Bayesian multisensory perception

Timothy M. Hospedales **Sethu Vijayakumar**
School of Informatics, University of Edinburgh
EH1 2QL, Scotland, UK
t.hospedales@ed.ac.uk, sethu.vijayakumar@ed.ac.uk

Abstract

We investigate a solution to the problem of multi-sensor perception by formulating it in the framework of Bayesian model selection. Humans robustly integrate and segregate multi-sensory data as appropriate, but previous theoretical work has focused largely on purely integrative cases, leaving segregation unaccounted for and unexploited by machine perception systems. We illustrate a unifying, principled Bayesian solution which accounts for both integration and segregation by reasoning explicitly about data association in a probabilistic framework. Unsupervised learning of such a model with EM is illustrated for a real world audio-visual application.

1 Introduction

There has been much recent interest in optimal multi-sensor fusion both for understanding human multi-sensory perception[4, 1] and for machine perception applications[3, 6]. Most of these have considered the simple cases in which the observations are known to be correlated (generated from the same latent source), and the task is merely to make the best estimate of the latent source state by fusing the observations. However, in most real world perceptual situations any given pair of observations are unlikely to have originated from the same latent source. A more general problem in multi-sensor perception is therefore to infer the *association* between observations and any latent states of interest as well as any potential integration or segregation that may be necessary as a result. This data association problem has been of more long standing interest, for example, in the radar community[2]. Aside from enabling correct sensor fusion, data association can be of inherent interest for understanding higher level semantics encoded in the observations. For example, a key task in interpreting a meeting for a human or machine is not just to infer who was there and what was said, but to correctly associate visual and acoustic observations to understand who said what.

In this paper, we illustrate the commonality of multi-sensor perception problems in these domains and provide a unifying, principled Bayesian account of their solution, reasoning explicitly about the association of observations with latent states. Moreover, we illustrate that using the EM algorithm, such inference can be performed simultaneously with parameter estimation for unsupervised learning of perceptual models.

2 Theory

We can frame the inference of data association equivalently as a model selection or a structure inference problem. A graphical model for the process of generating observations in two *different modalities* $D = \{x_1, x_2\}$ from a *single* source with latent state l is illustrated in Fig. 1(a). The source state is drawn independently along with binary visibility/occlusion variables (M_1, M_2) specifying its visibility in each modality. The observations are then generated with x_i being dependent on l if $M_i = 1$ or on some background distribution if $M_i = 0$. Equivalently, all the structure options could be explicitly enumerated into four separate models, and the generation process then first selects

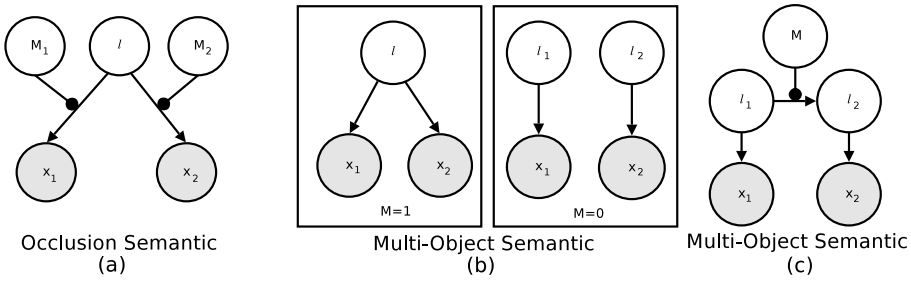

Figure 1: Graphical models to describe generation of multi-modal observations. (a) Occlusion semantic: Observations x_i are dependent on latent state l or a background distribution, depending on visibility structure variables, M_i. (b) Multi-object semantic: Observations x_i are determined by a single object of latent state l or two latent objects l_i depending on the model, M. (c) Multi-object semantic: Compact description.

one of the four models to describe how the observations will be generated before selecting l and generating the observations according to the dependencies encoded in that model. Inference in this model then consists of computing the posterior over the latent state and the generating model (either the two binary structure variables M_i or a single model index variable) given the observations. An observation in modality i is perceived as being associated with (having originated from) the latent source of interest with probability $p(M_i = 1|D)$, which will be large if the observation is likely under the foreground distribution and small if the observation is better explained by the background distribution.

To illustrate with a toy but concrete example, consider the problem of inferring a single dimensional latent state l representing a location on the basis of two point observations in separate modalities. l is governed by an informative[1] Gaussian prior centered at zero, i.e., $p(l) = \mathcal{N}(l|0, p_l)$ and the binomial visibility variables have prior probability $p(M_i) = \pi_i$. If the state is observed by sensor i ($M_i = 1$) then the observation in that modality is generated with precision p_i, such that $x_i \sim \mathcal{N}(x_i|l, p_i)$. Alternately, if the state is not observed by the sensor, its observation is generated by the background distribution $\mathcal{N}(x_i|0, p_b)$, which tends toward un-informativeness with $p_b \to 0$. The joint probability of the model can be written as

$$p(D, l, M) = \mathcal{N}(x_1|l, p_1)^{M_1}\mathcal{N}(x_1|0, p_b)^{(1-M_1)}\mathcal{N}(x_2|l, p_2)^{M_2}\mathcal{N}(x_2|0, p_b)^{(1-M_2)}\mathcal{N}(l|0, p_l)p(M_1, M_2)$$

If we are purely interested in computing the posterior over latent state, we integrate over models or structure variables. For the higher level task of inferring the cause or association of observations, we integrate over the state to compute the posterior model probability, benefiting from the automatic complexity control induced by Bayesian Occam's razor[5].

$$\begin{aligned}
p(M_1 = 0, M_2 = 0|x_1, x_2) &\propto \mathcal{N}(x_1|0, p_b)\mathcal{N}(x_2|0, p_b)(1 - \pi_1)(1 - \pi_2) \\
p(M_1 = 1, M_2 = 0|x_1, x_2) &\propto \frac{1}{Z_1}exp - \frac{1}{2}\left(x_1^2 p_1 p_l/(p_1 + p_l)\right)\mathcal{N}(x_2|0, p_b)\pi_1(1 - \pi_2) \qquad (1) \\
p(M_1 = 1, M_2 = 1|x_1, x_2) &\propto \frac{1}{Z_2}exp - \frac{1}{2}\left(\frac{x_1^2 p_1(p_2 + p_l) - 2x_1 x_2 p_1 p_2 + x_2^2 p_2(p_1 + p_l)}{p_1 + p_2 + p_l}\right)\pi_1\pi_2
\end{aligned}$$

Intuitively, the structure posterior (Eq. 1) is dependent on the relative data likelihood under the background and marginal foreground distributions. The posterior of the fully segregative model depends on the background distributions and hence tends toward being independent of the data except via the normalization constant. In contrast, the posterior of the fully integrative, pure fusion model depends on the three way agreement between the observations and the prior.

Fig. 2 illustrates a schematic of some informative types of behavior produced by this model. If the data and the prior are all strongly correlated (Fig. 2(a)) such that both observations are inferred with

[1]Think of this as a filtering task where we have the estimate of l from the previous frame

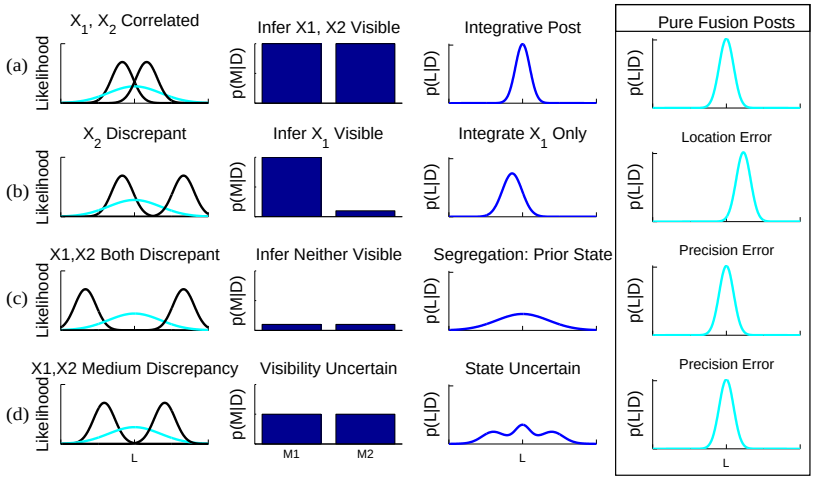

Figure 2: Inference in occlusion semantic toy model. Likelihoods of the observations in each of two modalities in black, prior in grey. Observations (a) x_1, x_2 strongly correlated, (b) x_2 strongly discrepant, (c) x_1, x_2 both strongly discrepant, (d) x_1, x_2 both moderately discrepant.

near certainty to be associated with the latent source of interest, the fused posterior over the location is approximately Gaussian with $p(l|x_1, x_2) = \mathcal{N}(l|\hat{l}, p_{l|x})$ where $p_{l|x} = p_1 + p_2 + p_l$, $\hat{l} = \frac{p_1 x_1 + p_2 x_2}{p_{l|x}}$. If x_2 is strongly discrepant with x_1 and the prior (Fig. 2(b)), it would be inferred with near certainty that sensor 2 was occluded and its observation generated by the background distribution. In this case, the posterior over the state is again near Gaussian but fusing only x_1 and the prior; $p_{l|x} = p_1 + p_l$, $\hat{l} = \frac{p_1 x_1}{p_{l|x}}$. If both x_1 and x_2 are strongly discrepant with each other and the prior (Fig. 2(c)), both sensors are likely to have been occluded, in which case the posterior over the latent state reverts to the prior $p_{l|x} = p_l$, $\hat{l} = 0$. Finally, if the correlation between the observations and the prior is only moderate (Fig. 2(d)) such that the posterior marginal over the structural visibility variables are not near certain, then the posterior marginal over the latent state is a (potentially quad-modal) mixture of 4 Gaussians corresponding to the four possible models. For real world data, occlusion, or other cause for meaningless observation is almost always possible, in which case assuming a pure fusion model (Fig. 2(box)) can result in dramatically inappropriate inference.

There is one more distinguishable way in which two point observations can be generated, i.e., each could be generated by a separate source instead of a single source. The choice of the multi-source versus the fused generating model (Fig. 1(b)) can also be expressed compactly as structure inference as before by also using two latent state variables as in the single source case, but requiring equality between them if $M = 1$ and independence if $M = 0$ (Fig. 1(c)).

It is possible to enumerate all five possible model structures and perform the Bayesian model selection given the data. However, usually the semantics of a given perceptual problem correspond to a prior over models which either allows the four discussed earlier ("occlusion semantic") or a choice between one or two sources ("multi-object semantic"). The occlusion semantic arises for example, in audio-visual processing where a source may independently be either visible or audible. The multi-object semantic arises, for example in some psychophysics experiments[7] where both sensors have definitely observed an interesting event, and the task is to decide what they observed, which is conditionally dependent on whether they observed the same source or not.

We will now illustrate the latter case with a toy but concrete example of generating observations in two different modalities x_1, x_2 which may both be due to a single latent source $(M = 1)$, or two separate sources $(M = 0)$. Using vector notation, the likelihood of the observation $\mathbf{x} = [x_1, x_2]^T$ given the latent state $\mathbf{l} = [l_1, l_2]^T$ is $\mathcal{N}(\mathbf{x}|\mathbf{l}, \mathbf{P}_x)$ where $\mathbf{P}_x = diag([p_1, p_2])$. Let us assume the prior distributions over the latent locations are Gaussian but tend to un-informativeness. In the multi-object model the prior over l_is $p(\mathbf{l}|M = 0) \sim \mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_0)$ is uncorrelated, so $\mathbf{P}_0 = p_0\mathbf{I}$ and $p_0 \to 0$. In the single object model, the prior over $l_i s$ $p(\mathbf{l}|M = 1) \sim \mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_1)$ requires the l_is to

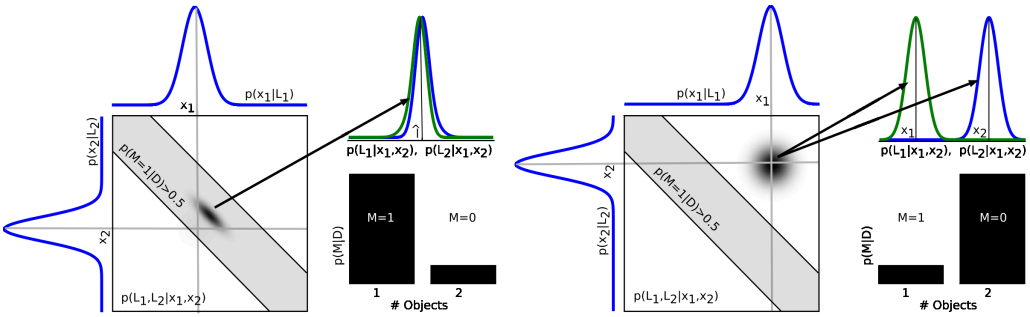

Figure 3: Inference in multi-object semantic toy model. (a) For correlated inputs, $x_1 \cong x_2$, the presence of one object is inferred and its location posterior is the probabilistic fusion of the observations. (b) For very discrepant inputs, $x_1 \neq x_2$, the presence of two objects is inferred and the location posterior for each is at the associated observation.

be equal so $\mathbf{P}_1$ is chosen to be strongly correlated. The joint probability of the whole model and the posterior over the structure are given in Eq. 2.

$$\begin{aligned} p(\mathbf{x}, \mathbf{l}, M) &= \mathcal{N}(\mathbf{x}|\mathbf{l}, \mathbf{P}_x)\mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_0)^{(1-M)}\mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_1)^{M}p(M) \\ p(M|\mathbf{x}) &\propto \int_{\mathbf{l}} \mathcal{N}(\mathbf{x}|\mathbf{l}, \mathbf{P}_x)\mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_0)^{(1-M)}\mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_1)^{M}p(M)d\mathbf{l} \\ p(M=0|\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\mathbf{0}, (\mathbf{P}_x^{-1} + \mathbf{P}_0^{-1})^{-1}) \\ p(M=1|\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\mathbf{0}, (\mathbf{P}_x^{-1} + \mathbf{P}_1^{-1})^{-1}) \end{aligned} \tag{2}$$

A schematic of interesting behavior observed is illustrated in Fig. 3. If x_1 and x_2 are only slightly discrepant (Fig. 3(a)), then the single object model is inferred with high probability. The posterior over $\mathbf{l}$ is also strongly correlated and Gaussian about the point of the fused interpretation; $p(\mathbf{l}|\mathbf{x}) \sim \mathcal{N}(\mathbf{l}|\hat{\mathbf{l}}, \mathbf{P}_{l|x})$ where $\hat{\mathbf{l}} = \mathbf{P}_{l|x}^{-1}\mathbf{P}_x\mathbf{x}$, $\mathbf{P}_{l|x} = \mathbf{P}_x + \mathbf{P}_1$. The location marginals for each l_i are therefore the same and aligned at $\hat{\mathbf{l}}$. If x_1 and x_2 are highly discrepant (Fig. 3(b)), then the two object model is inferred with high probability. In this case the posterior $p(\mathbf{l}|\mathbf{x})$ is spherical and aligned with the observations themselves rather than a single fused estimate; i.e. $\hat{\mathbf{l}} = \mathbf{P}_{l|x}^{-1}\mathbf{P}_x\mathbf{x} \simeq \mathbf{x}$, $\mathbf{P}_{l|x} = \mathbf{P}_x + \mathbf{P}_0$.

The inferences discussed so far have been exact. Obviously there are various potential approximations such as computing the *location posterior* given the MAP model, which may be acceptable in some cases, but crucially misrepresents the state posterior for regions of input space with intermediate discrepancy (c.f. Fig. 2(d)). Alternately, the *model probability* could be approximated using a MAP or ML estimate of the latent state variable. The agreement between the Bayesian and MAP solution depends on how much the latent state posterior is like a delta function, which depends on both the agreement between observations and the precision of their likelihoods. However, using the ML estimate of the state will not work at all as the most complex model will always be selected.

Previous probabilistic accounts of human multi-sensory combination (e.g. [4, 1]) are special cases of our theory, having explicitly or implicitly assumed a pure fusion structure. [8] describes a heuristic democratic *adaptive* cue integration perceptual model, but again assumes a pure fusion structure. Hence these do not, for example, exhibit the robust discounting (sensory fission or segregation) of strongly discrepant cues observed in humans[4]. As we have seen, such fission is necessary for perception in the real world as outliers can break mandatory fusion schemes. In a radar context, these issues have been addressed somewhat with heuristic schemes such as validation gates [2]. In contrast we provide a principled probabilistic, adaptive theory of sensor combination which can account for fusion, fission and the spectrum in-between. The combination strategy is handled by a Bayesian model selection without recourse to heuristics, and the remaining parameters can be learned directly from the data with EM.

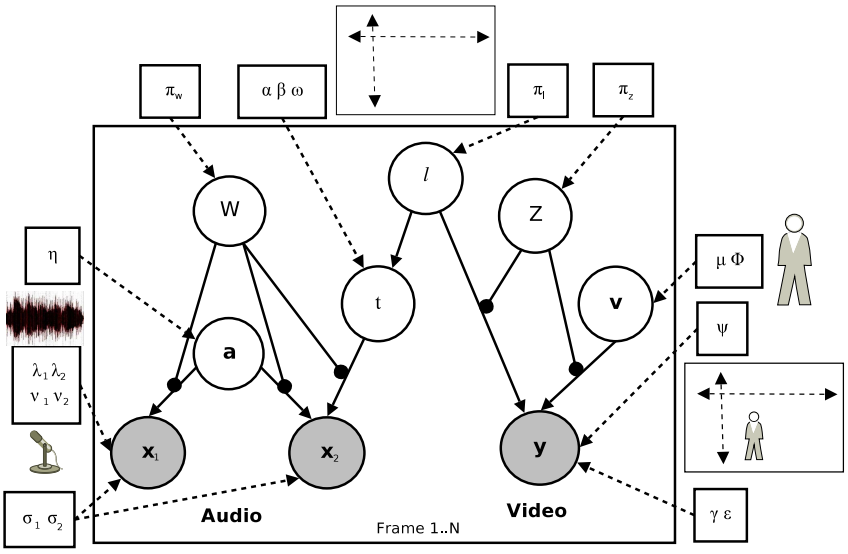

Figure 4: Graphical model for audio-visual inference & data association

The illustrative scenarios discussed here generalize in the obvious way to more observations. A more challenging question is that of realistic multi-dimensional observations which depend in complex ways on the latent state, a topic we will address in the real world application discussed next.

3 Bayesian Multi-sensory Perception for Audio-Visual Scene Understanding

To illustrate the application of these ideas to a real, large scale machine perception problem, we consider a task inspired by [3]; that of unsupervised learning and inference with audio-visual (AV) input. [3] demonstrated inference of an AV source location and learning of its template based on correlations between the input from a camera and two microphones - useful for example, in teleconferencing applications[6]. The AV localization part of this task (and optimal solution) is similar to the task in psychophysics experiments such as [1]. We now tackle the bigger scene understanding problem of inferring how the AV data should be associated (pure fusion was previously assumed), i.e, whether the source should be associated with both modalities, or only one, or if there is no source present at all. This is a problem of the "occlusion semantic" type as discussed previously.

3.1 Introduction

A graphical model to describe the generation of audio-visual data $D = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}\}$ is illustrated in Fig. 4. For a given frame, the discrete translation l representing the source state is selected from its prior distribution π_l and its observability in each modality (W, Z) are selected from their binomial priors. For simplicity, we only consider source translation along the azimuth. Consider first the all visible case $(\mathrm{W} = \mathrm{Z} = 1)$. The video appearance $\mathbf{v}$ is sampled from a diagonal Gaussian distribution $\mathcal{N}(\mathbf{v}|\mu, \phi)$ with parameters defining its soft template. The observed video pixels are generated by sampling from another Gaussian $\mathcal{N}(\mathbf{y}|\mathbf{T}_l\mathbf{v}, \Psi\mathbf{I})$ the mean of which is the sampled appearance translated by l using the transformation matrix $\mathbf{T}_l$. The latent audio signal $\mathbf{a}$ is sampled from a zero mean, uniform covariance Gaussian i.e., $\mathcal{N}(\mathbf{a}|\mathbf{0}, \eta\mathbf{I})$. The time delay between the signals at each microphone is drawn as a linear function of the translation of the source $\mathcal{N}(t|\alpha l+\beta, \omega)$. On the basis of the latent signal and the delay, the observation at each microphone is generated by sampling from a uniform diagonal Gaussian with the mean $\mathbf{a}$, shifted τ samples relative to each other; $\mathcal{N}(\mathbf{x}_1|\mathbf{a}, v_1\mathbf{I})$, $\mathcal{N}(\mathbf{x}_2|\mathbf{T}_t\mathbf{a}, v_2\mathbf{I})$. If the video modality is to be occluded $(\mathrm{Z} = 0)$, the observed video pixels are drawn from a Gaussian background distribution $\mathcal{N}(\mathbf{y}|\gamma\mathbf{1}, \epsilon\mathbf{I})$ independently of latent state and audio data. If the audio modality is to be silent $(\mathrm{W} = 0)$, the samples at each speaker are drawn from Gaussian background distributions $\mathcal{N}(\mathbf{x}_i|\mathbf{0}, \sigma_i\mathbf{I})$ independently of each other, the latent state and the video data. The joint probability of the model therefore factorizes as

$$
\begin{aligned}
p(\mathbf{x_1}, \mathbf{x_2}, \mathbf{y}, \mathbf{a}, t, l, \mathbf{v}, W, Z) &= p(\mathbf{x}_1|W, \mathbf{a})p(\mathbf{x}_2|W, \mathbf{a}, t)p(t|l)p(\mathbf{y}|Z, \mathbf{v}, l)p(\mathbf{v})p(\mathbf{a})p(l)p(\mathrm{W})p(\mathrm{Z}) \\
&= \mathcal{N}(\mathbf{x}_1|\mathbf{a}, v_1)^W \mathcal{N}(\mathbf{x}_1|\mathbf{0}, \sigma_1)^{(1-W)} \mathcal{N}(\mathbf{x}_2|T_t\mathbf{a}, v_2)^W \mathcal{N}(\mathbf{x}_2|\mathbf{0}, \sigma_2)^{(1-W)} \\
&\cdot \quad \mathcal{N}(\mathbf{a}|0, \eta)p(\mathrm{W})\mathcal{N}(\mathbf{y}|T_l\mathbf{v}, \Psi\mathbf{I})^Z \mathcal{N}(\mathbf{y}|\gamma\mathbf{1}, \epsilon\mathbf{I})^{(1-Z)} \mathcal{N}(\mathbf{v}|\mu, \phi)p(\mathrm{Z})p(l)
\end{aligned}
$$

3.2 Inference

The posterior marginal of interest for the scene interpretation task is that of the discrete location and visibility structure variables $p(l, \mathrm{W}, \mathrm{Z}|D)$. Because of the linear-Gaussian structure of the model, the latent appearance variables $\mathbf{a}$ and $\mathbf{v}$ can be analytically integrated, leaving only the inter-microphone delay t and source location l to be summed out numerically. Conditioned on the fused model, and other discrete variables $(\mathrm{Z} = 1, \mathrm{W} = 1, t, l)$ the posteriors over the latent signals are Gaussian, $\mathcal{N}(\mathbf{a}|\mu_{\mathbf{a}|\mathbf{x},t}, \nu_{\mathbf{a}})$ and $\mathcal{N}(\mathbf{v}|\mu_{\mathbf{v}|\mathbf{y},l}, \nu_{\mathbf{v}})$, with precision and mean given by $\mu_{\mathbf{a}|\mathbf{x},t} = v_{\mathbf{a}}^{-1}(\lambda_1\nu_1\mathbf{x}_1 + \lambda_2\nu_2\mathbf{T}_t^T\mathbf{x}_2)$, $\nu_{\mathbf{a}} = \eta + \lambda_1^2\nu_1 + \lambda_2^2\nu_2$, $\mu_{\mathbf{v}|\mathbf{y},l} = \nu_{\mathbf{v}}^{-1}(\phi\mu + \mathbf{T}_l^T\Psi\mathbf{y})$, $\nu_{\mathbf{v}} = \phi + \Psi$. The marginal video likelihood is also Gaussian with $\mu_{\mathbf{y}|l} = \mathbf{T}_l\mu$, $\nu_{\mathbf{y}|l} = (\Psi^{-1} + \mathbf{T}_l\phi_s^{-1}\mathbf{T}_l^T)^{-1}$. Expressions for the posterior of the fully fused model and the source location (Eq. 3) and the posterior of the fully fissioned model (Eq. 4) can be derived in terms of these statistics.

$$
\begin{aligned}
p(l, \mathrm{W} = 1, \mathrm{Z} = 1|D) &\propto \left(\int_{\mathbf{v}} p(\mathbf{y}, \mathbf{v}|l, Z = 1)\right)\left(\sum_t \int_{\mathbf{a}} p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{a}, t|l, W = 1)\right) p(\mathrm{W}{=}1, \mathrm{Z} = 1, l) \\
&\propto \left(\mathcal{N}(\mathbf{y}|\mu_{\mathbf{y}|l}, \nu_{\mathbf{y}|l})\right)\left(\sum_t p(t|l, D) exp(\mu_{\mathbf{a}|t,\mathbf{x}}^{\mathbf{T}}\nu_{\mathbf{a}}\mu_{\mathbf{a}|t,\mathbf{x}})\right)\pi_l\pi_w\pi_z \qquad (3) \\
p(l, \mathrm{W} = 0, \mathrm{Z} = 0|D) &\propto p(\mathbf{x}|\mathrm{W} = 0)p(\mathbf{y}|\mathrm{Z} = 0)p(\mathrm{W} = 0, \mathrm{Z} = 0)p(l) \\
&= \mathcal{N}(\mathbf{x}_1|\mathbf{0}, \sigma_1\mathbf{I})\mathcal{N}(\mathbf{x}_2|\mathbf{0}, \sigma_2\mathbf{I})\mathcal{N}(\mathbf{y}|\gamma\mathbf{1}, \epsilon\mathbf{I})\pi_l(1 - \pi_w)(1 - \pi_z) \qquad (4)
\end{aligned}
$$

For a single observed modality, the posterior is a mixture of these terms in a similar manner to Eqs. 1.

3.3 Learning

All the parameters in this model $\theta = \{\lambda_1, \lambda_2, \nu_1, \nu_2, \eta, \alpha, \beta, \omega, \pi_l, \mu, \phi, \Psi, \pi_w, \pi_z, \gamma, \epsilon, \sigma_1, \sigma_2\}$ are jointly optimized by a standard EM procedure of alternately inferring the posterior distribution $q(H|D)$ over hidden variables $H = \{\mathbf{a}, \mathbf{v}, l, t, \mathrm{W}, \mathrm{Z}\}$ given the observed data $D = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}\}$ and optimizing the expected complete log likelihood or free energy $\frac{\partial}{\partial\theta}\int_H q(H|D)lg\frac{p(H,D)}{q(H|D)}$. As this is a complex model of many parameters, in the interest of space, we present just two informative updates[2]. Eq. 5 gives the update for the mean μ of the source visual appearance distribution in terms of the posterior mean $\mu^j_{\mathbf{v}|\mathbf{y},l}$ of the video appearance given the data D^j for each frame j and translation l, as inferred during the E step. Intuitively, the result is a weighted sum of the appearance inferences over all frames and transformations, where the weighting is the posterior probability of transformation and visibility in each frame. N_f specifies the number of samples per audio frame.

$$
\mu \leftarrow \frac{\sum_{j,l} p(l, \mathrm{Z} = 1|D^j)\mu^j_{\mathbf{v}|\mathbf{y},l}}{\sum_j p(\mathrm{Z} = 1|D^j)} \qquad (5) \qquad\qquad \sigma_i^{-1} \leftarrow \frac{\sum_j q(\mathrm{W} = 0|D^j)(\mathbf{x}_i^j)^T\mathbf{x}_i^j}{N_f\sum_j q(\mathrm{W} = 0|D^j)} \qquad (6)
$$

The scalar precision parameter of background noise is given by Eq. 6. Again, it is intuitive that the estimate of the background variance should be a weighted sum of square signals at each frame where the weighting is the posterior probability of the source being silent in that frame. Note that because of the probabilistic formulation, the updates still work and make intuitive sense even if at some point during learning, all frames in the sequence are inferred with near certainty to be visible or not, though care must be taken in the numerical implementation.

[2]Full E and M step derivations are attached as supplementary material

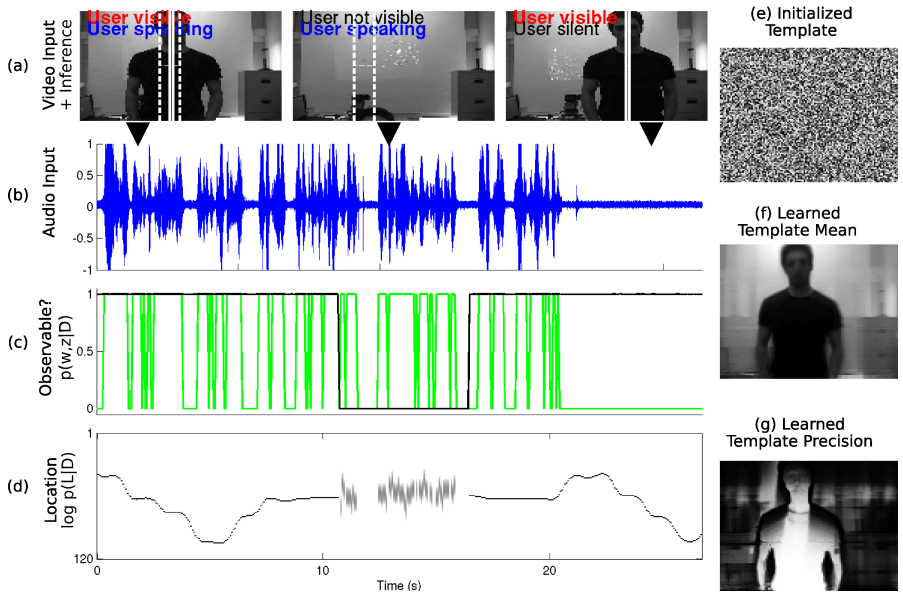

Figure 5: Audio-visual data association & inference results. Video samples (a) and audio data (b) from a sequence where the user is first visibly walking and speaking, then hidden but still speaking, and finally visible and walking but silent. (c) Posterior probability of visibility (dark) and audibility (light) during the sequence. (d) Posterior probability of source location during the sequence. (e) initial and (f,g) final video appearance parameters after learning.

3.4 Demonstration

Results for an AV sequence after 20 cycles of EM are illustrated in Fig. 5. In this sequence, the user is initially walking and talking, then hides from the camera while continuing to speak and then continues to walk while remaining silent. Fig. 5(a) illustrates three representative video frames from each of the described segments of the sequence and Fig. 5(b) illustrates the averaged audio input waveform. The posterior probability of the observability structure variables (W, Z) are shown in 5(c). The more reliable video modality is correctly inferred to be observable at the appropriate times. On the basis of the much noisier audio observations, the source is correctly inferred to be silent during the appropriate period. The fluctuations in the inference about audibility of the source in the rest of the sequence is appropriate behavior for the model as they are almost always at the pauses intrinsic to speech data, during which the observations at the microphones are uncorrelated background noise. To explicitly identify contiguous periods of speech including pauses, the current approach of treating each frame independently would need to be extended to include a suitable Markov model over the frames. The posterior log probability of the source location at each frame is visible in Fig. 5(d) - it is strongly peaked at the correct azimuth while the user is visible, and peaked less strongly about the correct location while the user is only audible.

This model retains all the properties of the inspiring formulation[3] which allow most of the expensive E and M step computations to be expressed in terms of FFTs. This enables very quick processing of 22 frame-cycles per second in matlab for data consisting of 120x100 pixel images and 1000 sample audio frames[3], allowing real time learning and inference.

To cope with intermittent cues, previous multi-modal machine perception systems in this context have relied on observations of discrepant modalities providing uninformative likelihoods, [6, 3], which may not always be the case. For example, the pure fusion model used in[3] fails in the illustrated video sequence. When the subject is audible but not visible, the visual inference of the next best location (the filing cabinet) dominates that of the audio, instead of being discounted due to the poor match of learned visual template and Bayesian Occam's razor. Importantly, by explicitly inferring association, the model "knows" when observations arise from the source of interest or not. This

[3]Video clip attached as supplement. Further examples and matlab code at http://<anonymized>

is important for models attempting to infer higher level structure in the data. For example, a speech transcription model should not associate a nearby background conversation of poorly matching template and uncorrelated spatial location with the visible user when he is silent. Finally, in contrast to data association methods studied in other fields [2], our model is more principled in formulation and can work directly with the data rather than candidate observations; hence, using signature or template information in a unified way along with correlation.

4 Discussion

In this paper, we introduced a principled formulation of multi-sensor perception in the framework of Bayesian inference and model selection in probabilistic graphical models. Bayesian models of multi-sensor *fusion* have previously been applied in machine perception applications and understanding human perception. However, for sensor combination with real world data, extra inference in the form of data association is necessary as most pairs of signals should not actually be integrated. In many cases, deciding how the observations should be associated is in itself important for understanding the higher level structure of the observed data.

Investigations of human multi-sensory perception have reported robust discounting of discrepant cues [7, 4] but principled theory to explain this has been lacking. We envisage that our theory can be used to understand a much greater range of integrative and segregative perceptual phenomena in a unified way - including, for the first time, higher level perceptual association. Performing psychophysical experiments to investigate whether human perceptual association is consistent with the optimal theory described here is a major research theme which we are currently investigating.

In the case of machine perception, the type of model described generalizes existing integrative models and provides a principled solution to questions of sensor combination including use of signature, fusion, fission and points between. As our AV application illustrates, computing the exact posterior over latent source and data association for real problems is potentially real-time even before employing approximations. The major complicating extension, which we have not considered here on a large scale, is that of multiple objects, potentially observed simultaneously with each sensor. In this case, the computation required for exhaustive reasoning grows exponentially in the maximum number of objects; so for more than a few objects the simple strategy employed here is not viable. For these problems, we are investigating using approximate greedy inference to identify the multi-modally observed objects one at a time in order of best correlation along the lines of [9].

References

[1] D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol*, 14(3):257–262, Feb 2004.

[2] Y. Bar-Shalom, T. Kirubarajan, and X. Lin. Probabilistic data association techniques for target tracking with applications to sonar, radar and eo sensors. *IEEE Aerospace and Electronic Systems Magazine*, 20(8):37–56, Aug. 2005.

[3] M. J. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):828–836, July 2003.

[4] M. O. Ernst and M. S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415:429–433, 2002.

[5] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[6] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proceedings of the IEEE*, 92(3):495–513, 2004.

[7] L. Shams, Y. Kamitani, and S. Shimojo. Illusions: What you see is what you hear. *Nature*, 408:788, December 2000.

[8] J. Triesch and C. von der Malsburg. Democratic integration: self-organized integration of adaptive cues. *Neural Comput*, 13(9):2049–2074, Sep 2001.

[9] C. K. I. Williams and M. K. Titsias. Greedy learning of multiple objects in images using robust statistics and factorial learning. *Neural Comput*, 16(5):1039–1062, May 2004.

Appendix: Equations

Inference

The joint distribution and posterior distributions can be factored as follows

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}, \mathbf{a}, t, l, \mathbf{v}, W, Z) &= p(\mathbf{x}_1\mathbf{x}_2|t, W, \mathbf{a}, l)p(\mathbf{y}|\mathbf{v}, Z, l)p(W, Z, \mathbf{v}, \mathbf{a}, l, t) \\ &= p(\mathbf{x}_1|W, \mathbf{a})p(\mathbf{x}_2|W, \mathbf{a}, t)p(t|l)p(\mathbf{y}|\mathbf{z}, v, l)p(\mathbf{a})p(\mathbf{v})p(l)p(W)p(Z) \\ &= \mathcal{N}(\mathbf{x}_1|\lambda_1\mathbf{a}, \nu_1)^w\mathcal{N}(\mathbf{x}_1|0, \sigma_1)^{(1-w)}\mathcal{N}(\mathbf{x}_2|\lambda_2\mathbf{T}_t\mathbf{a}, \nu_2)^w\mathcal{N}(\mathbf{x}_2|0, \sigma_2)^{(1-w)} \\ &\quad \cdot\mathcal{N}(t|\alpha l + \beta, \omega)\mathcal{N}(\mathbf{y}|\mathbf{T}_l\mathbf{v}, \Psi)^z\mathcal{N}(\mathbf{y}|\gamma, \epsilon)^{(1-z)} \\ &\quad \cdot\mathcal{N}(\mathbf{a}|0, \eta)\mathcal{N}(\mathbf{v}|\mu, \phi,)\pi_l\pi_w\pi_z \end{aligned}$$

$$p(\mathbf{a}, \mathbf{v}, t, l, W, Z|D) = p(\mathbf{v}|l, D)p(\mathbf{a}|t, D)p(t|l, D)p(t, l, W, Z|D)$$

Pre-transformation signals

The posterior over pre-transformation signals are products of Gaussians and therefore Gaussian. Define also for brevity, $z \equiv (Z = 1), \overline{z} \equiv (Z = 0)$ and $w \equiv (W = 1), \overline{w} \equiv (W = 0)$.

$$\begin{aligned} p(\mathbf{v}|l, \mathbf{y}, z) &\propto \mathcal{N}(\mathbf{y}|T_l\mathbf{v}, \Psi)\mathcal{N}(\mathbf{v}|\mu, \phi) \\ &= \mathcal{N}(\mathbf{v}|\mu_{\mathbf{v}|l}, \nu_{\mathbf{v}}) \\ \nu_{\mathbf{v}|l,z} &= \phi + \Psi \\ \mu_{\mathbf{v}|l,z} &= (\nu_{\mathbf{v}|l,z})^{-1}(\phi\mu + \mathbf{T}_l^T\Psi\mathbf{y}) \end{aligned}$$

$$\begin{aligned} p(\mathbf{a}|t, r, \mathbf{x}_1, \mathbf{x}_2, w) &\propto \mathcal{N}(\mathbf{x}_1|\lambda_1\mathbf{a}, \nu_1)\mathcal{N}(\mathbf{x}_2|\lambda_2\mathbf{T}_t\mathbf{a}, \nu_2)\mathcal{N}(\mathbf{a}|0, \eta) \\ &= \mathcal{N}(\mu_{\mathbf{a}|t,w}, v_{\mathbf{a}|w}) \\ \nu_{\mathbf{a}|w} &= \lambda_1^2\nu_1 + \lambda_2^2\nu_2 + \eta \\ \mu_{\mathbf{a}|t,w} &= (\nu_{\mathbf{a}|w})^{-1}(\lambda_1\nu_1\mathbf{x}_1 + \lambda_2\nu_2\mathbf{T}_t^T\mathbf{x}_2) \end{aligned}$$

Misc

Marginal likelihood of the video is also gaussian

$$\begin{aligned} p(\mathbf{y}|l, z) &= \int_{\mathbf{v}} \mathcal{N}(\mathbf{y}|\mathbf{T}_l\mathbf{v}, \Psi)\mathcal{N}(\mathbf{v}|\mu, \phi) \\ &= \mathcal{N}(\mathbf{y}|\mu_{\mathbf{y}|l,z}, \nu_{\mathbf{y}|l,z}) \\ \mu_{\mathbf{y}|l} &= \mathbf{T}_l\mu \\ \nu_{\mathbf{y}|l} &= (\Psi^{-1} + \mathbf{T}_l\phi_s^{-1}\mathbf{T}_l^T)^{-1} \end{aligned}$$

Inter-aural time delay

$$\begin{aligned} p(\tau|l, \mathbf{x}_1, \mathbf{x}_2, w) &\propto p(\tau|l, w)p(\mathbf{x}_1\mathbf{x}_2|\tau, w) \\ &= p(\tau|l, w)\int_{\mathbf{a}} p(\mathbf{x}_1\mathbf{x}_2\mathbf{a}|\tau, w) \\ &\propto p(\tau|l, w)exp(\lambda_1\lambda_2\nu_1\nu_2c_\tau) \\ c_\tau &= \sum_i \mathbf{x}_{1,i-\tau}\mathbf{x}_{2,i}/v_{\mathbf{a}|w} \end{aligned}$$

Location & structure

Consider for example the totally visible and totally non-visible cases

$$\begin{aligned}
p(l,w,z|D) &\propto p(\mathbf{y}|l,z)p(\mathbf{x}_1,\mathbf{x}_2|l,w)p(w,z,l) \\
&= \left(\int_{\mathbf{v}} p(\mathbf{y},\mathbf{v}|l,z)\right)\left(\int_{\mathbf{a}}\sum_t p(\mathbf{x}_1,\mathbf{x}_2,\mathbf{a},t|l,w)\right)p(w)p(z)p(l) \\
&\propto \left(\mathcal{N}(\mathbf{y}|\mu_{\mathbf{y}|l,w},v_{\mathbf{y}|l,w})\right)\left(\sum_t p(t|l,D)exp\frac{1}{2}(\mu^T_{\mathbf{a}|\tau,w}v_{\mathbf{a}|w}\mu_{\mathbf{a}|\tau,w})\right)\pi_l\pi_w\pi_z \\
p(l,\overline{w},\overline{z}|D) &\propto p(\mathbf{y}|l,\overline{z})p(\mathbf{x}_1,\mathbf{x}_2|l,\overline{w})p(\overline{w},\overline{z},l) \\
&= p(\mathbf{x}|\overline{w})p(\mathbf{y}|\overline{z})p(\overline{w})p(\overline{z})p(l) \\
&= \mathcal{N}(\mathbf{x}_1|\mathbf{0},\sigma_1\mathbf{I})\mathcal{N}(\mathbf{x}_2|\mathbf{0},\sigma_2\mathbf{I})\mathcal{N}(\mathbf{y}|\gamma\mathbf{1},\epsilon\mathbf{I})\pi_l\pi_{\overline{w}}\pi_{\overline{z}}
\end{aligned}$$

Paramater Updates

For the video parameters, where N_{xy} is the number of pixels in each image and $\mathbf{y}^j(i)$ is pixel i in image frame j.

$$\begin{aligned}
\mu &\leftarrow \frac{\sum_{l,j} p(l,z|\mathbf{D}^j)\mu^j_{\mathbf{v}|l,z}}{\sum_j p(z|\mathbf{D}^j)} \\
\phi^{-1} &\leftarrow \frac{\sum_{l,j} p(l,z|\mathbf{D}^j)diag((\mu^j_{\mathbf{v}|l,z}-\mu)(\mu^j_{\mathbf{v}|l,z}-\mu)^T+\nu^{-1}_{\mathbf{v}|z})}{\sum_j p(z|\mathbf{D}^j)} \\
\gamma &\leftarrow \frac{\sum_j p(\overline{z}|\mathbf{D}^j)\sum_i \mathbf{y}^j(i)}{N_{xy}\sum_j p(\overline{z}|\mathbf{D}^j)} \\
\epsilon^{-1} &\leftarrow \frac{\sum_j p(\overline{z}|\mathbf{D}^j)(\mathbf{y}^j-\gamma)^T(\mathbf{y}^j-\gamma)}{N_{xy}\sum_j p(\overline{z}|\mathbf{D}^j)}
\end{aligned}$$

For the audio parameters, where N_f is the number of samples in each frame

$$\begin{aligned}
\eta^{-1} &\leftarrow \frac{\sum_{t,j} p(\tau,w|\mathbf{D}^j)(\mu^2_{\mathbf{a}|t,w}+Tr(\nu^{-1}_{\mathbf{a}|w}))}{N_f\sum_j p(w|\mathbf{D}^j)} \\
\lambda_1 &\leftarrow \frac{\sum_{t,j} p(t,w|\mathbf{D}^j)\mathbf{x}_1^T\mu_{\mathbf{a}|\tau,w}}{\sum_{t,j} p(t,w|\mathbf{D}^j)(\mu^2_{\mathbf{a}|\tau,w}+Tr(\nu^{-1}_{\mathbf{a}|w}))} \\
\lambda_2 &\leftarrow \frac{\sum_{t,j} p(t,w|\mathbf{D}^j)\mathbf{x}_2^T\mathbf{T}_\tau\mu_{\mathbf{a}|\tau,w}}{\sum_{t,j} p(t,w|\mathbf{D}^j)(\mu^2_{\mathbf{a}|\tau,w}+Tr(\nu^{-1}_{\mathbf{a}|\tau,w}))} \\
\nu_1^{-1} &\leftarrow \frac{\sum_{t,j} p(t,w|\mathbf{D}^j)((\mathbf{x}_1-\lambda_1\mu_{\mathbf{a}|\tau,w})^2+\lambda_1^2Tr(\nu^{-1}_{\mathbf{a}|w}))}{N_f\sum_j p(w|\mathbf{D}^j)} \\
v_2^{-1} &\leftarrow \frac{\sum_{t,j} p(t,w|\mathbf{D}^j)((\mathbf{x}_2-\lambda_2\mathbf{T}_\tau\mu_{\mathbf{a}|\tau,w})^2+\lambda_2^2Tr(\nu^{-1}_{\mathbf{a}|w}))}{N_f\sum_j p(w|\mathbf{D}^j)} \\
\sigma_i^{-1} &\leftarrow \frac{\sum_j p(\overline{w}|\mathbf{D}^j)\mathbf{x}_i^T\mathbf{x}_i}{N_f\sum_j p(\overline{w}|\mathbf{D}^j)}
\end{aligned}$$

Audio-visual link parameters

$$
\begin{aligned}
\beta &\leftarrow \frac{\sum_{\tau,l,j} p(t,w,l|D^j)(t-\alpha l)}{\sum_j p(w|D^j)} \\
\alpha &\leftarrow \frac{\sum_{t,j,l} p(\tau,w,l|\mathbf{D}^j)(lt - l\frac{\sum_{t,j} p(t,w|\mathbf{D}^j)t}{\sum_j p(w|\mathbf{D}^j)})}{\sum_{j,l} p(w,l|\mathbf{D}^j)(l^2 - l\frac{\sum_{j,l} p(w,l|\mathbf{D}^j)l}{\sum_j p(w|\mathbf{D}^j)})} \\
\omega^{-1} &\leftarrow \sum_{t,j,l} p(t,l,w|D^j)(t^2 - 2\tau\alpha l - 2t\beta + \alpha^2 l^2 + 2\alpha l\beta + \beta^2) / \sum_j p(w|D^j)
\end{aligned}
$$