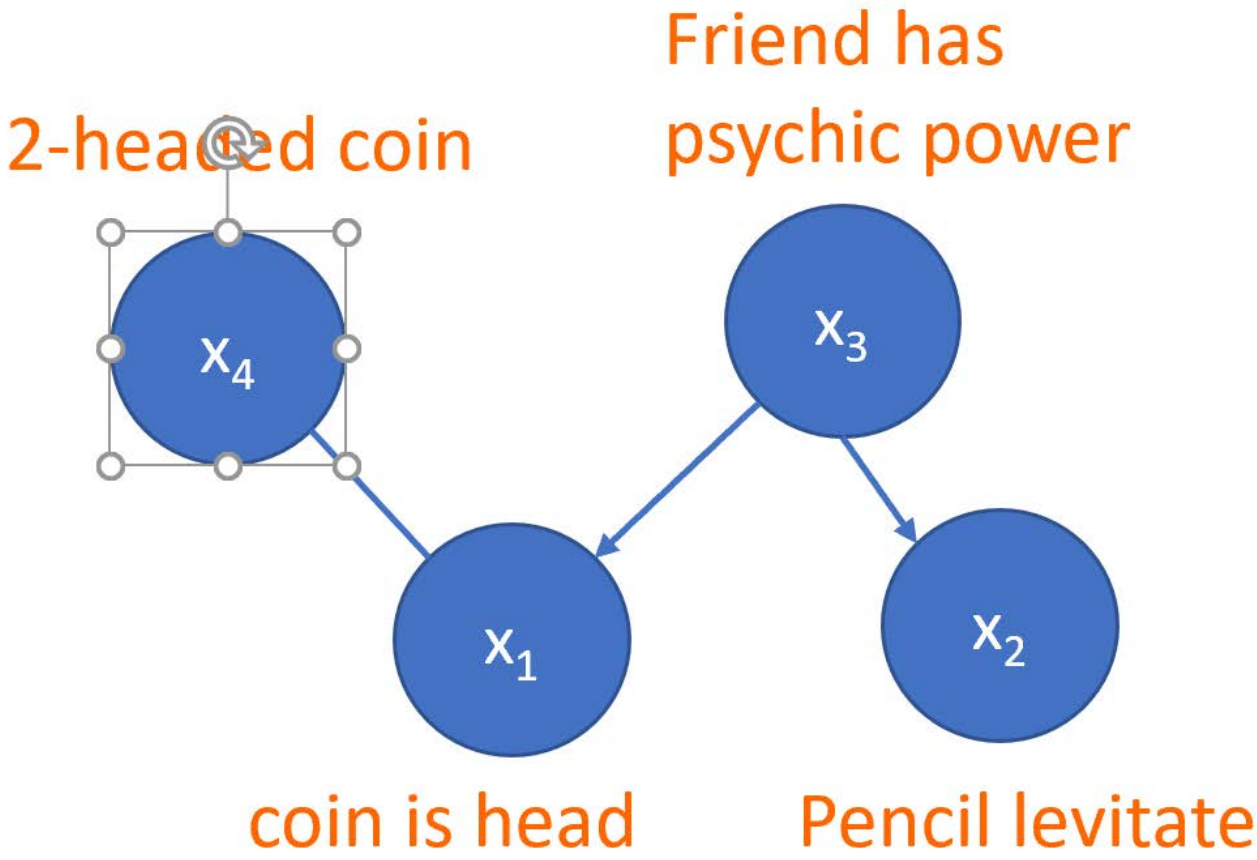

Figure 5: Graphical models showing the relationship between the variables. To confuse the reader, we replace "burglary" with "two-headed coin", "alarm goes off" by "coin is heads", "earthquake" by "friend has psychic powers", and "radio reports earthquake" by "pencil levitates". The story here is that 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 coin toss example can be represented by a graph with nodes $1, 4, 3, 2$. Each node has a binary-valued state variable X_1, X_4, X_3, X_2. If $X_1 = 1$ the coin toss is heads ($X_1 = 0$ if it did not). Similarly the variable X_4, X_3 indicate whether the coin is double headed, your friend has psychic powers or not. $X_2 = 1$ if your friend can levitate the pencil ($X_2 = 0$ is there is not).
- The joint distribution $P(X_1, X_4, X_3, X_2)$ can be decomposed as $P(X_1|X_4, X_3)P(X_4)P(X_3)P(X_2|X_3)$. 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_4, X_3 of the observed data X_1, X_2.
- If you do not ask your friend to levitate a pencil, then the graph is simplified $P(X_1, X_4, X_3) = P(X_1|X_4, X_3)P(X_4)P(X_3)$. Without the evidence from levitation we will probably interpret the head toss as due to a 2-headed coin. The new evidence (from the levitation) 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 $P(X_1, X_4, X_3)$ need? How do we compute the posteriors $P(X_4, X_3|X_1)$?
- A general distribution $P(X_1, X_4, X_3)$ has $2^3 - 1 = 7$ parameters (each variable X takes two possible values, yielding 2^3 but we subtract 1 because of the constraint $\sum_{X_1,X_4,X_3} P(X_1, X_4, X_3) = 1$). But if $P(X_1, X_4, X_3) = P(X_1|X_4, X_3)P(X_4)P(X_3)$ then it takes only $4 + 1 = 1 = 6$ parameters (note $P(X_1|X_4, X_3)$ has 4 parameters, because $\sum_{X_1} P(X_1|X_4, X_3) = 1$ for each possible value of X_4, X_3, and there are four possible values for X_4, X_3. 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 $P(X_4, X_3|X_1)$. This is $P(X_4, X_3|X_1) = \frac{P(X_1,X_4,X_3)}{P(X_1)} = \frac{P(X_1,X_4,X_3)}{\sum_{X_4,X_3} P(X_1,X_4,X_3)} = \frac{P(X_1|X_4,X_3)P(X_4)P(X_3)}{\sum_{X_4,X_3} P(X_1|X_4,X_3)P(X_4)P(X_3)}$.
- To compute the posteriors for earthquakes and burglaries separately, we compute $P(X_3|X_1) = \sum_{X_4} P(X_4, X_3|X_1)$ and $P(X_4|X_1) = \sum_{X_3} P(X_4, X_3|X_1)$. (The general rule is "sum out variables you are not interested in).

Graphical models and explaining away (I)(C)

- Similarly for the full model $P(X_1, X_4, X_3, X_2) = P(X_1|X_4, X_3)P(X_2|X_3)P(X_4)P(X_3)$, we find that there are $4 + 2 + 1 + 1 = 8$ parameters (instead of $2^4 - 1 = 15$).
- The posteriors $P(X_4, X_3|X_1, X_2)$ are computed to be $\frac{P(X_1|X_4,X_3)P(X_2|X_3)P(X_4)P(X_3)}{\sum_{X_4,X_3} P(X_1|X_4,X_3)P(X_2|X_3)P(X_4)P(X_3)}$.

Graphical models and explaining away (II)

- Explaining away explains some visual phenomena. 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 specific 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.

- *Model selection* occurs when we specify several possible models and have to select one of them to explain the data.
- Recall that $P(S|I) = P(I|S)P(S)/P(I)$. The term $P(I) = \sum_S P(I|S)P(S)$ is the probability of the data according to the model. If the model is correct, then $P(I)$ will be high and, if not, $P(I)$ will be low. (Caveat – this is over-simplifying. Usually we also want prior probabilities for each model).
- Suppose we have two models $P_1(S|I) = P_1(I|S)P_1(S)/P_1(I)$ and $P_12(S|I) = P_2(I|S)P_2(S)/P_2(I)$ which compete to explain the image. We select the first model is $P_1(I) > P_2(I)$ and the second if $P_2(I) > P_1(I)$.
- For example, suppose we have two models for using shading to estimate the shape of an object. Both have the sane likelihood terns $P(I|S)$ but the first model has a prior $P_1(S)$ that the shape of the object is piecewise smooth, while the second has a prior $P_2(S)$ that the object is a face. For most stimuli $P_1(I) > P_2(I)$, because most objects are not faces, But if the image is a face, or an inverted face, then $P_1(I) > P_2(I)$, so we will estimate the object to be a face even if it is inverted,

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 sub regions of each image. For example, the lighting and geometry in most images are too complex to make shape from shading a reliable cue except in, at best, limited regions of an image. Model selection can, in principle, be used to address this problem.
- Similarly shape from texture algorithms require assumptions about the texture on an object which are only valid for a very small set of objects. So applying them to complex real world images requires using model selection, or an alternative approach, to decide if shape from texture can be applied to any region.
- 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 regions of the image, if any, the cues are valid.

Model selection illustration

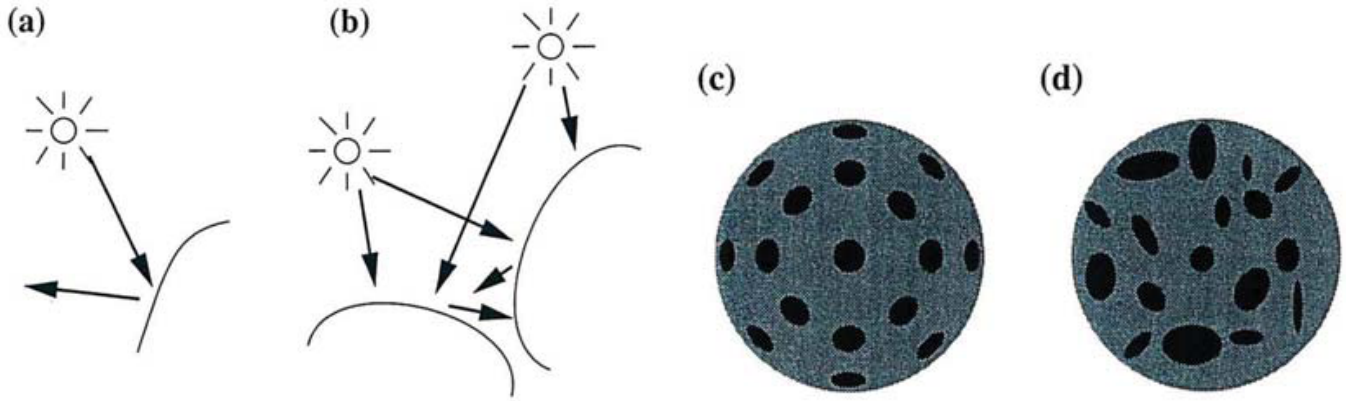

Figure 6: Model selection may need to be applied to decide if a cue can be used. Shape from shading cues will work for case (a) because the shading pattern is simply due to a smooth convex surface illuminated by a single source. But for case (b) the shading pattern is complex – due to mutual reflection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

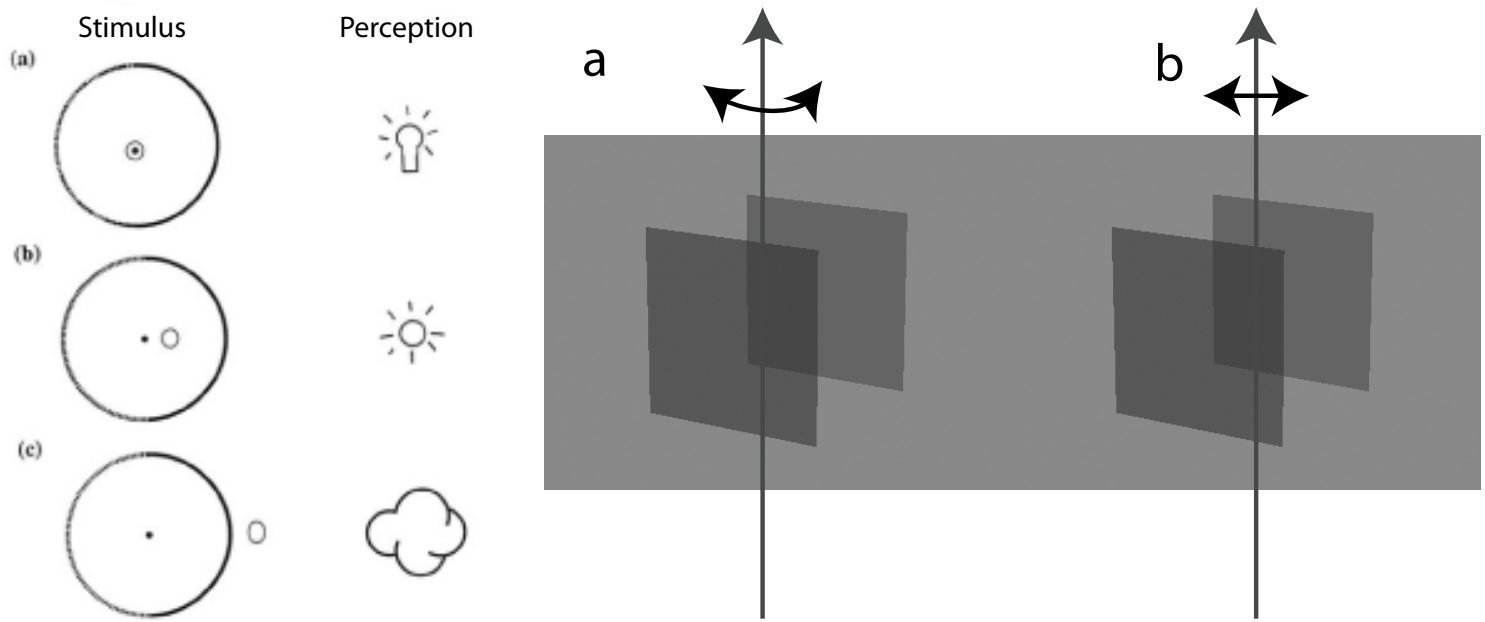

Figure 7: Examples of strong coupling with model selection. Left: A sphere is viewed binocularly, and small changes in the position of the specularity lead to very different percepts (Blake and Bülthoff, 1990). Right: Similarly altering, the transparency of the moving surfaces can make the two surfaces 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.
  [**?**]ecent foundational neural network models, like depth anything, have the ability to estimate shape and depth fairly reliably after being trained on enormous datasets. These are black box methods so it is unclear why they are successful and what cues they exploit. One conjecture is that they exploit some combination of shading, texture, and perspective while also exploit prior knowledge about the shapes of objects.