Sensory Integration and Kalman Filtering

Robert Jacobs
Department of Brain & Cognitive Sciences
University of Rochester
Rochester, NY 14627, USA

October 26, 2008

Perceptual environments tend to be redundant—we receive sensory signals from multiple sensory modalities including vision, audition, olfaction, and haptics (touch). Redundancy also arises within individual sensory modalities. For example, there are many visual cues to depth and shape such as stereo, motion, texture, and shading. What do we do with all this information?

Perhaps the simplest and most commonplace sensory integration rule studied in the literature is a linear rule: one's percept based on signals from multiple sensory modalities is a linear combination of the percepts based on each of the signals from the individual sensory modalities.

For the sake of concreteness, let's think about the problem of estimating visual depth. A commonly assumed framework for how an observer might go about judging the depth of a visual object defined by multiple visual cues is the following two-stage process. First, depth estimates based on individual cues are derived. Next, a weighted combination of these estimates is calculated and used as the observer's composite depth percept; the cue weights are based on the relative reliabilities of the cues in the current visual context. For example, consider an observer judging the depth of an object defined by motion and texture cues. During stage one, the observer calculates depth estimates based on each individual cue. Let $d_M(m)$ denote the observer's depth-from-motion estimate, and let $d_T(t)$ denote the observer's depth-from-texture estimate. During stage two, the observer combines these estimates into a unified depth percept, denoted $d(m,t)$, using, for instance, a linear cue combination rule:

$$d(m,t) = w_M \; d_M(m) \; + \; w_T \; d_T(t) \tag{1}$$

where the linear coefficients for the motion and texture cues, denoted w_M and w_T, are chosen based on the estimated reliabilities of these cues.

But what is "cue reliability"? There are several ways of defining cue reliability in the perceptual literature. For the purposes of this note, we will use one particular definition: A cue is reliable if the distribution of inferences based on that cue has a small variance.

Let's think about visual depth perception again. It is important to realize that **every visual cue is ambiguous**. There are many reasons underlying this ambiguity, including physical factors, such as atmospheric or optical blurring, and biological factors, such as noise inherent to human nervous systems. Therefore, there is no "correct" interpretation of a cue. Consider an observer viewing a coffee cup. The visual environment provides many cues to the shape and depth of this cup. For the moment, let's consider one particular cue, such as a shading cue, to the depth of the coffee cup (this depth may be defined, for instance, as the distance from the point on the cup closest to the observer to the point furthest away). The horizontal axis of the graph in Figure 1 gives possible

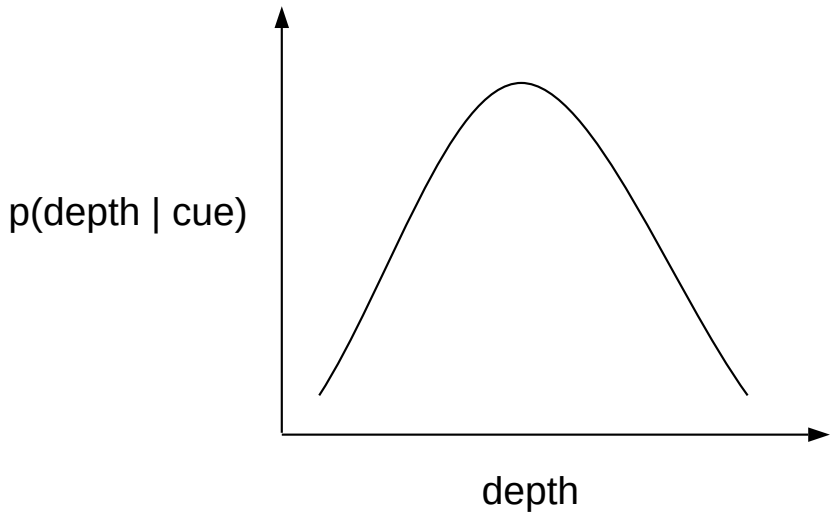

Figure 1: The visual environment typically provides many cues to the depth of an object. For the moment, let's consider one particular cue. The horizontal axis of this graph gives possible values of the object depth, and the vertical axis gives the conditional probability of each possible depth conditioned on the value of the cue. Note that the cue is ambiguous with respect to object depth because it is consistent, to a lesser or greater degree, with many possible depths.

values of the depth, and the vertical axis gives the conditional probability of a depth conditioned on the value of the cue. This probability distribution is not a delta function; that is, the cue is not consistent with one, and only one, depth value. Rather, the cue is consistent, to a lesser or greater degree, with a range of depth values. If the variance of the probability distribution is relatively small, then the observer may tend to believe that the cue is reliable because the cue specifies the cup's depth as lying within a narrow range. Consequently, he or she should make extensive use of the information provided by this cue. If, on the other hand, the variance is relatively large, then the observer may tend to believe that the cue is unreliable because it is consistent with many possible depths. In this case, the observer may tend to ignore the information provided by this cue, or at least discount the information provided by this cue relative to the information provided by other, more reliable, cues. An observer that follows the logic outlined here would be acting in accord with a mathematical model known as a Kalman filter, as described below.

Let's make this more concrete. According to our model, more reliable cues are assigned a larger weight in a linear cue combination rule, and less reliable cues are assigned a smaller weight. Continuing with our example from above, let d denote a possible depth of a visual object, and let m and t denote the values of the motion and texture cues. In addition, let d_m^* denote the optimal estimate of visual depth based solely on the motion cue [this is the depth d that maximizes the probability of a depth value given the motion cue, $P(d|m)$], let d_t^* denote the optimal depth estimate based solely on the texture cue [the depth d that maximizes $P(d|t)$], and let d^* denote the optimal depth estimate based on both motion and texture cues [the depth d that maximizes $P(d|m,t)$]. Given certain mathematical assumptions, Bayes' rule can be used to show the following result:

$$d^* = w_m \, d_m^* \; + \; w_t \, d_t^* \tag{2}$$

where

$$w_m = \frac{\frac{1}{\sigma_m^2}}{\frac{1}{\sigma_m^2} + \frac{1}{\sigma_t^2}} \quad \text{and} \quad w_t = \frac{\frac{1}{\sigma_t^2}}{\frac{1}{\sigma_m^2} + \frac{1}{\sigma_t^2}}, \tag{3}$$

and σ_m^2 and σ_t^2 are the variances of the distributions $P(d|m)$ and $P(d|t)$ respectively. This version of the Kalman filter (Equations 2 and 3) has several appealing properties. First, the optimal estimate of depth based on both motion and texture cues is a linear combination of the optimal estimates based on the individual cues. Second, the linear coefficients, the weights w_m and w_t, are non-negative and sum to one. Lastly, the weight on a cue, such as the motion weight w_m, is large when the cue is relatively reliable (the variance σ_m^2 is smaller than the variance σ_t^2), and small when the cue is relatively unreliable (σ_m^2 is larger than σ_t^2). This point is illustrated in Figure 2.

I hope that you now have good intuitions about sensory integration according to our simple model. My goal in the rest of this note is to mathematically justify this model. Mathematically, a good place to start thinking about sensory integration and Kalman filtering is the following simple scenario. Imagine that you want to estimate the mean of a Normal distribution (whose variance is known) based on a single sensory observation. Let θ denote the mean, σ^2 denote the variance, and x denote the observation. Then the likelihood distribution is

$$p(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\theta)^2\}. \tag{4}$$

You have prior beliefs about the mean of the distribution which you characterize via a Normal prior distribution:

$$p(\theta) = \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\}. \tag{5}$$

Using Bayes' rule, the posterior distribution of the mean θ is

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{p(x)}. \tag{6}$$

Note that $p(\theta|x)$ is a function of θ, but $p(x)$ is a constant which does not depend on θ, meaning

$$\begin{aligned}
p(\theta|x) &\propto p(x|\theta)\, p(\theta) && (7)\\
&\propto \exp\{-\frac{1}{2\sigma^2}(x-\theta)^2\}\ \exp\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\} && (8)\\
&= \exp\{-\frac{1}{2\sigma^2}(x^2 - 2x\theta + \theta^2) - \frac{1}{2\tau_0^2}(\theta^2 - 2\theta\mu_0 + \mu_0^2)\} && (9)\\
&= \exp\{-\frac{x^2}{2\sigma^2} - \frac{1}{2}\theta^2(\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}) + \theta(\frac{x}{\sigma^2} + \frac{\mu_0}{\tau_0^2}) - \frac{\mu_0^2}{2\tau_0^2}\} && (10)\\
&= \exp\{-\frac{1}{2}\theta^2(\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}) + \theta(\frac{x}{\sigma^2} + \frac{\mu_0}{\tau_0^2})\}\ \exp\{-\frac{x^2}{2\sigma^2} - \frac{\mu_0^2}{2\tau_0^2}\} && (11)\\
&\propto \exp\{-\frac{1}{2}\theta^2(\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}) + \theta(\frac{x}{\sigma^2} + \frac{\mu_0}{\tau_0^2})\} && (12)
\end{aligned}$$

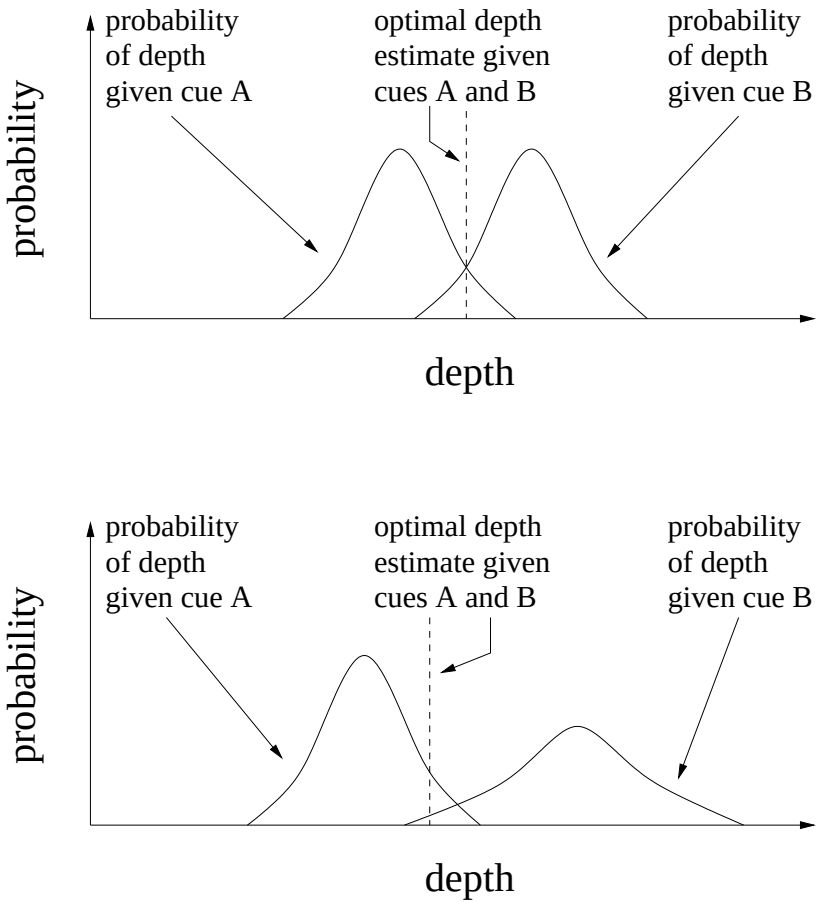

Figure 2: Consider a Kalman filter estimating the depth of an object based on two cues, labeled cues A and B. The horizontal axis of each graph in this figure represents possible depth values, and the vertical axis represents the probability of a depth value. In each graph is shown the probability of depth given cue A, the probability of depth given cue B, and the filter's optimal depth estimate given both cues. The optimal depth estimate based on both cues is a weighted average of the means of the distributions based on single cues. If the distribution of depth given one cue is equal to the distribution given the other cue, then the weights used in the average are equal and the optimal estimate is halfway between the two means (top graph). If, however, the depth distribution given cue A has a smaller variance than the distribution given cue B, then the mean based on cue A is assigned a larger weight than the mean based on cue B and the optimal estimate is closer to the mean based on cue A (bottom graph).

where Equations 8 and 12 have dropped constants that do not depend on θ, and where Equations 9 and 11 have made use of the identity $e^a e^b = e^{a+b}$. As a matter of notation, let

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}x}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \tag{13}$$

and

$$\begin{aligned} \tau_1^2 &= \frac{1}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} && (14) \\ &= \frac{\tau_0^2\sigma^2}{\tau_0^2 + \sigma^2}. && (15) \end{aligned}$$

Using this notation, we re-write Equation 12 as

$$p(\theta|x) \propto \exp\{-\frac{1}{2}\theta^2\frac{1}{\tau_1^2} + \theta\frac{\mu_1}{\tau_1^2}\}. \tag{16}$$

Next we add into the exponent $-\frac{1}{2}\frac{\mu_1^2}{\tau_1^2}$ which is a constant that does not depend on θ, yielding

$$p(\theta|x) \propto \exp\{-\frac{1}{2\tau_1^2}(\theta - \mu_1)^2\}. \tag{17}$$

Because $p(\theta|x)$ is a density that must integrate to one, we get the result

$$p(\theta|x) = \frac{1}{\sqrt{2\pi\tau_1}}\exp\{-\frac{1}{2\tau_1^2}(\theta - \mu_1)^2\} \tag{18}$$

meaning that $p(\theta|x)$ is a Normal distribution with mean μ_1 and variance τ_1. Equation 13 is one way of expressing the posterior mean μ_1. An alternative way (which is closer to the typical notation for a Kalman filter; see below) is to express μ_1 as the prior mean μ_0 adjusted toward the observation x:

$$\begin{aligned} \mu_1 &= \mu_0 + \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}}(x - \mu_0) && (19) \\ &= \mu_0 + \frac{\tau_0^2}{\sigma^2 + \tau_0^2}(x - \mu_0). && (20) \end{aligned}$$

Using engineering terminology, $x - \mu_0$ is the "innovation" and $\frac{\tau_0^2}{\sigma^2+\tau_0^2}$ is the "Kalman gain".

What does all of this have to do with sensory integration? The mathematics above tells us how to combine two probability distributions. In this case, the two distributions are a likelihood distribution $p(x|\theta)$ and a prior distribution $p(\theta)$. Of course, the same mathematics could be used to combine any two (Normal) distributions, such as the distribution of an object's shape indicated

by vision and the distribution of the shape indicated by haptics, or the distribution of an event's spatial location indicated by vision and the distribution of the location indicated by audition.

For the sake of concreteness, let's consider the latter example. On a given trial of an experiment, an event occurs in the environment. A subject needs to localize this event based on the visual and auditory cues arising from the event, denoted v and a respectively. Let l_v denote a perceived spatial location based on vision, l_a denote a perceived location based on audition, and let l^* denote the true (but unknown) spatial location (which we will estimate based on both sensory cues). Assume l_v is a random variable with a Normal distribution with mean μ_v and variance σ_v^2, whereas l_a is a random variable with a Normal distribution with mean μ_a and variance σ_a^2. Our goal is to compute the posterior distribution $p(l^*|v,a)$. Using Bayes' rule,

$$p(l^*|v,a) \propto p(v,a|l^*)\, p(l^*). \tag{21}$$

Assuming that the visual and auditory cues are conditionally independent given the true location l^*, we arrive at the equation

$$p(l^*|v,a) \propto p(v|l^*)\, p(a|l^*)\, p(l^*) \tag{22}$$

where, using Bayes' rule

$$p(v|l^*) = \frac{p(l^*|v)p(v)}{p(l^*)} \tag{23}$$

$$p(a|l^*) = \frac{p(l^*|a)p(a)}{p(l^*)}. \tag{24}$$

Now let's assume that the prior distributions of the location $p(l^*)$, of the visual cue $p(v)$, and of the auditory cue $p(a)$ are uniform (i.e. constants), meaning that all possible locations, all possible visual cues, and all possible auditory cues are equally likely. In this case,

$$p(l^*|v,a) \propto p(v|l^*)\, p(a|l^*). \tag{25}$$

Using the mathematics above, we find that $p(l^*|v,a)$ is a Normal distribution. Its mean μ^* is

$$\mu^* = w_v\mu_v + w_a\mu_a \tag{26}$$

where the linear coefficients (or cue weights) w_v and w_a are

$$w_v = \frac{\frac{1}{\sigma_v^2}}{\frac{1}{\sigma_v^2}+\frac{1}{\sigma_a^2}} \tag{27}$$

$$= \frac{\sigma_a^2}{\sigma_v^2+\sigma_a^2} \tag{28}$$

and

$$w_a = \frac{\frac{1}{\sigma_a^2}}{\frac{1}{\sigma_v^2}+\frac{1}{\sigma_a^2}} \tag{29}$$

$$= \frac{\sigma_v^2}{\sigma_v^2+\sigma_a^2}. \tag{30}$$

Its variance σ^{2*} is given by

$$\sigma^{2*} = \frac{1}{\frac{1}{\sigma_v^2} + \frac{1}{\sigma_a^2}} \tag{31}$$

$$= \frac{\sigma_v^2 \sigma_a^2}{\sigma_v^2 + \sigma_a^2}. \tag{32}$$

The framework above can be extended in many (many, many, many) ways. Perhaps most importantly, we can extend the mathematics so that we are integrating three (or more) Normal distributions, meaning that we could apply this framework to the study of sensory integration when three (or more) sensory cues are available. In addition, we could consider extensions to likelihood functions which are not Normal, and prior distributions which are not uniform. These topics are beyond the scope of this note.

We now need to move closer to the typical theory of Kalman filters as it is discussed in the engineering literature. There are at least two important aspects of typical Kalman filters which are missing from the discussion above. The first important aspect is that Kalman filters operate "on-line". This implies that to compute a new best estimate of a quantity (and its uncertainty), we can update our previous estimate using a new measurement. This implies that we don't have to consider all the previous data all over again to compute the optimal estimates—we only need to consider the optimal estimate from the previous time step and the new measurement (we will use recursion). The second important aspect is that the quantity that we are trying to estimate might have some temporal dynamics. We need to estimate this quantity at each moment in time in a way that takes into account the dynamics.

Let's illustrate these ideas with a very simple example (due to Max Welling, Department of Computer Science, UC Irvine). Consider a ship at sea which has lost its bearings. The captain of the ship needs to estimate its current position using the stars. In the meantime, the ship moves due to the waves of the sea. Here is a simple model for the temporal dynamics of the ship's location:

$$y_{t+1} = y_t + c + w_t \tag{33}$$

where y_t is the ship's position at time t, c is a drift term (constant velocity) due to the waves, and w_t is a noise term (random variable with a Normal distribution with mean zero and variance σ_w^2). A noisy measurement of the ship's location is described by

$$x_t = y_t + v_t \tag{34}$$

where v_t is a noise term (random variable with a Normal distribution with mean zero and variance σ_v^2). In other words, the measurement of position is equal to the true position plus noise. Let's assume that we have some estimate $\hat{y}_t$ of the ship's location at time t and some uncertainty σ_t^2 in this estimate. How does our estimate change as the ship sails for one second? Of course, the ship will drift on average in the direction of its velocity by a distance c. Using our model of the temporal dynamics, we get

$$\hat{y}_{t+1}^{-} = \hat{y}_t + c. \tag{35}$$

The superscript '-' indicates that this is our prediction of the position at time $t+1$ before we've received the measurement x_{t+1}. Our uncertainty in this prediction must increase over time because we don't know the true drift of the ship (we don't know the noise term w_t):

$$\sigma_{t+1}^{2-} = \sigma_t^2 + \sigma_w^2. \tag{36}$$

Note that if we do not receive any measurements, the uncertainty will keep growing (the ship will keep drifting over time in ways that we cannot perfectly predict). If we receive a measurement, however, we can use this information. Our final estimate at time $t+1$ is a weighted average between the measurement of position and the guess of position based on our model of the temporal dynamics:

$$\hat{y}_{t+1} = \frac{\sigma_v^2}{\sigma_{t+1}^{2-} + \sigma_v^2}\hat{y}_{t+1}^- + \frac{\sigma_{t+1}^{2-}}{\sigma_{t+1}^{2-} + \sigma_v^2}x_{t+1}. \tag{37}$$

(Please understand this equation. It is perfectly analogous to the optimal linear cue combination rule described above.) Note that if we have infinite confidence in the measurement ($\sigma_v^2 = 0$), then the new estimate is equal to the measurement. On the other hand, if we have infinite confidence in the estimate based on the model of dynamics, then the measurement is ignored. For the new uncertainty, we obtain

$$\sigma_{t+1}^2 = \frac{\sigma_{t+1}^{2-}\sigma_v^2}{\sigma_{t+1}^{2-} + \sigma_v^2} \tag{38}$$

This is also easy to interpret as it says that if one of the uncertainties disappears, then the total uncertainty disappears. Note that the uncertainty always decreases or stays the same by adding a measurement. The estimates for the location and uncertainty, incorporating the measurement, can be written as follows:

$$\hat{y}_{t+1} = \hat{y}_{t+1}^- + K_{t+1}(x_{t+1} - \hat{y}_{t+1}^-) \tag{39}$$

$$\sigma_{t+1}^2 = (1 - K_{t+1})\sigma_{t+1}^{2-} \tag{40}$$

$$K_{t+1} = \frac{\sigma_{t+1}^{2-}}{\sigma_{t+1}^{2-} + \sigma_v^2} \tag{41}$$

From this we can see that the estimate is corrected by the measurement error (the "innovation": $x_{t+1} - \hat{y}_{t+1}^-$) multiplied by a gain factor (the Kalman gain: K_{t+1}). If the gain is zero, no attention is paid to the measurement. If its one, we simply use the measurement as our new estimate. Similarly for the uncertainties.