Modeling Cue Integration in Cluttered Environments

Maneesh Sahani and Louise Whiteley

This is the draft, uneditted version of a chapter to published in 2011 by Oxford University Press in the book Sensory Cue Integration*, M. Landy, K. Körding, and J. Trommershäuser, eds.*

Introduction

A complex, cluttered environment is replete with cues, and deciding which of them go together can be at least as great a challenge as properly integrating the ones that do. Although human observers are often imperfect in such settings, they usually outperform machine-perception algorithms on tasks such as demarcating and identifying objects seen against a cluttered background, or following a conversation in a noisy crowd. This failure to build machines capable of matching human performance may be seen as a sign that the field of perceptual studies has not yet found the right way in which to analyze such complex situations. Indeed, one of the largest challenges to making a successful analysis of perception and cue integration in such contexts comes at the outset in setting up an appropriate model. In this chapter we will take some modest strides towards building new cue integration models with the expressive power necessary to capture at least some of these complex settings.

Setting up an appropriate model is, of course, only the first step. Solving the problems of grouping and integration efficiently is still challenging. Indeed, the apparent imperfections of observers might suggest that the problem is, in fact, very difficult in a fundamental sense. We will see that drawing exact inferences within our model is algorithmically complex in a precise quantitative sense, and requires resources in terms of computational hardware or time that grow unfeasibly large for moderate problem sizes. Thus the actual process of inference will require approximation in many real-world settings. We will introduce one very simple such approximation here.

The structure of model we choose will be based deliberately, but loosely, on the sort of spatial 'feature maps' that are seen in neural systems, particularly those involved in vision. This neuromorphic structure is important for at least two reasons. First, and most obvious, it might allow for connections to be built between perceptual modeling at a behavioral level, and the underlying neural function. Second, and perhaps more subtly, the structure of any approximation will very likely be determined by the structure of the representations used. Thus, even to describe human behavior alone we may very well have to construct our models along neurally plausible lines. Indeed we have argued elsewhere (Whiteley, 2008) that the phenomenon of visual attention may be understood as a mechanism that has evolved to refine perceptual approximations within a model such as the one described here.

Integrating two independent cues. In the simplest probabilistic formulation of cue integration, we begin by identifying a physical *attribute* or feature of the environment that an observer might wish to estimate from sensory data. Let us call this attribute a. It might be the size or location of an object, its slant or reflectance, its material or chemical composition, or its relationship to other objects in the environment. The observer now gathers sensory data, which we call $\mathbf{s}$ (the bold symbol representing a vector, or more generally a set). In their raw form these data are extremely numerous — they include the activity of billions of sensory receptors in the eye, ear, nose and throughout the body. For the purposes of analysis, however, we reduce these data to a smaller set of, often scalar, *cues*: c_i. These are functions of the sensory input $\mathbf{s}$, which together carry much or all of the information about the attribute which was present in the sensory data—in a sense, they act much like *sufficient statistics* in the theory of statistical inference[1]

It is important to realize that the computation required to obtain the cue from the sensory activity might be rather involved. For example, the projected aspect ratio of an object might be a cue carrying information about its slant. That aspect ratio is not sensed directly by any of the observer's receptors, but must be computed from the retinal image by a part of the observer's sensory system which is able to extract the boundary of the object. In many cases, this computation of cues from the sensory input appears to introduce "internal noise" to the cue value. There are other cues for slant, for example the gradient in apparent size of texture elements along a surface, and these may be computed by different parts of the sensory system. We often choose to analyze cues which convey independent information about the value of the external attribute. In probabilistic terms, the distributions of the cue values are independent given the attribute value:

$$P(c_1, c_2|a) = P(c_1|a)P(c_2|a)\,, \tag{5.1}$$

where these distributions capture both essential variation in the value of the cue due to uncontrolled aspects of the world, as well as the internal "noise" added by the nervous system. Then, armed with knowledge of the background or "prior" distribution of the attribute value, $P(a)$, cue integration proceeds by Bayes' rule:

$$P(a|c_1, c_2) = \frac{P(c_1, c_2|a)P(a)}{\displaystyle\int da\, P(c_1, c_2|a)P(a)} \propto P(c_1|a)P(c_2|a)P(a)\,. \tag{5.2}$$

This yields the familiar multiplicative form of cue combination, and if the likelihoods[2] $P(c_1|a)$ and $P(c_2|a)$, and the prior $P(a)$ are all Gaussian in form then the mean of integrated estimate will be a linear combination of the mean estimates derived from each cue alone.

This simple framework has been successfully used to describe a great deal of human behavior. It has also been extended

[1] These definitions—of *attribute* as a physical property or feature of the environment that is to be estimated and *cue* as a function of the sensory input that carries information about that property—will be used throughout this chapter to formalize the probabilistic setting.

[2] The probability $P(c|a)$ viewed as a function of a for fixed c is known as a likelihood function.

in many ways, a number of which are discussed at length in the other chapters of this book. Here, we will review two extensions which lay the groundwork for the more elaborately structured model that we will develop in the next section.

Linked attributes. In the first case, consider the situation where the computed value of a cue depends on more than one aspect of the external environment—that is, there are two separate attributes, a_1 and a_2 which combine to determine the value of a single cue. For example, the apparent aspect ratio of a surface depends not only on its slant but also on its true shape. Thus, a more complete analysis of the situation above would require that we consider both attributes a_1, the slant as before and a_2 the veridical aspect ratio. The distribution for c_1 then depends on both, and, specifically, the prior on shapes a_2 may have a substantial impact on the effective likelihood function $P(c_1|a_1)$:

$$P(c_1|a_1) = \int da_2 \, P(c_1|a_1, a_2)P(a_2) \,. \tag{5.3}$$

This influence on the interpretation of cues due to priors over the attributes in the environment is discussed extensively in Chapter 8. The situation is far from uncommon: color cues depend on both the reflectance of a surface and the color of the illuminant; spectral cues to sound location depend on both the position and spectrum of the source; otolithic cues to attitude depend on both head orientation and linear acceleration (this final example being of considerable importance to aeroplane pilots during take off).

Binding uncertainty. In the second case, the observer might extract two cues from the sensory input, but be unsure about whether they provide information about the same external attribute. This situation occurs most commonly when the two cues might have been derived from information about different objects. A simple example, considered in Chapter 2, is when both visual and acoustic information about location are available, but it is not known whether the light and sound came from the same place. In this case, the observer might consider the two possibilities separately. That is, estimates are formed simultaneously under two *models* m_1 and m_2, each associated with a prior probability of being the true situation. In the first, there is only one object in one location, and both cues depend on it:

$$P(c_1, c_2, a_1|m_1) = P(c_1|a_1, m_1)P(c_2|a_1, m_1)P(a_1|m_1) \,. \tag{5.4}$$

In the second, each cue derives from a different source, and therefore a different location.

$$P(c_1, c_2, a_1, a_2|m_2) = P(c_1|a_1, m_2)P(c_2|a_2, m_2)P(a_1|m_2)P(a_2|m_2) \,. \tag{5.5}$$

Inference, for example about the location of the light source, is then performed by averaging over the two possible models:

$$P(a_1|c_1, c_2) \propto P(c_1, c_2|a_1, m_1)P(a_1|m_1)P(m_1) + \int da_2 \, P(c_1|a_1, m_2)P(c_2|a_2, m_2)P(a_1, a_2|m_2)P(m_2) \,. \tag{5.6}$$

Again, this same formal structure applies in more than one situation: examples include inference about the direction of object motion, where local motion cues might be assigned to one or more (possibly transparent) objects; and a host of auditory grouping phenomena, where sound energy must be sorted into auditory streams.

Multiple objects and clutter

A typical realistic environment contains many objects. Each of these is described by a number of different attributes or features, and in each case a number of different potential sensory cues provide information about the objects and their attributes. Extracting cues, and sorting out which one provides information about which attribute may be challenging. Worse, the attributes of one object may influence the interpretation of cues to another. An extreme example is the color of a light source, which affects the interpretation of reflectance cues for all of the surfaces around it. Perhaps less obviously, and yet possibly more commonly, aspects of the form of objects that are placed close to one another, particularly in the visual periphery, may be difficult to resolve even if a single object under the same viewing conditions is easily seen. This phenomenon is known as "crowding" (e.g. Levi, 2008). We will refer to environments in which this sort of cue-interference takes place as "cluttered" (Baldassi et al., 2006; van den Berg et al., 2009).

At first glance, it might seem that the model-averaging approach described above (Eqs. 5.4–5.6) might be extended to describe inference with an arbitrary numbers of objects. For instance, one might first specify a prior over the number of objects present in the scene, where we define an 'object' rather loosely as a source of attributes that generate cues which need integrating. Conditioned on this number of objects one might then specify a distribution over the attributes that will be present, and conditioned on those a distribution over cues. In principle, the cues may provide information about a single attribute value associated with a single object (although we would not know which one); or may provide simultaneous information for more than one attribute, as in the aspect ratio example above; or indeed may provide simultaneous information about attributes for more than one object.

There are, however, a number of difficulties that would be encountered in attempting this approach.

1. First, as in the two-source example, the most natural way to approach inference within such a model structure would be to integrate cues separately for each possible number of objects, and then average the results of inference weighted by the probability of each model. Whilst not unreasonable when there are only two possibilities, this evaluation of different discrete hypotheses seems both behaviorally and neurally implausible when many different object numbers must be considered.

2. Second, closer examination of the model m_2 in our multimodal example, reveals that it does not just specify that there were two objects present. It also assigns the visual cue to one object's location attribute, and the auditory cue to the other's (that is, c_1 depends on a_1 and c_2 on a_2). In this simple case nothing would have been added by also considering the symmetric alternative where object 2 was the light source and object 1 the origin of the sound, but in the general case the mis-assignment of cues to object attributes may have substantial impact. Thus, either we must specify a separate model for each possible assignment of observed cues to objects, or we must consider all such assignments within a model of a given size. In either case, the computational resource needed to consider all possible assignments grows combinatorially in N_{obj}.

3. This situation grows yet more difficult if we attempt to model the interference effects of clutter, for this would require that we consider not only the assignment of cues to objects, but also all the potential linkages between attributes that might influence the interpretation of the cues.

4. Finally, this representation does not make explicit the information needed to work out the correct association of cues. Visual cues, for example, are most likely to be grouped together—and to interfere—if they originate from the

same region of space, auditory cues if their onsets are simultaneous. In its simplest form, this model has no way of encoding such spatial or temporal information, and uncertainty about that information, alongside the cues.

In fact, this final point provides a basis on which to formulate an alternative model structure. Whilst much of the development below would apply to any sensory process—and indeed we will apply it to the multimodal location example below—it will prove more straightforward to focus our development on vision, where the preferred attribute is space.

Sensory maps. In the visual domain, objects, their attributes in the real world, the sensory input, and the low- or mid-level cues extracted from that sensory input are all distributed over visual space. It is therefore natural to replace the discrete attribute and cue values of the combinatorial model with functions of space. We will refer to these functions as 'maps'. Each map must represent both the presence and the values of attributes and cues at each possible location. Since a cluttered environment may contain spatially overlapping objects that possess multiple different attribute values, a simple function that indicates an attribute value at each point in space would be insufficient. For example, the estimated direction of local motion at a point in visual space may be an important cue to the movement of an extended object. However, a single function giving local motion direction would be unable to express the potential absence or indeed simultaneous presence of more than one local motion vector (as might occur for transparent objects; Sahani and Dayan 2003). Instead, each feature map is a function of both location and attribute (when modeling physical features in the environment) or cue (when modeling internal feature representations) value. The value of the function indicates the 'strength' of the attribute or cue with that value at that location. Exactly what we mean by strength depends on the cue or attribute under consideration. For a local motion cue, the 'strength' may correspond to motion energy extracted from the visual input at a point in space. For a color attribute associated with the reflectance of a physical surface, the 'strength' may be the color saturation and what we we have called the attribute value may be the hue. In other settings, there may be no graded 'strength' variable. In this case, the feature map may be represented as a sum of Dirac deltas (or a binary-valued function if discretized). In many cases, the true attribute map will be sparse, with only a few values exhibiting any strength at a few locations, where objects are present. Cue maps, on the other hand, may often be dense, as the computation of a cue at each point in space may yield a non-zero value due to noise even when the corresponding attribute is absent.

It may be valuable at this point to review the roles of the variables involved in the model. The attribute functions represent the veridical value of object attributes at different points in space. Examples might include surface reflectance properties, depth, slant, and veridical motion. These attributes combine to generate sensory input, and this sensory input forms the basis for the computation of the cue (or sensory feature) maps. Examples of these latter maps would be luminance and color opponent information, binocular disparity, orientation and local (aperture-viewed) motion. Each of these can be found by local computations on the retinal images, and each provides a cue to the value of the corresponding attribute at corresponding points in space. Other types of cue involve non-local computation; for example, computing the apparent aspect ratio requires integration across an object's boundaries. These 'high-level' cues are difficult to handle within the framework we are developing, and will be considered only briefly in the final section.

In our discussion of simple cue integration, the perceived location of a flash or sound was a cue just like any other. In our new formulation, space plays a special role. Each sensory map provides information about the location of its associated cues. Thus a flash of light might result in a luminance cue map with a single 'bump' at the corresponding point in space. Information from these different maps must then be combined, with the encoded locations providing a strong signal as to which cues are most likely to go together. In some cases, this integration might be relatively involved. For example, by combining information about orientation and local motion energy across a region of space, one may estimate veridical local motion, resolving the so-called aperture problem. This might allow generalization of motion results such as those of Weiss et al. (2002) to cluttered environments with multiple moving objects (c.f. Weiss, 1998).

Structuring the prior. In the scheme we have sketched, attributes and cues have both been replaced by map functions. But what of the distribution over the number of objects and their locations? Certainly, we could continue to use a 'nested' prior as in the multiple model case, in which the distribution is expressed hierarchically—first a distribution over the number of objects, and then, conditioned on that number, a distribution over their locations. However, it is possible to structure this prior in a way that is closer to the attribute and sensory cue maps. In this view, the priors over the number of objects and their locations are combined into a single distribution, known as a *spatial point process*—a probability distribution over sets of points in space. The number of points in a set drawn from this distribution would correspond to the number of objects, whilst the points themselves correspond to the object locations. Indeed, it is often convenient to represent a draw from a spatial point process as a sum of Dirac delta functions, and this would bring the object representation into essentially the same form as that of the attributes and cues. In practice, we will simulate models in which space has been discretized, and thus the prior will express a probability distribution over a binary vector.

Approximation. The sensory map representation has been chosen to make models of cue combination in cluttered scenes with multiple objects easier to express and analyze. Could it also resolve the combinatorial issues we raised when discussing the hierarchical object-based representation? Unfortunately, these problems are fundamental to the problem of grouping cues, and simply recasting the model does not allow us to escape them. What the representation does do is provide a framework within which it is easier to see how to make approximations (although see work by Lücke and Sahani (2008) for an example of the opposite approach, where a model expressed in terms of maps is approximated by iterating a small number of possible contributory objects). This point is a bit difficult to discuss in depth at the abstract level we have used to this point, but will be looked at in more detail below, once we have developed a more concrete (but simplified) model.

The Model

The previous section has outlined a general but still very abstract scheme for the representation of object locations, attributes and spatially-distributed cues in complex environments with multiple objects and cluttered cues. In this section, we will implement this scheme within a simple discretized model. While this will allow us to simulate some of the essential features of multi-object cue integration, our goal here is not to develop the best possible model within the framework—that is a matter for future research. Instead, we will keep the model (and the accompanying approximation) as simple as possible, with a view to illustrating the representational and modeling capabilities of the framework.

The model rests on a simple 'grid world' consisting of a single, discretized spatial dimension ($\mathbf{x}$) and two discretized feature dimensions labeled α and β. For illustrative purposes, we may think of these as corresponding to orientation (α) and color (β). Fig. 5.1 illustrates a single state of the grid world, and what the two 'attribute maps' that describe the grid world would look like for this state. Four spatial locations, labeled x_1, x_2, x_3, and x_4 each take one value of orientation and color, indicated by the gray entries in the orientation feature map; $a^{\alpha}(\mathbf{x}, \alpha)$, and color feature map; $a^{\beta}(\mathbf{x}, \beta)$. Grayscale values indicate the strength of the attribute, so for orientation this corresponds to contrast and for color this corresponds to luminance. In Fig. 5.1 all contrasts and luminances are equal, as indicated by the shared gray level of the 'on' entries in the maps.

[Fig. 5.1 about here]

Below we describe a simple mathematical model for how states of the grid world produce noisy observations (corresponding to noisy sensory cue maps), and how inverse inference works back from these observations to an approximate posterior belief about the state of the world that caused them. We then show how simulated judgments made from this posterior mimic behavioral results observed in two different experiments.

Fig. 5.2 illustrates the generation of noisy observations in the model. The prior over objects is sparse, so that only a small number of objects—and therefore attributes—will be present at any one time. The location, orientation, and color of the objects present is encoded in a binary vector $\mathbf{u}$. Each '1' element in this vector corresponds to an object with a particular location and pair of attribute values, as depicted in the 3D grid at the left of the figure—the vector in the mathematical model is formed by unwrapping this 3D grid into a long vector of pixels. Two rectangular projection matrices, P^α and P^β, can then be used to obtain two 2D representations, one for each feature type, as shown in the next panel of the figure. The corresponding rasterized vectors indicate the spatial locations of non-zero orientation values ($\mathbf{u}^\alpha$) and color values ($\mathbf{u}^\beta$):

$$\begin{aligned} \mathbf{u}^\alpha &= \mathrm{P}^\alpha \mathbf{u}\,, \\ \mathbf{u}^\beta &= \mathrm{P}^\beta \mathbf{u}\,. \end{aligned} \tag{5.7}$$

[Fig. 5.2 about here]

The binary vectors $\mathbf{u}^\alpha$ and $\mathbf{u}^\beta$ thus indicate the *location* of the peaks in the attribute map functions. The *amplitudes* of these peaks – that is, the strength of the corresponding features – are each drawn independently from a zero-mean Gaussian distribution. Equivalently, we can view the attribute maps themselves[3] as drawn from zero-mean multivariate normal distributions with diagonal covariances U^α and U^β, whose diagonal elements are given by the vectors $\mathbf{u}^\alpha$ and $\mathbf{u}^\beta$:

$$\begin{aligned} \mathbf{a}^\alpha &\sim \mathcal{N}\left(\mathbf{0}, \mathrm{U}^\alpha\right) \quad \text{with} \quad \mathrm{U}^\alpha = \mathrm{diag}[\mathbf{u}^\alpha]\,, \\ \mathbf{a}^\beta &\sim \mathcal{N}\left(\mathbf{0}, \mathrm{U}^\beta\right) \quad \text{with} \quad \mathrm{U}^\beta = \mathrm{diag}[\mathbf{u}^\beta]\,. \end{aligned} \tag{5.8}$$

(The symbol $\sim$ means 'is distributed according to' and $\mathcal{N}(\mu, \Sigma)$ is a normal or Gaussian distribution with mean μ and (co)variance Σ.) Thus, where there is no object with the corresponding attribute value—*i.e.* where there is a '0' entry in $\mathbf{u}^\alpha$ or $\mathbf{u}^\beta$—the value of the corresponding attribute map is zero, as generated by a zero mean, zero variance Gaussian. However, when there is such an object—*i.e.* there is a '1' entry in $\mathbf{u}^\alpha$ or $\mathbf{u}^\beta$—then the attribute value is drawn from a zero mean Gaussian with a variance of 1, thus generating a range of attribute strengths. Note that because the Gaussian is zero mean, high feature strengths may be represented by high positive *or* high negative values. This accords with neurally inspired representations of features in terms of a pair of opposing axes—for example a blue-yellow axis for color or a positive-negative polarity axis for orientation contrast.

[3] In this discretized model, feature maps have of course become feature vectors, but we continue to use the same terminology as the effects of interest are the same.

The third panel in Fig. 5.2 shows an example of attribute feature maps drawn from this distribution, $\mathbf{a}^\alpha$ and $\mathbf{a}^\beta$, where now gray represents 0. The relationship between these and the corresponding cue feature maps $\mathbf{c}$ is modeled very simply. Each attribute vector is multiplied by a corresponding weight matrix Λ, whose effect is to convolve the attribute map with a kernel. The Gaussian shape of kernel, as shown above the arrows linking $\mathbf{a}$ to $\mathbf{c}$ in Fig. 5.2, is inspired by the typical form of a neuronal receptive field and has the effect of smearing out the sparse entries of $\mathbf{a}$. These cue-maps are also affected by internal noise, which we model as an independent normally distributed perturbation with a (diagonal) covariance matrix Ψ. Thus, the mapping from attribute to cue maps can be written:

$$\begin{aligned}\mathbf{c}^\alpha &\sim \mathcal{N}\left(\Lambda^\alpha \mathbf{a}^\alpha, \Psi^\alpha\right) , \\ \mathbf{c}^\beta &\sim \mathcal{N}\left(\Lambda^\beta \mathbf{a}^\beta, \Psi^\beta\right) . \end{aligned} \tag{5.9}$$

An example of observations generated by this process is shown in the rightmost panel of Fig. 5.2. The cue space is taken to be higher dimensional (here represented by a finer discretization) than the attribute space. This transformation from a state of the world, $\mathbf{a}$, to noisy observations, $\mathbf{c}$, can be thought of as representing all the sources of stochasticity that render perceptual inference necessarily probabilistic – including noise in the external world, unreliable neural firing, and coarse response properties. It also mixes attribute values from nearby points in the same cue values, thus providing a very simple model of cue interference in a cluttered scene.

The model can be written more compactly by concatenating the two feature dimensions. First, attribute vectors are generated from binary feature-location vectors through zero mean Gaussians:

$$\mathbf{a} = \begin{bmatrix}\mathbf{a}^\alpha \\ \mathbf{a}^\beta\end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \mathrm{U}\right) \qquad \text{with} \qquad \mathrm{U} = \begin{bmatrix}\mathrm{U}^\alpha & 0 \\ 0 & \mathrm{U}^\beta\end{bmatrix} . \tag{5.10}$$

Second, observed cues are generated from the attribute functions, which are passed through a weight matrix to form the mean of a Gaussian with diagonal noise:

$$\mathbf{c} = \begin{bmatrix}\mathbf{c}^\alpha \\ \mathbf{c}^\beta\end{bmatrix} \sim \mathcal{N}\left(\Lambda\mathbf{a}, \Psi\right) \qquad \text{with} \qquad \Lambda = \begin{bmatrix}\Lambda^\alpha & 0 \\ 0 & \Lambda^\beta\end{bmatrix} , \quad \Psi = \begin{bmatrix}\Psi^\alpha & 0 \\ 0 & \Psi^\beta\end{bmatrix} . \tag{5.11}$$

These equations represent a 'generative model' for noisy cue observations, expressed as an indirect prior on attribute feature maps $p(\mathbf{a}) = \int d\mathbf{u}\, p\left(\mathbf{a}|\mathbf{u}\right) p_0\left(\mathbf{u}\right)$, where p_0 is the sparse prior; and a simple likelihood term $p\left(\mathbf{c}|\mathbf{a}\right)$. Perceptual inference involves inverting the generative model by Bayes' rule, to compute a posterior belief about the true attribute map given the noisy cues it generated:

$$p(\mathbf{a}|\mathbf{c}) \propto p(\mathbf{c}|\mathbf{a})\, p(\mathbf{a}) \tag{5.12}$$

$$\propto p(\mathbf{c}|\mathbf{a}) \int d\mathbf{u}\, p\left(\mathbf{a}|\mathbf{u}\right) p_0\left(\mathbf{u}\right) . \tag{5.13}$$

Written this way, the operation seems no more difficult that of Eq. 5.2. Indeed, it might seem simpler because the vector form obscures the presence of multiple cues and attributes and the binding problem, thus subsuming both simple cue combination (Eq. 5.2) and the two extensions discussed in the Introduction to this chapter (Eqs. 5.3 and 5.6). Has the map-based formulation really resolved the computational difficulties of the combinatorial representation?

The answer, of course, is no. These combinatorial difficulties are fundamental to the multi-object problem. In the map formulation they manifest themselves in the difficulty of computing the integral over $\mathbf{u}$ (or the sum over its values in a discretized settings). The point-process prior embodies knowledge about the sparse distribution of objects made up of spatially co-located features, and thus may be highly structured. For example, in the simple discrete model, it may take the form of a mixture of sparse distributions each with a fixed number of objects present. Each setting of the object vector $\mathbf{u}$ induces a different Gaussian distribution of attribute vectors, and so the net prior distribution on $\mathbf{a}$ is a mixture of Gaussians. If the binary vector $\mathbf{u}$ is n entries long it has 2^n possible settings, giving a mixture of 2^n Gaussians—summing over each of these is indeed intractable.

Have we gained anything at all then? In fact, we have gained two things. First, by expressing all features as functions of space, we have avoided the need for an explicit search over all possible assignments of cues to objects—an operation that had added a further layer of combinatorial complexity to the non-spatial model. Second, as we will see in the next section, the current framework lends itself to a straightforward approximation.

Approximate inference

Many probabilistic models used in machine perception, or as the bases for models of biological perception, have led to intractable problems of inference. As a result, probabilists have devoted substantial effort to finding reliable, fast and accurate approximations to the exact inferential operation. Many different schemes of approximation have been proposed, each with its own advantages and disadvantages. In many cases, the theoretical basis for the approximations may be as involved as the development of the underlying models themselves—or, indeed, more so. As our goal in this chapter is to be illustrative rather than optimal, we will develop an approximation that is particularly simple but not particularly accurate. It is related, albeit distantly, to a family of powerful state-of-the-art approximation algorithms known as Expectation Propagation or EP (Minka, 2001). A full application of EP with iterative refinement of the approximation may, in fact, be as good a deterministic approximation as would be available within the grid world model. On the other hand, our non-iterative version will be didactically helpful, but not at all optimal.

Our approach is to ignore the correlations in the prior, exploiting the Gaussian properties of the generative model to arrive at a Gaussian approximation. We start by noting that, conditioned on the value of $\mathbf{u}$, $\mathbf{a}$ and $\mathbf{c}$ are jointly Gaussian with zero mean:

$$\begin{aligned} p\left(\begin{bmatrix}\mathbf{a}\\ \mathbf{c}\end{bmatrix}\middle|\mathbf{u}\right) &= p\left(\mathbf{c}|\mathbf{a}\right)\ p\left(\mathbf{a}|\mathbf{u}\right) \\ &= \mathcal{N}\left(\Lambda\mathbf{a},\Psi\right)\ \mathcal{N}\left(\mathbf{0},\mathrm{U}\right) \\ &= \mathcal{N}\left(\mathbf{0},\begin{bmatrix}\mathrm{U} & \mathrm{U}\Lambda^{\mathsf{T}}\\ \Lambda\mathrm{U} & \Lambda\mathrm{U}\Lambda^{\mathsf{T}}+\Psi\end{bmatrix}\right)\ . \end{aligned} \tag{5.14}$$

We could, in principle, find $p(\mathbf{a}|\mathbf{c})$ from the joint distribution over $\mathbf{a}$ and $\mathbf{c}$ obtained by integrating Eq. 5.14 with respect to the prior distribution on $\mathbf{u}$. This integral, however, is the exponentially large mixture of zero-mean Gaussians we described above, with one Gaussian for each possible setting of $\mathbf{u}$:

$$p\left(\begin{bmatrix}\mathbf{a}\\ \mathbf{c}\end{bmatrix}\right) = \sum_{\mathbf{u}} p_0(\mathbf{u})\, p\left(\begin{bmatrix}\mathbf{a}\\ \mathbf{c}\end{bmatrix}\middle|\mathbf{u}\right)\ , \tag{5.15}$$

and is, of course, intractable. We therefore approximate the joint posterior by minimizing the Kullback-Leibler (KL) divergence between the true joint distribution above and a Gaussian approximation, $q(\cdot)$. This optimal approximating distribution also has a zero mean, and can be found simply by replacing the covariance matrix U in the conditional distribution (Eq. 5.14) with its average under the prior. We write this average covariance matrix as $\overline{\mathrm{U}}_0$, giving:

$$
\begin{aligned}
q\left(\begin{bmatrix}\mathbf{a}\\ \mathbf{c}\end{bmatrix}\right) &= \operatorname*{argmin}_{\tilde{q}\in\mathcal{N}} \ \mathbf{KL}\left[\sum_{\mathbf{u}} p_0(\mathbf{u})\ \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\mathrm{U} & \mathrm{U}\Lambda^{\mathsf{T}}\\ \Lambda\mathrm{U} & \Lambda\mathrm{U}\Lambda^{\mathsf{T}}+\Psi\end{bmatrix}\right) \ \middle| \ \tilde{q}(\cdot)\right]\\
&= \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\overline{\mathrm{U}}_0 & \overline{\mathrm{U}}_0\Lambda^{\mathsf{T}}\\ \Lambda\overline{\mathrm{U}}_0 & \Lambda\overline{\mathrm{U}}_0\Lambda^{\mathsf{T}}+\Psi\end{bmatrix}\right) \ . \qquad (5.16)
\end{aligned}
$$

The covariance matrix U is diagonal, with '1' entries on the diagonal indicating the presence of a particular feature-location combination, and '0' entries elsewhere. The *average* of each diagonal element under the prior will thus be a number between 0 and 1 equal to the marginal prior probability of generating that particular feature-location combination (*i.e.* the probability of the relevant entry on the diagonal of $\overline{\mathrm{U}}_0$ being 'on').

Eq. 5.16 is a joint normal distribution over $\mathbf{a}$ and $\mathbf{c}$. To obtain the posterior distribution needed to model perceptual inference we must transform this into a conditional distribution on $\mathbf{a}$ given $\mathbf{c}$. This involves some algebra, but is a standard operation of probabilistic calculus and the result can be found in tables of probabilistic identities[4]. The resulting posterior distribution is

$$
q\left(\mathbf{a}|\mathbf{c}\right) \;=\; \mathcal{N}\left(\overline{\mathrm{U}}_0\Lambda^{\mathsf{T}}\left(\Lambda\overline{\mathrm{U}}_0\Lambda^{\mathsf{T}}+\Psi\right)^{-1}\mathbf{c}, \overline{\mathrm{U}}_0 - \overline{\mathrm{U}}_0\Lambda^{\mathsf{T}}\left(\Lambda\overline{\mathrm{U}}_0\Lambda^{\mathsf{T}}+\Psi\right)^{-1}\Lambda\overline{\mathrm{U}}_0\right) \ . \qquad (5.17)
$$

We must emphasize again at this point that Eq. 5.17 is far from being an optimal approximation to Eq. 5.13. Nonetheless, we will use it below to successfully model behavior in a perceptual experiment involving very brief presentations of visual stimuli in which features are sometimes incorrectly bound. In the final section, we will briefly discuss the possibility that exactly this sort of crude approximation may underlie rapid inattentive perception, which is prone to perceptual errors; while more elaborate attentive processes may depend on refining the form of the approximation to match the particular stimulus or task.

Modeling behavior

We now turn to the question of how the framework developed above may be used to model data from two different behavioral experiments, both of which involve the combination of cues from multiple sources. As has been the case throughout the chapter, our goal here is purely didactic. The same data can be, and indeed have been, modeled just as successfully using simpler frameworks—these simpler frameworks were, however, tailored specifically to the particular experimental settings. By showing that similar results can be obtained from a more generic framework we hope to illustrate its power, and to be able to draw useful parallels between disparate phenomena.

[4] The result we require is that if $\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}A & C\\ C^{\mathsf{T}} & B\end{bmatrix}\right)$ then $\mathbf{x}|\mathbf{y} \sim \mathcal{N}\left(CB^{-1}\mathbf{y}, A - CB^{-1}C^{\mathsf{T}}\right)$.

Localizing simple cues

The first experiment we look at concerns a version of the "ventriloquist" effect. When presented simultaneously with a flash of light and a short burst of sound, observers often mislocate the source of the sound in a way that seems to be influenced by the location of the light. This phenomenon was studied by Körding et al. (2007), who argued that the way in which the size of the effect falls off with the separation between the light and sound sources made it unlikely that this was a simple case of observers invariably combining visual and auditory cues to location. Instead, they developed a simple binding model of the type described in the introduction to this chapter, in which an ideal observer estimated the source of the sound, taking into account two possibilities: one, that the sound and light came from the same source, and the second that each originated independently of the other. The estimated location of the sound was then derived from the cues by weighting the effects of both models (Eq. 5.6).

The map-based framework that we have developed in the preceding sections may also be used to model these data, and in this section we will explore how. In fact, the model we will require here is considerably simpler to the one that we have laid out to this point, as we have no need to consider multiple different values of visual or auditory attributes. Each flash of light and burst of sound looks and sounds the same as any other. Thus, the only variable that we need keep track of is space. In this version of the model, then, the binary vector $\mathbf{u}$ extends only over space. A value of '1' indicates that a source (either visual, auditory, or both) is present at the corresponding location. In the simulations below, we choose a simple independent prior on the elements of $\mathbf{u}$. As there is only one possible value for each attribute, the attribute-specific vectors, $\mathbf{u}^{\alpha}$ and $\mathbf{u}^{\beta}$, are identical to $\mathbf{u}$ in this case (that is, the projection matrices P^{α} and P^{β} are both identity matrices). The attribute maps, $\mathbf{a}^{\alpha}$ and $\mathbf{a}^{\beta}$, have the same dimensionality as $\mathbf{u}$ and indicate the "strength" (here, brightness or loudness) of the corresponding feature.

[Fig. 5.3 about here]

To simulate the experiment within this model, we construct attribute maps that contain exactly one source each, positioned in the same way as the visual and auditory stimuli (see Fig. 5.3). Körding et al. used 5 possible locations for the light and the sound, and so the dimension, D, of the attribute map vectors is taken to be 5. As the brightness and loudness of the stimuli used in the experiment did not vary, we can set the strength of a feature that is present to '1' without losing generality. Thus, the simulated attribute maps resemble source maps $\mathbf{u}$. The difference is that although two sources, and therefore two '1's, might be needed in the source vector $\mathbf{u}$, only one of these will appear in each of the corresponding visual and auditory attribute location vectors. The attribute value for the other source will be '0'. These attribute maps are then used to generate noisy cue maps as in the full model, now using one-dimensional Gaussian weights in the matrices Λ^{α} and Λ^{β}. The difference in accuracy between the auditory and visual localization systems is reflected in both different Gaussian extents within the weight matrices, and different degrees of noise (set by Ψ^{α} and Ψ^{β}) in the cue maps.

We then simulate the observer's inferential process as follows. Based on the noisy cue-maps we construct a separate (marginal) posterior distribution over each attribute map, without specific reference to the structure of the experiment. That is, we do not assume during inference that $\mathbf{u}$ has either one or two non-zero entries, nor that $\mathbf{a}^{\alpha}$ and $\mathbf{a}^{\beta}$ each contain exactly one non-zero attribute. The posterior is thus given by

$$\begin{aligned} p(\mathbf{a}^\alpha|\mathbf{c}^\alpha, \mathbf{c}^\beta) &\propto p(\mathbf{c}^\alpha, \mathbf{c}^\beta, \mathbf{a}^\alpha) \\ &= \sum_{\mathbf{u}\in\{0,1\}^D} p(\mathbf{u})\, p(\mathbf{c}^\alpha, \mathbf{c}^\beta, \mathbf{a}^\alpha|\mathbf{u}) \\ &= \sum_{\mathbf{u}\in\{0,1\}^D} p(\mathbf{u})\, p(\mathbf{c}^\alpha|\mathbf{u})\, p(\mathbf{c}^\beta|\mathbf{u})\, p(\mathbf{a}^\alpha|\mathbf{c}^\alpha, \mathbf{u})\,, \end{aligned} \tag{5.18}$$

with the corresponding expression for $\mathbf{a}^\beta$. For each setting of $\mathbf{u}$, the conditional distributions of both the attribute and cue maps are Gaussian, and so the summand is straightforward to evaluate. Writing $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$ for the multivariate normal pdf with mean μ and covariance matrix Σ as before, but now evaluated at $\mathbf{x}$, we have:

$$p(\mathbf{a}^\alpha|\mathbf{c}^\alpha, \mathbf{c}^\beta) = \sum_{\mathbf{u}\in\{0,1\}^D} \pi(\mathbf{u})\, \mathcal{N}\left(\mathbf{a}^\alpha; \Sigma_a^\alpha (\Lambda^\alpha)^\mathsf{T} (\Psi^\alpha)^{-1} \mathbf{c}^\alpha, \Sigma_a^\alpha\right), \tag{5.19}$$

with $(\Sigma_a^\alpha)^{-1} = (\Lambda^\alpha)^\mathsf{T}(\Psi^\alpha)^{-1}\Lambda^\alpha + U^{-1}$, $\pi(\mathbf{u}) = p(\mathbf{u}) \cdot \mathcal{N}\left(\mathbf{c}^\alpha; \mathbf{0}, \Lambda^\alpha U (\Lambda^\alpha)^\mathsf{T} + \Psi^\alpha\right) \cdot \mathcal{N}\left(\mathbf{c}^\beta; \mathbf{0}, \Lambda^\beta U (\Lambda^\beta)^\mathsf{T} + \Psi^\beta\right)$ and $U = \mathrm{diag}[\mathbf{u}]$ as before. Thus, the posterior distribution over the attribute map is given by a mixture of 2^D (here 32) Gaussians. In this case, the small number of possible locations (and the fact that all lights and all sounds are identical) means that exact computation of this posterior is possible, and so we will not need to invoke the approximation scheme for these data.

We now have a modeling choice to make. How should this posterior distribution be translated into a reported location for the sound? The best justified approach would be to define a "cost" associated with each possible answer, and then choose the report that minimizes the expected cost under the calculated posterior. However, it is not obvious which the right cost function should be. Two obvious candidates might be a zero-one match-based cost, where the penalty for any error is the same; or a "squared-error" cost, where the penalty is related to the square of the distance between the report and the true source location. But it is hard to know which of these, or indeed the many other possible cost functions, observers might have employed—particularly as they were given no feedback about performance during the experiment. Thus, we make a slightly more *ad hoc* modeling choice here. We can easily find the mean of the posterior distribution on the attribute map: it is simply the average of the means of the Gaussians within the mixture, weighted by the proportions $\pi(\mathbf{u})$. We take the reported location to be that associated with the peak of this mean function.

[Fig. 5.4 about here]

Fig. 5.4a shows distribution of reported locations in the experimental data of Körding et al., as well the distribution recovered from the simple binding model reported in the same study. The simulated distribution of reported locations resulting from our map-based model appears in Fig. 5.4b. Both models reproduce the essential features of the behavioral data. A characteristic feature of the behavioral data that is difficult to capture in a simple cue-combination model is the way in which the relative bias in the reported sound location falls off as the true separation between the light and sound is increased. This is much easier to explain with a multiple object model, as pointed out by Körding et al. (see Fig. 5.4c). In essence, the chance of mistakenly thinking that there was a single common source for both light and sound is reduced for larger separations, thus reducing the extent of the bias. The results from the present model also display this feature (Fig. 5.4d).

As we said at the outset of this section, our goal in developing a model for these data was purely didactic. In this one experiment we do not expect the predictions of this type of model to differ substantially from that of the simple multiple objects model. The difference, of course, is in potential. Our cues and attributes are spatial maps rather than scalar location variables. This means the inferential model takes on a generic form, without specific reference to the fact that the experiment involved a single sound and light. (We did assume that there were only 5 possible source locations, although this was only a matter of convenience. Similar results are obtained carrying out inference over more possible locations.) As such, we have implicitly integrated over the possibility of any number (up to 5) of sources of each type being present. In practice, in these data this difference will be very small, as these other alternatives carry very little posterior probability. In principle, however, a similar model be used with no changes to handle a much more crowded environment.

Finally, before moving on to model a second experiment, it is worth reviewing some of the modeling choices we have made. First, we used a simple independent prior on the source locations ($p(\mathbf{u})$) with a high appearance probability. The experiment itself did not accord with this prior, and it is unlikely that observers' general experience would either. A more structured prior might therefore be preferable from a modeling standpoint, although determining the appropriate form would itself be something of a challenge. Second, the Gaussian distribution on attribute values, conditioned on the presence of the source, does not describe the current experiment well. In fact, about two-thirds of the objects in the experiment are associated with a 0 value of either loudness or luminance. The model, on the other hand, expects that most objects will exhibit both audible and visible attributes. It is thus arguable that a more appropriate model would include a conditional distribution over attribute values that was sparse, even when a corresponding object is present. Finally, as pointed out above, there are a number of ways in which the posterior distribution over attribute maps may be reduced to a single answer in the simulated experiment. In each case, one can see how to structure the model to more accurately reflect the experiment, and possible the observers' broader experience. However, the simple version we have described suffices to show how the general framework may be applied.

Localizing misbindings

A similar approach may be taken to model a related but slightly more elaborate experiment due to Hazeltine et al. (1997), in which the reported location of an object was conditioned on the misbinding of two visual features. The study of visual feature misbinding or 'illusory conjunctions' dates to an experiment of Treisman and Schmidt (1982) who noted that observers asked to recall a briefly glimpsed display of colored letters would often report having seen one letter with the color of another. Despite some controversy (Donk, 1999; Prinzmetal et al., 2001; Donk, 2001) a number of similar experiments have elicited errors in binding judgments in a number of different ways (see *e.g.* Cohen and Ivry 1989; Ashby et al. 1996; Prinzmetal et al. 1986 as well as Hazeltine et al.). In the experiment we seek to model observers briefly viewed a horizontal array of colored letters, followed by a masking display. They were asked to report the location of the green letter by pointing at the screen, and then to say whether that green letter was an 'O'. In half the trials, the green letter was indeed an O, in a quarter an O of another color appeared somewhere else in the display, and in the remaining quarter no Os were present. Hazeltine et al. were interested in trials where the green letter was misidentified as an O whilst an O of another color was present, and in whether the reported location of the green letter was displaced towards the location of the actual O, suggestive of an illusory conjunction in space (see Fig. 5.6a). A second experiment yielded similar results when the roles of letter identity and color were reversed – subjects reporting the location of the letter O and then whether or not it was green – showing that the role of the two features was symmetric.

[Fig. 5.5 about here]

The model depends on the same sort of 'grid world' as before, this time exploiting the different feature dimensions as well as just location, but therefore requiring approximation. We consider 3 objects placed close together in the center of a field with 9 possible locations[5], each with a different combination of 9 possible feature values. The simulation is illustrated in Fig. 5.5, where the 3 different feature conjunctions are represented by the three white entries in each of the two $\mathbf{u}$ matrices at the far left. In this case the number of possible source configurations is $2^{9\times 9\times 9}$ and exact inference is clearly intractable, despite the apparently modest size of the model. Therefore, inference was performed approximately, with posterior distributions over both feature maps, $\mathbf{a}^{\alpha}$ and $\mathbf{a}^{\beta}$, being inferred according to Eq. 5.17. The means of these posteriors (equal to the modes, $\hat{\mathbf{a}}^{\alpha}$ and $\hat{\mathbf{a}}^{\beta}$, as the distribution is Gaussian) were used to make judgments. A particular discretized value of β (corresponding to "the letter is green" in the experiment) was taken to identify the target. We modeled the reported location of this target as the center of mass along the location dimension, x, for that target value (i.e. in a restricted region of the feature map, as indicated by the red horizontal box in $\hat{\mathbf{a}}^{\beta}$). Each location within the α feature map (each column of $\hat{\mathbf{a}}^{\alpha}$) was then weighted accorded to its value in the restricted region of the β map. We then summed the weighted map over space to obtain a marginal distribution over α, from which the highest mean feature strength could be selected. This perceived feature value was compared to the target value of α (corresponding to "is the green letter an O?") to yield a binary yes or no response.

We found the same displacement effect as reported by Hazeltine et al., illustrated in Fig. 5.6. The plots in Fig. 5.6 include data from trials where the target value of β is *not* co-located with the target value of α—*i.e.* , the green letter was not an O—and show the reported locations of the (green) target for both correct rejections (where the observer correctly says that the green letter was not an O; white circles), and false positives (where the observer incorrectly reports that the green letter *was* an O; black squares). The abscissae of the graphs are oriented so that the the 'distractor' letter (the O) was located to the right of the target on the plot, and so there is an attraction towards the location of the distractor when the observer incorrectly judges that both target feature values came from the same object. Thus, once again the simple model—this time with a simple approximation—is able to capture the essential features of the behavioral data.

[Fig. 5.6 about here]

Final thoughts

In this chapter we have laid out one approach to describing the inferential problem encountered when integrating multiple different cues that may arise from many different objects. By switching representation from a set of discrete single-valued cues to a spatial representation based on attribute and cue 'maps' we were able to naturally model observers' behavior in some simple multi-object and multi-cue settings.

Whilst effective in these settings, the framework is still far from providing a complete description of perceptual inference and integration in cluttered scenes. A first clear shortcoming is that the model is tuned to a particular level of analysis, corresponding to what is often termed 'mid-level' perception, but cues must be integrated at all levels of perception. Second, while the framework makes it somewhat easier to phrase grouping and integration questions at this middle level, it cannot resolve the fundamental intractabilities associated with answering these questions. Thus, the identification of good approximate algorithms, and possibly of approximations that adapt to the task context, is an active area of research.

[5] Hazeltine et al. used an array of 5 letters, but we can obtain similar results using the smaller, three-object array.

Object-based (high-level) cues. The framework we have developed here works best when the cues used for inference are inherently localized in space (in the visual case) or with respect to some other dimension of importance in grouping. We pointed out above that this is not true of all possible cues. For example, finding the apparent aspect ratio of an object requires a non-local computation over the object's boundary, and the resulting cue value is not uniquely identified with any particular position. However, whilst such non-locality is difficult to fit into the precise formulation we have developed, by itself it would not prove a fundamental obstacle to extension.

A more significant distinction may be drawn, on the other hand, between "high-level" cues, which can only be identified once a scene has *already* been parsed into objects, but which then provide information about the properties of those objects; and "mid-level" ones that require pre-segmentation computations on the sensory image, and which provide information which is helpful for the segmentation process itself as well as for the determination of object properties. As can be seen in the other chapters of this book, the integration of these high-level cues often seems to follow principles similar to those used at the lower levels. Despite this fact, their dependence on object segmentation makes it difficult to analyze them in parallel with the pre-segmentation cues. A model that incorporated both would be most naturally structured hierarchically, with the extraction and integration of these high-level cues being performed using the output of the mid-level processing modeled here. This is not to say that high-level cues cannot be integrated with mid-level ones, but that such integration is likely to occur either by the propagation of mid-level information to the higher levels or by constraint-based message passing between layers.

Approximation and attention. We argued that the combinatorial complexity of cue integration in a cluttered environment compels both human observers and machine-based algorithms to rely on approximations to optimal inference. The exact signatures of such approximation in behavior are not easy to tease out. In one sense, they should show up as sub-optimality in perception. But whilst humans certainly appear to make 'mistakes' in perception, it is difficult to know whether these 'mistakes' arise from a mismatch between the stimulus and the observer's expectations, from a mismatch between the observer's goals and the expectations of the experimenter, or from a genuine sub-optimality of processing. On the other hand, one possible signature of this approximation process may come not from 'mistakes' as such, but from the phenomenon of sensory attention.

One way in which human observers appear to be suboptimal is in their inability to process or 'attend' to all aspects of a cluttered scene at once. Intriguingly, attention has often been linked to the problem of feature integration, and some of the largest behavioral and neural effects of attention are seen when stimuli are crowded. Sensory attention is often described as a response to a limitation of some resource. In such accounts, however, the questions of precisely what is that resource, and why it should be limited, often go unanswered.

The sort of analysis we have described here may suggest an answer to this question, as explored by Whiteley (2008) in her Ph. D. thesis. The limited resource is *computational*: it is the capacity to perform inference within combinatorially complex settings. To do so optimally would require either physical resources or time that would grow combinatorially in the number of possible objects that must be considered. Faced with limited physical systems and with the need to perceive and act rapidly, observers have evolved to approximate. However, a single fixed approximation of the sort that was considered here may not be ideal. A more refined approach would be to tailor the approximation to the job at hand: that is, to adjust it with both the current sensory environment and the current task set. Indeed, a number of different approximations may be attempted one after the other, to achieve a sort of serial search. In this view it is this process of adapting the approximation that is the effect of, and reason for, sensory attention.

Bibliography

Ashby, F. G., Prinzmetal, W., Ivry, R., and Maddox, W. T. (1996). A formal theory of feature binding in object perception. *Psychological Review*, 103(1):165–192.

Baldassi, S., Megna, N., and Burr, D. C. (2006). Visual clutter causes high-magnitude errors. *PLoS Biology*, 4(3):e56.

Cohen, A. and Ivry, R. (1989). Illusory conjunctions inside and outside the focus of attention. *Journal of Experimental Psychology: Human Perception and Performance*, 15(4):650–663.

Donk, M. (1999). Illusory conjunctions are an illusion: The effects of target-nontarget similarity on conjunction and feature errors. *Journal of Experimental Psychology: Human Perception and Performance*, 25(5):1207–1233.

Donk, M. (2001). Illusory conjunctions die hard: a reply to Prinzmetal, Diedrichsen, and Ivry (2001). *Journal of Experimental Psychology: Human Perception and Performance*, 27(3):542–546.

Hazeltine, R. E., Prinzmetal, W., and Elliott, W. (1997). If it's not there, where is it? Locating illusory conjunctions. *Journal of Experimental Psychology: Human Perception and Performance*, 23(1):263–277.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE*, 2(9):e943.

Levi, D. M. (2008). Crowding–an essential bottleneck for object recognition: a mini-review. *Vision Research*, 48(5):635–54.

Lücke, J. and Sahani, M. (2008). Maximal causes for non-linear component extraction. *Journal of Machine Learning Research*, 9:1227–1267.

Minka, T. (2001). Expectation propagation for approximate Bayesian inference. In Breese, J. A. and Koller, D., editors, *UAI*, volume 17, pages 362–369, Washington, USA. Morgan Kaufman.

Prinzmetal, W., Diedrichsen, J., and Ivry, R. (2001). Illusory conjunctions are alive and well: A reply to Donk (1999). *Journal of Experimental Psychology*, 27(3):538–541.

Prinzmetal, W., Presti, D. E., and Posner, M. I. (1986). Does attention affect visual feature integration? *Journal of Experimental Psychology: Human Perception and Performance*, 12(3):361–369.

Sahani, M. and Dayan, P. (2003). Doubly distributional population codes: Simultaneous representation of uncertainty and multiplicity. *Neural Computation*, 15(10):2255–2279.

Treisman, A. and Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14(1):107–141.

van den Berg, R., Cornelissen, F. W., and Roerdink, J. B. T. M. (2009). A crowding model of visual clutter. *Journal of Vision*, 9(4):1–11.

Weiss, Y. (1998). *Bayesian motion estimation and segmentation*. PhD thesis, Massachusetts Institute of Technology.

Weiss, Y., Simoncelli, E. P., and Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598–604.

Whiteley, L. E. (2008). *Uncertainty, Reward, and Attention in the Bayesian Brain*. PhD thesis, University College London.

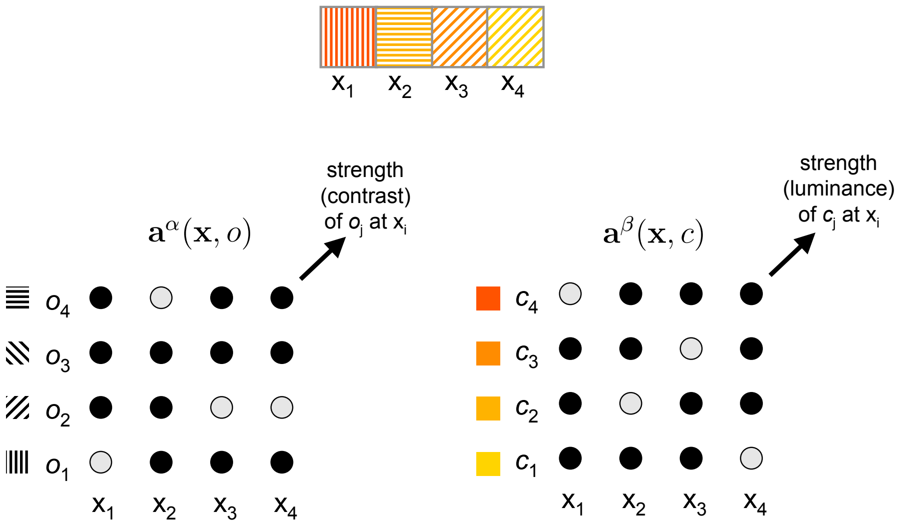

x_1 x_2 x_3 x_4
$\mathbf{a}^{\alpha}(\mathbf{x}, o)$
strength (contrast) of o_j at x_i
o_4
o_3
o_2
o_1
x_1 x_2 x_3 x_4
$\mathbf{a}^{\beta}(\mathbf{x}, c)$
strength (luminance) of c_j at x_i
c_4
c_3
c_2
c_1
x_1 x_2 x_3 x_4

Figure 5.1: **'Grid world' setting for simulations.** Illustrates the simple 'grid world' in which the simulations take place. The picture at the top represents a state of the environment, with oriented colored patches at each of 4 locations. The two arrays below represent the orientation $\mathbf{a}^{\alpha}$ and color $\mathbf{a}^{\beta}$ attribute maps corresponding to this state. Each attribute map indicates the strength, indicated by the gray level of the corresponding circle, with which each feature value (o or c) is present at the corresponding location.

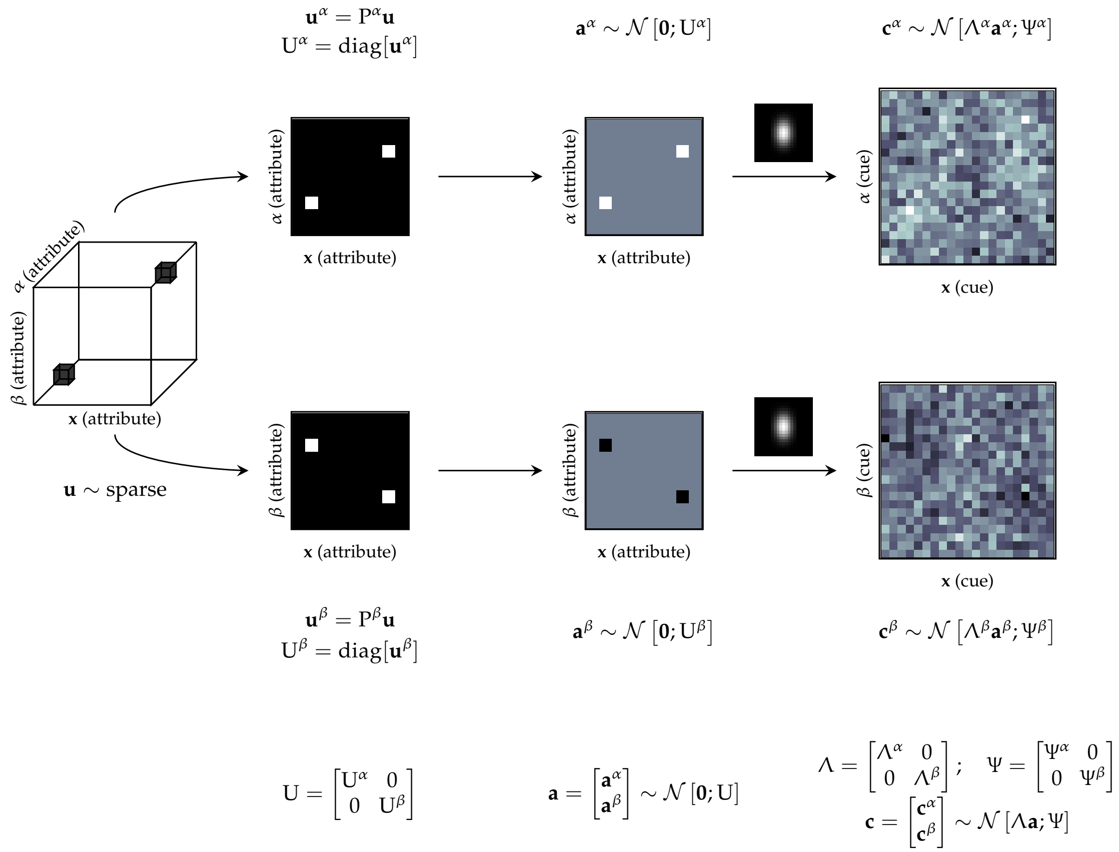

$\mathbf{u}^\alpha = \mathrm{P}^\alpha\mathbf{u}$
$\mathrm{U}^\alpha = \mathrm{diag}[\mathbf{u}^\alpha]$
$\mathbf{a}^\alpha \sim \mathcal{N}\left[\mathbf{0}; \mathrm{U}^\alpha\right]$
$\mathbf{c}^\alpha \sim \mathcal{N}\left[\Lambda^\alpha\mathbf{a}^\alpha; \Psi^\alpha\right]$
α (attribute)
β (attribute)
x (attribute)
α (cue)
x (cue)
β (cue)
$\mathbf{u} \sim$ sparse
$\mathbf{u}^\beta = \mathrm{P}^\beta\mathbf{u}$
$\mathrm{U}^\beta = \mathrm{diag}[\mathbf{u}^\beta]$
$\mathbf{a}^\beta \sim \mathcal{N}\left[\mathbf{0}; \mathrm{U}^\beta\right]$
$\mathbf{c}^\beta \sim \mathcal{N}\left[\Lambda^\beta\mathbf{a}^\beta; \Psi^\beta\right]$
$\mathrm{U} = \begin{bmatrix} \mathrm{U}^\alpha & 0 \\ 0 & \mathrm{U}^\beta \end{bmatrix}$
$\mathbf{a} = \begin{bmatrix} \mathbf{a}^\alpha \\ \mathbf{a}^\beta \end{bmatrix} \sim \mathcal{N}\left[\mathbf{0}; \mathrm{U}\right]$
$\Lambda = \begin{bmatrix} \Lambda^\alpha & 0 \\ 0 & \Lambda^\beta \end{bmatrix}; \quad \Psi = \begin{bmatrix} \Psi^\alpha & 0 \\ 0 & \Psi^\beta \end{bmatrix}$
$\mathbf{c} = \begin{bmatrix} \mathbf{c}^\alpha \\ \mathbf{c}^\beta \end{bmatrix} \sim \mathcal{N}\left[\Lambda\mathbf{a}; \Psi\right]$

Figure 5.2: **The generative model of 'grid world'** This schematic illustrates the generation of noisy observations, $\mathbf{c}$, from an underlying state of the world, $\mathbf{u}$, in the simple grid world in which simulations take place. The white squares in $\mathbf{u}^\alpha$ and $\mathbf{u}^\beta$ indicate which feature-location pairs are 'on', and the strength of these features (for example, contrast for orientation and luminance for color) is generated from a zero-mean Gaussian distribution, producing attribute maps $\mathbf{a}^\alpha$ and $\mathbf{a}^\beta$. In the representations of the maps, gray indicates zero strength, and black and white indicate the extremes of an opponent axis. To generate noisy observed cues, $\mathbf{a}^\alpha$ and $\mathbf{a}^\beta$ are multiplied by Gaussian weight matrices, Λ^α and Λ^β, which smear out the attribute map entries. One component of Λ is illustrated above the arrows that join $\mathbf{a}$ to $\mathbf{c}$, and the matrix consists of one such component centered on each location-feature pair. Independent Gaussian noise, Ψ, then corrupts the cues $\mathbf{c}^\alpha$ and $\mathbf{c}^\beta$.

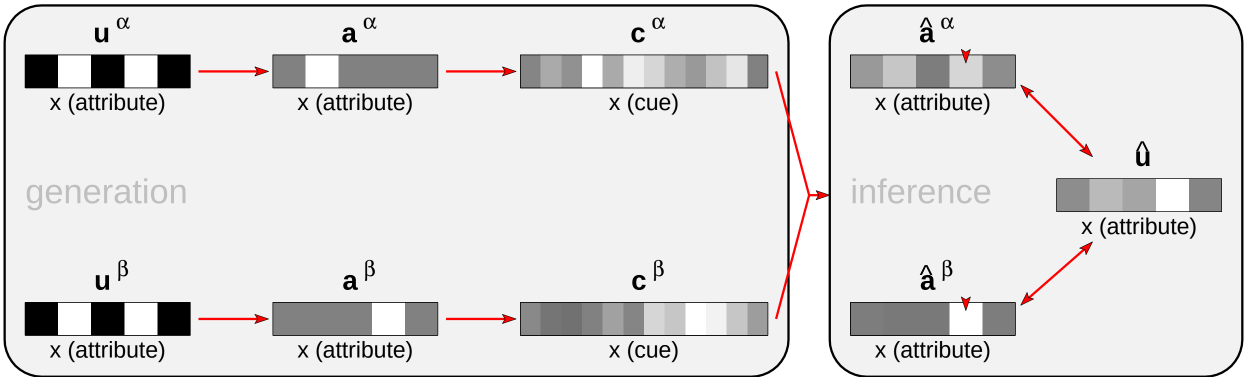
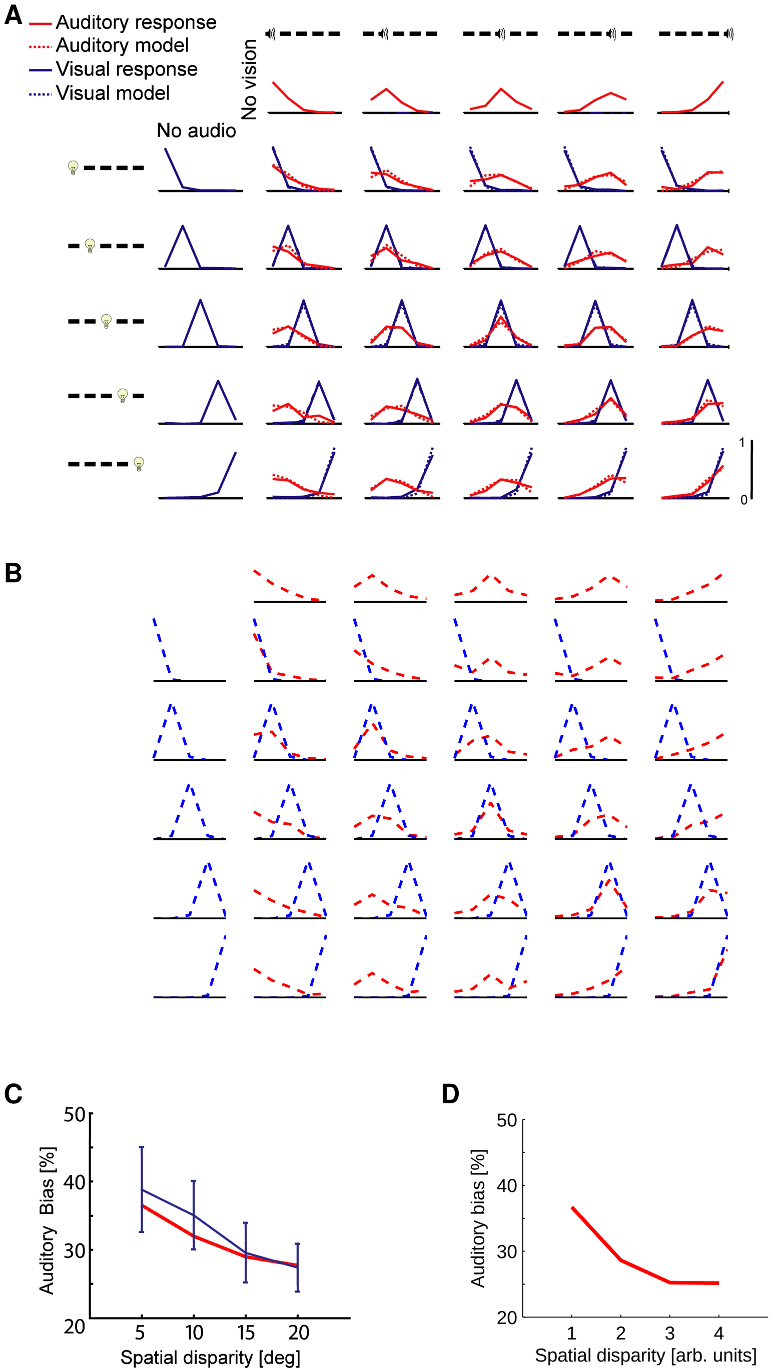


Figure 5.3: **Visual and auditory stimuli: example observations and inference. Generation** of one simulated example of attribute and cue maps, each a function of space ($\mathbf{x}$) alone, with the sound source at position 2 and light at position 4. Feature-presence maps ($\mathbf{u}$) both reflect both sources, but for each feature (sound, α, or light, β) only one source has a non-zero attribute value ($\mathbf{a}$). Cue maps ($\mathbf{c}$) are more densely sampled and noisy. The auditory cue ($\mathbf{c}^{\alpha}$) is more extensively smeared and more noisy than the visual. Red arrows show the flow of dependence. **Inference** of attribute and source maps given the cues. The shaded maps represent the mean of the inferred distributions over each corresponding variable, where in each case the other two variables have been integrated or summed over (e.g. Eq. 5.19). Bidirectional red arrows reflect the interdependence of the estimates. Red arrowheads over the mean estimated attribute maps indicate the locations of the maxima, which correspond to the simulated reported locations. Thus, in this simulation, the sound source was mislocalized to the position of the light.

Figure 5.4 *(following page)*: **Combining visual and auditory stimuli. A,** (Fig. 2c of Körding et al. 2007). Observers were asked to indicate at which of five possible locations they saw a light or heard a sound. Each graph represents one configuration of true light and sound source locations: the sound source location in each column (and its absence in the left-most column) is indicated by the diagrams at the top; the light source location in each row (and its absence in the top row) in indicated by the diagrams on the left. Solid lines show the distribution of observers' responses across the five alternatives in each case—reported sound source locations in red, and light locations in blue. Dotted lines show the corresponding distributions for a simple binding model simulated by Körding et al.. **B,** The corresponding distributions, arranged and colored in the same way, generated by simulations of the map-based model developed here. **C,** (Fig. 2e of Körding et al. 2007). The average auditory bias, defined as the (signed) error in location of the sound source divided by the (signed) displacement between sound and light sources, and shown as a function of the distance between the sound and light sources. Experimental data are in blue and the prediction of the binding model in red. **D,** The predicted bias as a function of sound and light separation for the map-based model.


A
Auditory response
Auditory model
Visual response
Visual model
No vision
No audio
1
0
B
C
Auditory Bias [%]
50
40
30
20
5
10
15
20
Spatial disparity [deg]
D
Auditory bias [%]
50
40
30
20
1
2
3
4
Spatial disparity [arb. units]

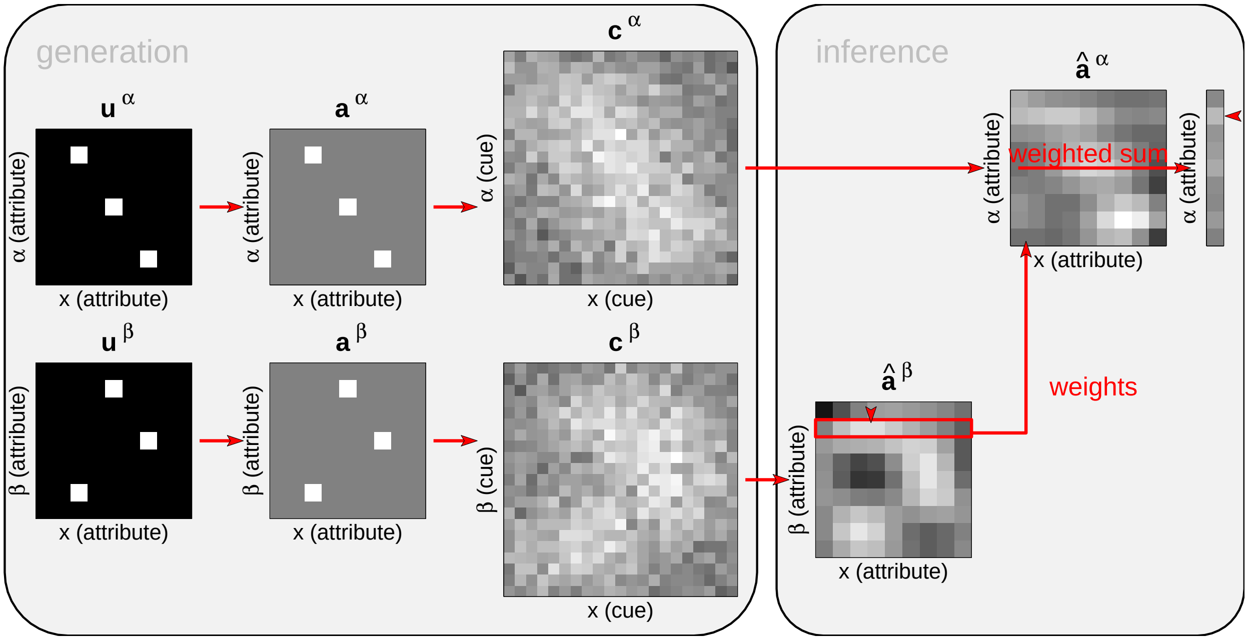


Figure 5.5: **Illusory conjunctions: example observations and inference. Generation** of one simulated example of attribute and cue maps. The attribute maps indicate three letters in the display, each with a unique combination of color (β) and letter identity (α), which generate smeared and noisy cue maps as in Fig. 5.2. **Inference** of attribute maps, reported location and binding, given the noisy cues. Mean attribute maps ($\hat{\mathbf{a}}$) are estimated under the approximation of Eq. 5.17. The red box over $\hat{\mathbf{a}}^{\beta}$ indicates the instructed color to be located ('green'); red arrowhead indicates the reported location (thresholded center-of-mass). The marginal false color map to the right of $\hat{\mathbf{a}}^{\alpha}$ shows the summed attribute strength, weighted by the values of $\hat{\mathbf{a}}^{\beta}$ within the red box. The red arrowhead indicates the reported letter identity. In this simulation the green letter was mislocalized towards the left, and mis-associated with the identity of the leftmost letter.

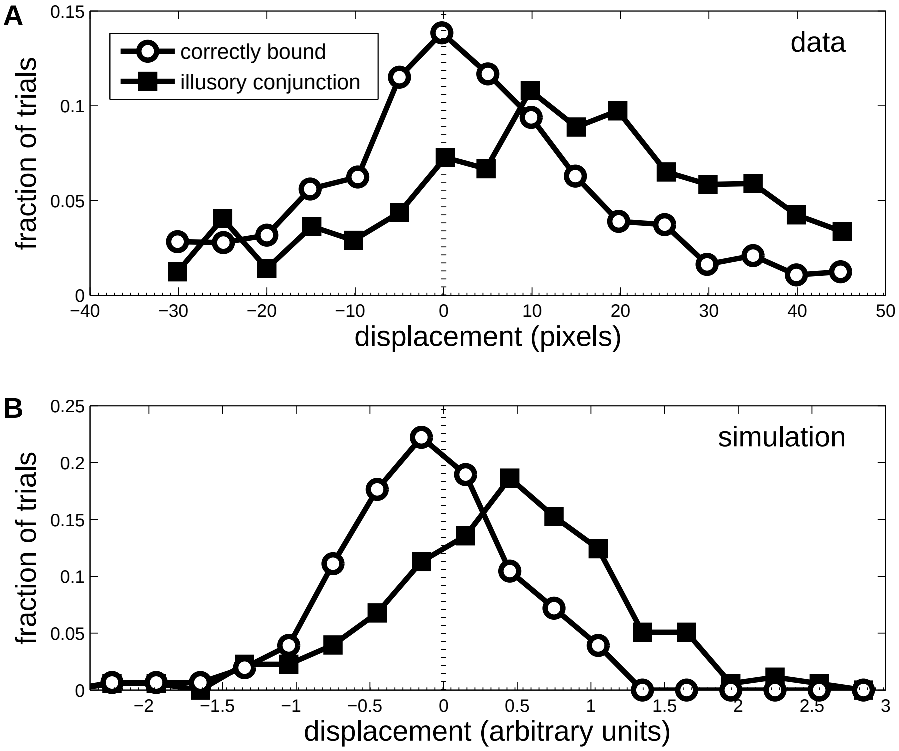


Figure 5.6: **Localization of illusory conjunctions. A,** Results from Hazeltine et al. (1997), replotted from their Fig. 1. Observers were asked to locate the green letter, and then say if it was the letter O—trials shown here are those in which the green letter was not an O, but an O of another color was present in the display (at +28 pixels). Graphs show binned histograms of reported locations when observers incorrectly identified the green letter as an O (black squares) and when they correctly reported that it was not an O (white circles). The distribution of reported locations of misbound letters is displaced towards the location of the 'distractor' O. **B,** Results of the map-based model simulation, sorted and binned as in **A**. The distractor is at position 2. There is less dispersion in responses, but the bias in reported locations associated with misbinding is evident.