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

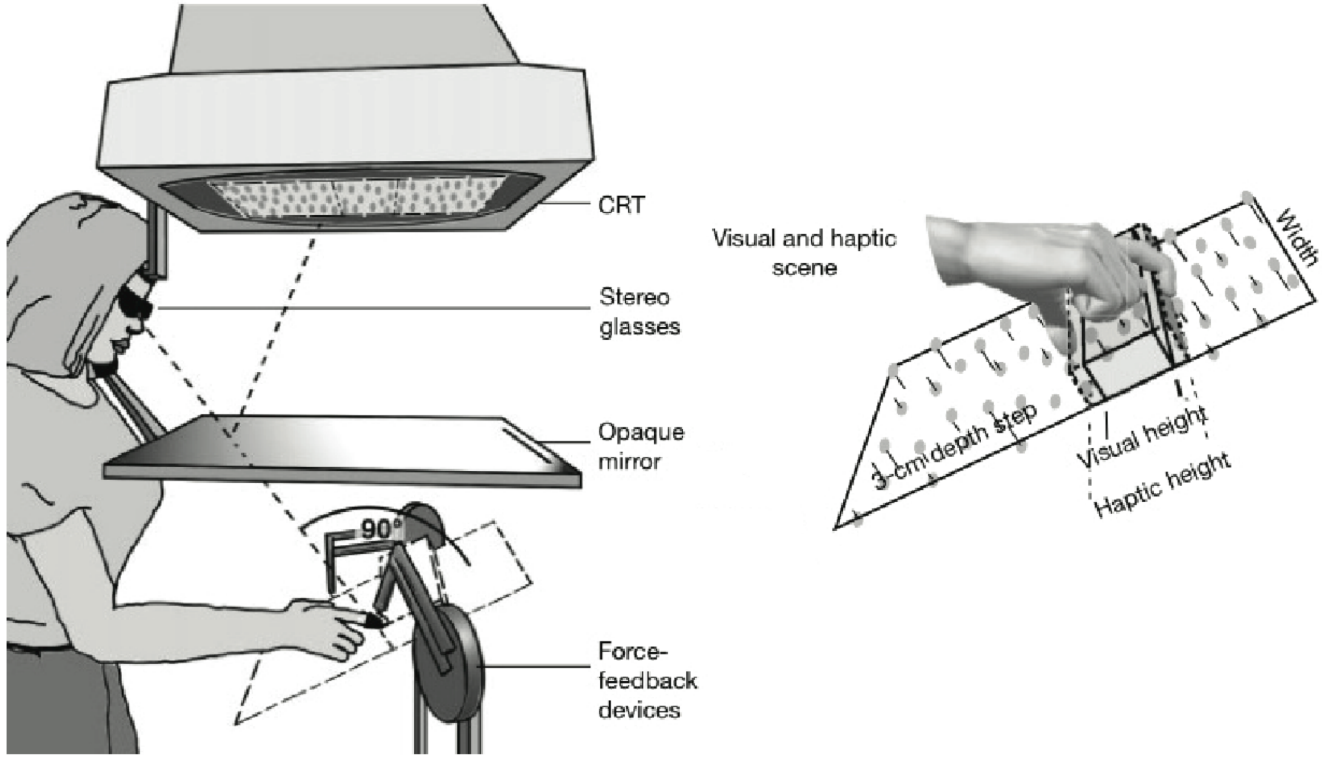

Figure 33: The work of Ernst and Banks shows that cues are sometimes combined by weighted least squares, where the weights depend on the variance of the cues. Figure adapted from Ernst & Banks (2002).

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

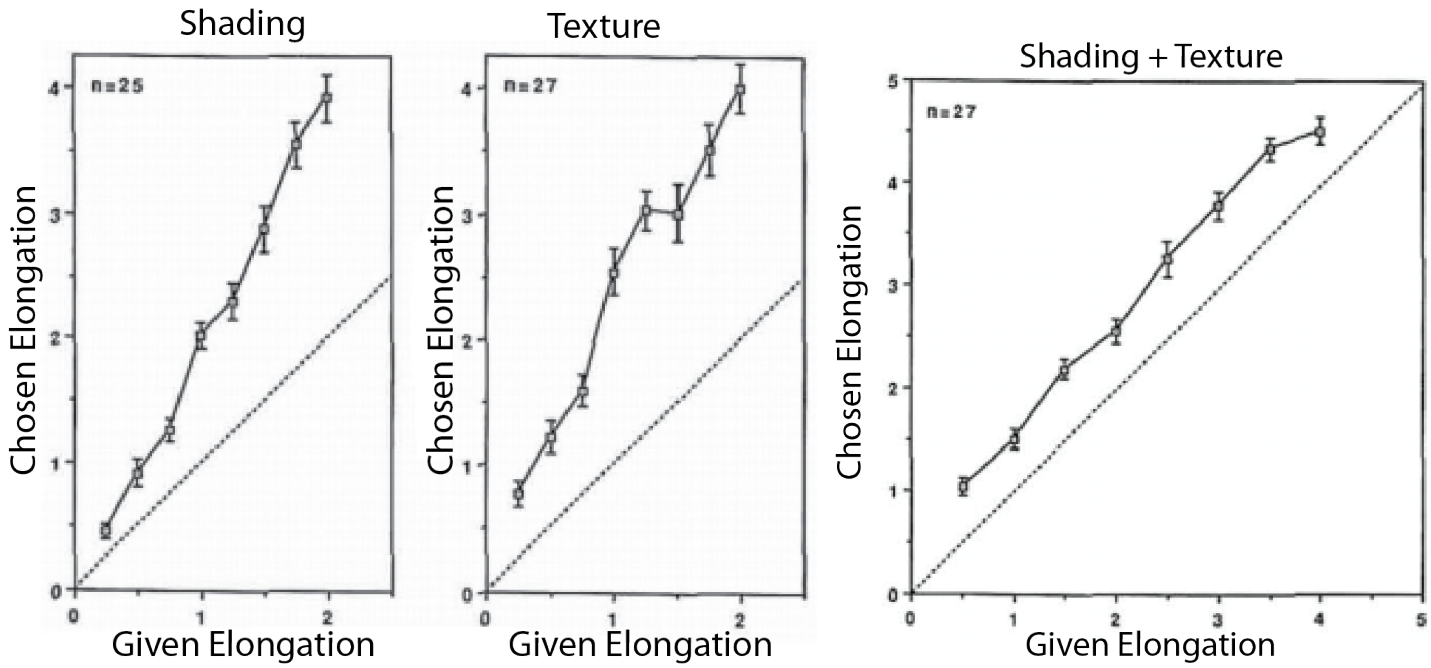

Figure 34: Cue coupling results that are inconsistent with linear weighted average (Bulthoff et al., 1990). Left: If depth is estimated using shading cues only, then humans underestimate the perceived orientation (i.e., they see a flatter surface). Center: Humans also underestimate the orientation if 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 $\vec{S}_2^* = \frac{t_2\vec{C}_2 + t_p\vec{S}_p}{t_1+t_2+t_p}$.

Lecture 12.6

- This lecture discusses the dependencies between visual cues, and 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 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

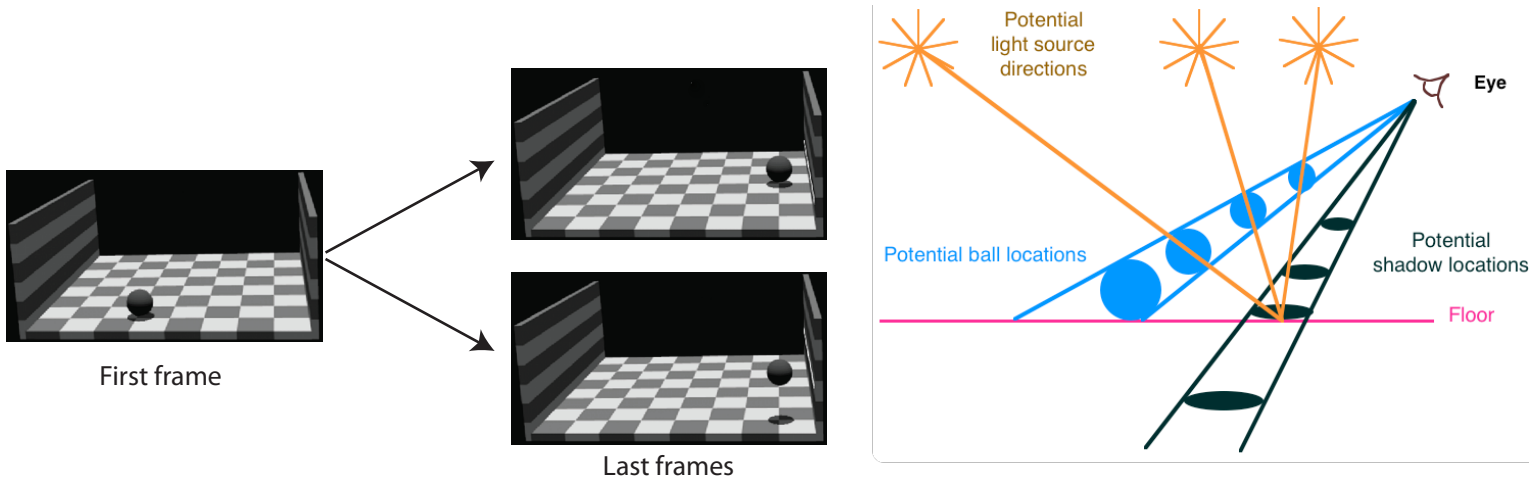

Figure 35: In the "ball-in-a-box" experiments, the motion of the shadow affects the perceived motion of the ball. The ball is perceived to rise from the ground if the shadow follows a horizontal trajectory in the image; but it is perceived to move towards the back of the box if the shadow follows a diagonal 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 Griffiths & Yuille (2006) 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:

$$P(\{X_\mu : \mu \in \mathcal{V}\}) = \prod_{\mu \in \mathcal{V}} P(X_\mu|\vec{X}_{pa(\mu)}). \tag{43}$$