In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '07) Hyderabad, India (2007)

Structure Inference for Bayesian Multisensory Perception and Tracking

Timothy M. Hospedales Joel J. Cartwright Sethu Vijayakumar
School of Informatics, University of Edinburgh, EH9 3JZ, Scotland, UK
t.hospedales@ed.ac.uk, j.j.cartwright@sms.ed.ac.uk, sethu.vijayakumar@ed.ac.uk

Abstract

We investigate a solution to the problem of multi-sensor perception and tracking by formulating it in the framework of Bayesian model selection. Humans robustly associate 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, Bayesian solution to multi-sensor perception and tracking which accounts for both integration and segregation by explicit probabilistic reasoning about data association in a temporal context. 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-modal perception [Ernst and Banks, 2002; Alais and Burr, 2004] and for machine perception applications [Beal *et al.*, 2003; Perez *et al.*, 2004]. Most of these have considered the simple cases in which the observations are known to be generated from the same latent source, and the task is to make the best estimate of the latent source state by fusing the observations. (We call models assuming such a fused structure pure fusion models.) However, in most real world 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 fusion (integration) or fission (segregation) that may be necessary. This data association problem has been of more long standing interest in the radar community [Bar-Shalom *et al.*, 2005]. The data association may not merely be a nuisance variable required for correct sensor combination. It can be of intrinsic interest for understanding the data. 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 these multi-sensor perception problems 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, inference can be performed simultaneously with parameter estimation for unsupervised learning of perceptual models.

2 Theory

Humans and machines equipped with multiple sensor modalities need to combine information from various senses to obtain an accurate unified perception of the world. Perception requires computing tangible quantities of interest in the world (e.g. people's locations) as well as the association between sensor observations (e.g. who said what). To formalize the perceptual problems faced by a human or machine equipped with multiple sensor modalities, we use a probabilistic generative modelling framework. The task of perception then becomes that of performing *inference* in the generative model, where object states and data association are both inferred.

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 $(\mathrm{M}_1, \mathrm{M}_2)$ specifying its visibility in each modality. The observations are then generated with x_i being dependent on l if $\mathrm{M}_i = 1$ or on a background distribution if $\mathrm{M}_i = 0$. Alternately, all the structure options could be explicitly enumerated into four separate models, and the generation process then first selects one of the four generative models before selecting l and generating the observations according to the dependencies encoded in that model. Inference then consists of computing the posterior over the latent state and the generating model (either as specified by 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 $p(\mathrm{M}_i = 1|D)$, which will be large if the observation is likely under the foreground distribution and small if it is better explained by the background distribution.

An Illustrative Example

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

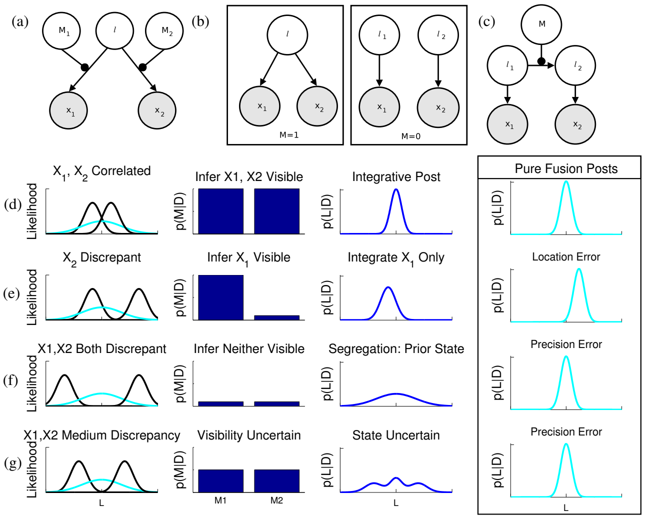


Figure 1: (a) Graphical model to describe un-reliable generation of multi-modal observations x_i, *occlusion semantic* (b,c) generation with one or two source objects, *multi-object semantic* (c-g) Inference in occlusion semantic model. Observations (d) x_1, x_2 strongly correlated, (e) x_2 strongly discrepant, (f) x_1, x_2 both strongly discrepant, (g) x_1, x_2 both moderately discrepant. Likelihoods of the observations in each of two modalities in black, prior in grey.

in separate modalities. For the purpose this illustration, let l be governed by an informative Gaussian prior centered at zero, i.e., $p(l) = \mathcal{N}(l|0, p_l)$ and the binomial visibility variables have prior probability $p(M_i) = \pi_i$. (Note that we use precisions rather than covariances throughout). If the state is observed by sensor i ($\mathrm{M}_i = 1$) 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 precision $p_b \to 0$. The joint probability can then be written as in Eq. 1. 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 [MacKay, 2003]. Defining for brevity $m_i \equiv (M_i = 1)$ and $\overline{m}_i \equiv (M_i = 0)$, we can write down the posteriors as in Eq. 2.

$$p(D, l, \mathrm{M}) = \mathcal{N}(x_1|l, p_1)^{\mathrm{M}_1}\mathcal{N}(x_1|0, p_b)^{(1-\mathrm{M}_1)}\mathcal{N}(x_2|l, p_2)^{\mathrm{M}_2}$$
$$\cdot\mathcal{N}(x_2|0, p_b)^{(1-\mathrm{M}_2)}\mathcal{N}(l|0, p_l)p(\mathrm{M}_1)p(\mathrm{M}_2) \qquad (1)$$

$$\begin{aligned} &p(\overline{m}_1, \overline{m}_2|D) \propto \mathcal{N}(x_1|0, p_b)\mathcal{N}(x_2|0, p_b) \\ &p(m_1, \overline{m}_2|D) \propto exp\left(-\frac{1}{2}x_1^2 p_1 p_l/(p_1 + p_l)\right)\mathcal{N}(x_2|0, p_b) \qquad (2) \\ &p(m_1, m_2|D) \propto \\ &\qquad exp\left[-\frac{1}{2}\frac{x_1^2 p_1(p_2 + p_l) - 2x_1 x_2 p_1 p_2 + x_2^2 p_2(p_1 + p_l)}{p_1 + p_2 + p_l}\right] \end{aligned}$$

Intuitively, the structure posterior (Eq. 2) 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 model depends on the three way agreement between the observations and the prior. The assumption of a one dimensional Gaussian prior and likelihoods is to facilitate illustrative analytical solutions; this is not in general a restriction of our framework as can be seen in Sec. 3.

Fig. 1 illustrates a schematic of some informative types of behavior produced by this model. If the data and the prior are all strongly correlated (Fig. 1(d)) such that both observations are inferred with near certainty to be associated with the latent source of interest, the fused posterior over the location is approximately Gaussian with $p(l|x_1, x_2) \approx \mathcal{N}(l|\hat{l}, p_{l|x})$ 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. 1(e)), it would be inferred that sensor 2 was occluded and its observation irrelevant. In this case, the posterior over the location 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. 1(f)), both observations are likely to be background originated, 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. 1(g)) such that the posterior over the structural visibility variables is not near certain, then the posterior over the latent state is a (potentially quad-modal) mixture of Gaussians corresponding to the 4 possible models. For real world data, occlusion, or other cause for meaningless observation is almost always possible, in which case assuming a typical pure fusion model (equivalent to constraining $\mathrm{M}_1 = \mathrm{M}_2 = 1$) can result in dramatically inappropriate inference (Fig. 1(box)).

Incorporating Temporal Dependencies

Now we consider the case where the latent state of interest and data association are correlated in time. A graphical model to describe the generation of such data is illustrated in Fig. 2(a). The state l and model variables M_i are each connected through time, producing a factorial hidden Markov model [Ghahramani and Jordan, 1997]. To generate from this model, at each time t the location and model variables are selected on the basis of their states at the previous time and the transition probabilities $p(l^{t+1}|l^t)$ and $p(\mathrm{M}_i^{t+1}|\mathrm{M}_i^t)$. Conditional on these variables, each observation is then generated in the same way as for the previous independently and identically distributed (IID) case. Inference may then consist of computing the posterior over the latent variables at each time given all T available observations, $p(l^t, \mathrm{M}^t|x_1^{1:T}, x_2^{1:T})$ (i.e., smoothing) if processing is off-line. If the processing must be on-line, the posterior over the latent variables given all the data up to the current time $p(l^t, \mathrm{M}^t|x_1^{1:t}, x_2^{1:t})$ (i.e., filtering) may be employed. Multi-modal source tracking is performed by computing the posterior of l, marginalizing over possible associations. We have seen previously that the posterior distribution over location at a given time is potentially non-Gaussian (Fig. 1(e)). To represent such general distributions, we can discretize the state space of l. In this simple case, exact numerical inference on the discretized distribution is tractable. Given state transition matrices $p(l^{t+1}|l^t)$ and $p(\mathrm{M}^{t+1}|\mathrm{M}^t)$, we can write down recursions for inference in this factorial hidden Markov model in terms of the posteriors $\alpha^t \triangleq p(l^t, \mathrm{M}^t_{1,2}|D^{1:t})$ and $\gamma^t \triangleq p(l^t, \mathrm{M}^t_{1,2}|D^{1:T})$ as

$$\alpha^t \propto \tag{3}$$

$$\sum_{l^{t-1},M^{t-1}_{1,2}} p(D^t|l^t, M^t_{1,2}) p(l^t|l^{t-1}) \prod_{i=1}^{2} p(M^t_i|M^{t-1}_i)\alpha^{t-1}$$

$$\gamma^t \propto \tag{4}$$

$$\sum_{l^{t+1},M^{t+1}_{1,2}} \frac{p(l^{t+1}|l^t)\prod_{i=1}^{2} p(M^t_i|M^{t-1}_i)\alpha^t}{\sum_{l^t,M^t_{1,2}} p(l^{t+1}|l^t)\prod_{i=1}^{2} p(M^t_i|M^{t-1}_i)\alpha^t}\gamma^{t+1}$$

Filtering makes use of the forward α recursion in Eq. 3 and smoothing the backward γ recursion in Eq. 4, which are analogues of the α and γ recursions in standard HMM inference. The benefits of temporal context for inference of source state and data association are illustrated in Fig. 2(b-g). Fig. 2(b) illustrates data from a series of T observations, $x^t_i \sim \mathcal{N}(l^t, p)$, $D = \{x^t_1, x^t_2\}^T_{t=1}$, in two independent modalities, of a continuously varying latent source l. These data include some occlusions/sensor failures, where the observation(s) are generated from a background distribution, and an unexpected discontinuous jump of the source. The temporal state evolution models for l and M are simple diffusion models. The robustness of source location inference by a pure fusion model without temporal context (Fig. 2(e)) is very limited, as it must always averages over observations, which is inappropriate when they are actually disassociated. A data association model (Fig. 2(f)) is slightly more robust, inferring that the generation structure was likely to have been different when the observations are discrepant. However, without temporal context, it cannot identify which observation was discrepant, and hence produces a non-Gaussian, multi-modal posterior for l. Including some temporal history, an on-line filtering data association model can infer which observations are discrepant, and discount them, producing much smoother inference (Fig. 2(g)). In this case, after the discontinuity in state, the fully disassociated observation structure is inferred and based on the temporal diffusion model, approximately constant location is inferred until enough evidence is accumulated to support the new location. Finally, an off-line smoothing data association model (Fig. 2(c)) infers a robust, accurate trajectory. For this case, the marginal posterior of the association variables is shown in Fig. 2(d). The illustrative scenarios discussed here generalize in the obvious way to more observations. With many sensors, the dis-association of a smaller number of discrepant sensors can be inferred even without prior information. In a pure fusion scheme, a single highly discrepant sensor can throw off the others during averaging.

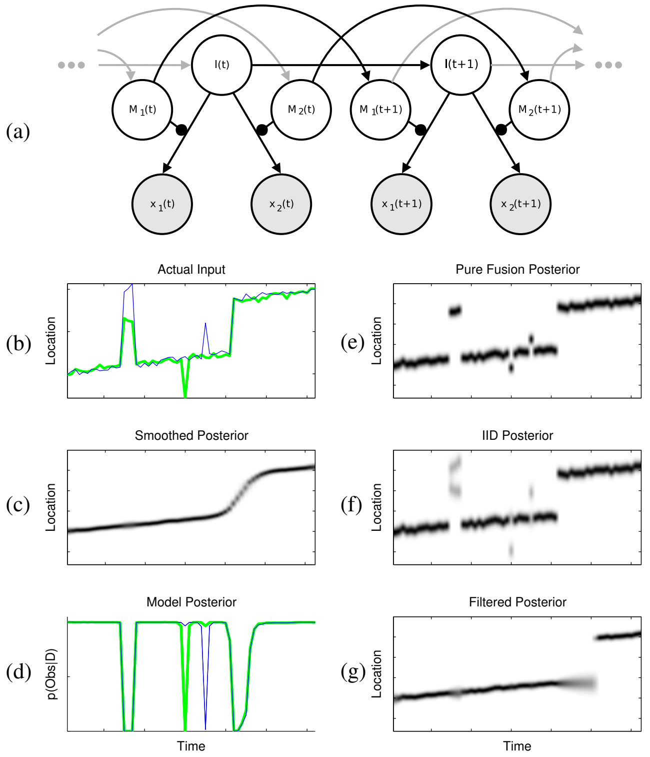


Figure 2: (a) Graphical model to describe generation of observations x_i with temporal dependency. (b) Synthetic input dataset in modality x_1 and x_2. (c) Posterior probability of l and (d) posterior probability of model structure for the temporal data association model. Posterior probability of l in (e) pure fusion model (f) IID data association model (g) filtered data association model.

An Illustrative Example with Multiple Objects

There is another simple way in which two multi-modal 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, frequently 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[Shams *et al.*, 2000] 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

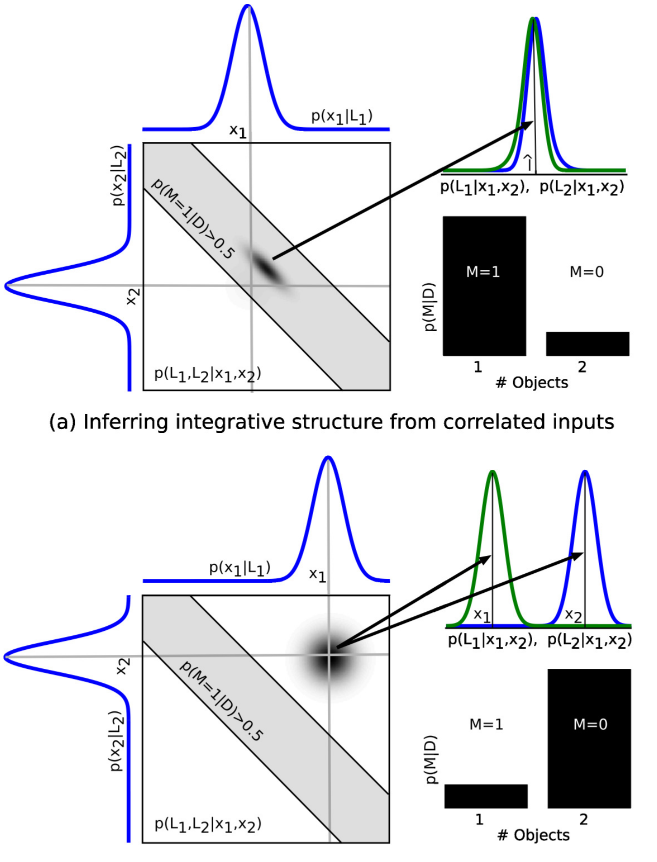
$\mathbf{P}_x = diag([p_1, p_2])$. Let us assume the prior distributions over the latent locations are Gaussian but tend to uninformativeness. In the multi-object model the prior over l_is $p(\mathbf{l}|M = 0) = \mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_0)$ is uncorrelated, so $\mathbf{P}_0 = p_0\mathbf{I}$ and $p_0 \rightarrow 0$. In the single object model, the prior over l_is $p(\mathbf{l}|M = 1) = \mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_1)$ requires the l_is to be equal so $\mathbf{P}_1$ is chosen to be strongly correlated. The joint probability of the whole model and the structure posterior are given in Eq. 5.

$$
\begin{aligned}
&p(\mathbf{x}, \mathbf{l}, M) = \mathcal{N}(\mathbf{x}|\mathbf{l}, \mathbf{P}_x)\mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_0)^{(1-M)}\mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_1)^M p(M) \\
&p(M|\mathbf{x}) \propto \int_{\mathbf{l}} \mathcal{N}(\mathbf{x}|\mathbf{l}, \mathbf{P}_x)\mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_0)^{(1-M)}\mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_1)^M p(M) \\
&p(M = 0|\mathbf{x}) \propto \mathcal{N}(\mathbf{x}|\mathbf{0}, (\mathbf{P}_x^{-1} + \mathbf{P}_0^{-1})^{-1})p(M = 0) \\
&p(M = 1|\mathbf{x}) \propto \mathcal{N}(\mathbf{x}|\mathbf{0}, (\mathbf{P}_x^{-1} + \mathbf{P}_1^{-1})^{-1})p(M = 1) \qquad (5)
\end{aligned}
$$

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; $p(\mathbf{l}|\mathbf{x}) \approx \mathcal{N}(\mathbf{l}|\hat{\mathbf{l}}, \mathbf{P}_{l|x})$ 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. $\hat{\mathbf{l}} = \mathbf{P}_{l|x}^{-1}\mathbf{P}_x\mathbf{x} \approx \mathbf{x}$, $\mathbf{P}_{l|x} = \mathbf{P}_x + \mathbf{P}_0$.

The inferences discussed so far have been exact. There are various potential approximations such as computing the *location posterior* given the MAP model, which may be acceptable, but which crucially misrepresents the state posterior for regions of input space with intermediate discrepancy (c.f. Fig. 1(d)). Alternately, the *model probability* could be approximated using a MAP or ML estimate of the state. The agreement between the Bayesian and MAP solution depends on how sharp the state posterior is, which depends on both the agreement between the observations and the precision of their likelihoods. Using the ML estimate of the state will not work at all as the most complex model is always selected.

Previous probabilistic accounts of human multi-sensory combination (e.g. [Ernst and Banks, 2002; Alais and Burr, 2004]) are special cases of our theory, having explicitly or implicitly assumed a pure fusion structure. [Triesch and von der Malsburg, 2001] describe a heuristic democratic *adaptive* cue integration perceptual model, but again assume 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 [Ernst and Banks, 2002]. As we have seen, such fission is necessary for perception in the real world as outliers can break pure fusion schemes. We provide a principled probabilistic, adaptive theory of temporal 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. 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.



(a) Inferring integrative structure from correlated inputs

(b) Inferring segregative structure from decorrelated inputs

Figure 3: Inference in multi-object semantic toy model. (a) For correlated inputs, $x_1 \approx x_2$, the presence of one objects 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.

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 [Beal *et al.*, 2003]; that of unsupervised learning and inference with audio-visual (AV) input. [Beal *et al.*, 2003] demonstrated inference of an AV source location and learning of its auditory and visual templates based on correlations between the input from a camera and two microphones - useful for example, in teleconferencing applications . The AV localization part of this task is similar to the task required in psychophysics experiments such as [Alais and Burr, 2004] where humans are also reported to exhibit near Bayes optimal sensor fusion. We now tackle the bigger scene understanding problem of inferring how the AV data should be associated through time (pure fusion was previously assumed), i.e, whether the source should be associated with both modalities,

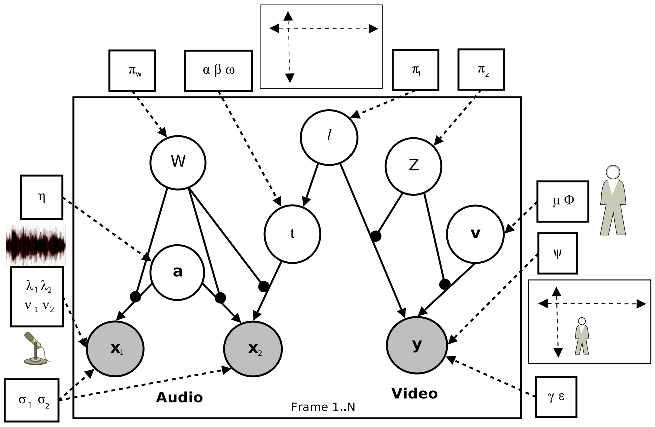


Figure 4: Graphical model for AV data generation.

or only one, or if there is no source present at all.

3.1 Introduction

A graphical model to describe the generation of a *single* frame of AV data $D = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}\}$ is illustrated in Fig. 4. A 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 $\mathcal{N}(t|\alpha l + \beta, \omega)$. Given the latent signal and the delay, the observation $\mathbf{x}_i$ at each microphone is generated by sampling from a uniform diagonal Gaussian with the mean $\mathbf{a}$, with $\mathbf{x}_2$ shifted t samples relative to $\mathbf{x}_1$; $\mathcal{N}(\mathbf{x}_1|\mathbf{a}, v_1\mathbf{I})$, $\mathcal{N}(\mathbf{x}_2|\mathbf{T}_t\mathbf{a}, v_2\mathbf{I})$. If the video modality is 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 l and audio data. If the audio modality is silent $(\mathrm{W} = 0)$, the samples at each speaker are drawn from background distributions $\mathcal{N}(\mathbf{x}_i|\mathbf{0}, \sigma_i\mathbf{I})$ independently of each other, l and the video.

To describe the generation of a series of correlated frames, the IID observation model in Fig. 4 is replicated and a factored Markov model is defined over the location and association variables $(l, \mathrm{W}, \mathrm{Z})$ exactly as the toy model was developed previously (refer Fig. 2(a)). Using j to index time, the state evolution distribution over the location shift is defined in the standard way $p(l^{j+1}|l^j) = \Gamma_{l^j, l^{j+1}}$, where the subscripts pick out the appropriate element of the matrix Γ. The observability transitions are defined similarly as $p(\mathrm{W}^{j+1}|\mathrm{W}^j) = \Theta_{w^j, w^{j+1}}$ and $p(\mathrm{Z}^{j+1}|\mathrm{Z}^j) = \Omega_{z^j, z^{j+1}}$. Suppressing indexing by j for clarity, the joint probability of the model including visible $D = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}\}_{j=1}^J$ and hidden variables $H = \{\mathbf{a}, \mathbf{v}, t, l, W, Z\}_{j=1}^J$ factorizes as

$$\begin{aligned} p(D, H) = &\prod_{j=1}^{J-1} p(l^{j+1}|l^j)p(W^{j+1}|W^j)p(Z^{j+1}|Z^j) \\ &\cdot \prod_{j=1}^{J} p(\mathbf{x}_1|W, \mathbf{a})p(\mathbf{x}_2|W, \mathbf{a}, t)p(\mathbf{a})p(t|l)p(\mathbf{v})p(\mathbf{y}|Z, \mathbf{v}, l) \\ = &\Bigg(\prod_{j=1}^{J} \mathcal{N}(\mathbf{x}_1|\mathbf{a}, v_1)^W \mathcal{N}(\mathbf{x}_1|\mathbf{0}, \sigma_1)^{(1-W)} \mathcal{N}(\mathbf{a}|0, \eta) \\ &\cdot \mathcal{N}(\mathbf{x}_2|T_t\mathbf{a}, v_2)^W \mathcal{N}(\mathbf{x}_2|\mathbf{0}, \sigma_2)^{(1-W)} \mathcal{N}(\tau|\alpha l + \beta, \omega) \\ &\cdot \mathcal{N}(\mathbf{y}|T_l\mathbf{v}, \Psi\mathbf{I})^Z \mathcal{N}(\mathbf{y}|\gamma\mathbf{1}, \epsilon\mathbf{I})^{(1-Z)} \mathcal{N}(\mathbf{v}|\mu, \phi)\Bigg) \\ &\cdot \prod_{j=1}^{J-1} \Gamma_{l^j, l^{j+1}} \Theta_{w^j, w^{j+1}} \Omega_{z^j, z^{j+1}} \qquad (6) \end{aligned}$$

3.2 Inference

Consider first the inference for a single frame of data. The posterior marginal of interest for this 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 out, leaving only the inter-microphone delay t to be summed out numerically when computing $p(l, \mathrm{W}, \mathrm{Z}|D)$. 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 $\mu_{\mathbf{a}|\mathbf{x},t} = \nu_{\mathbf{a}}^{-1}(\lambda_1\nu_1\mathbf{x}_1 + \lambda_2\nu_2\mathbf{T}_t^T\mathbf{x}_2)$, $\nu_{\mathbf{a}} = \eta + \lambda_1^2\nu_1 + \lambda_2^2\nu_2$, $\mu_{\mathbf{v}|\mathbf{y},l} = \nu_{\mathbf{v}}^{-1}(\phi\mu + \mathbf{T}_l^T\Psi\mathbf{y})$, $\nu_{\mathbf{v}} = \phi + \Psi$. The marginal video likelihood is also Gaussian with $\mu_{\mathbf{y}|l} = \mathbf{T}_l\mu$, $\nu_{\mathbf{y}|l} = (\Psi^{-1} + \mathbf{T}_l\phi^{-1}\mathbf{T}_l^T)^{-1}$. Expressions for the likelihood of the fully fused model and the source location (Eq. 7) and the likelihood of the fully fissioned model (Eq. 8) can be derived in terms of these statistics. Defining again $z_i \equiv (Z_i = 1)$ and $\overline{z}_i \equiv (Z_i = 0)$ etc, we can write

$$\begin{aligned} p(D|w, z, l) &\propto \int_{\mathbf{v}} p(\mathbf{y}, \mathbf{v}|l, z) \sum_t \int_{\mathbf{a}} p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{a}, t|l, w) \\ &\propto \mathcal{N}(\mathbf{y}|\mu_{\mathbf{y}|l}, \nu_{\mathbf{y}|l}) \sum_t p(t|l, D) exp(\mu_{\mathbf{a}|t,\mathbf{x}}^T \nu_{\mathbf{a}} \mu_{\mathbf{a}|t,\mathbf{x}}) \quad (7) \\ p(D|\overline{w}, \overline{z}) &\propto p(\mathbf{x}|\overline{w})p(\mathbf{y}|\overline{z}) \\ &= \mathcal{N}(\mathbf{x}_1|\mathbf{0}, \sigma_1\mathbf{I})\mathcal{N}(\mathbf{x}_2|\mathbf{0}, \sigma_2\mathbf{I})\mathcal{N}(\mathbf{y}|\gamma\mathbf{1}, \epsilon\mathbf{I}) \qquad (8) \end{aligned}$$

For a single observed modality, the likelihood is a mixture of terms from Eqs. 7 and 8, along the lines of Eq. 2. To infer the posterior over location and observability for IID frames, these likelihoods can be combined with the discrete prior distributions over the location and observabilities. To infer the probability of location and observability over time given the data, the likelihoods are used in recursions exactly like those in Eqs. 3 and 4.

3.3 Learning

All the parameters in this model $\theta = \{\lambda_{1,2}, \nu_{1,2}, \eta, \alpha, \beta, \omega, \pi_l, \mu, \phi, \Psi, \Gamma, \Theta, \Omega, \pi_w, \pi_z, \gamma, \epsilon, \sigma_{1,2}\}$ are jointly optimized by a standard EM procedure of alternately inferring the posterior distribution $q(H|D)$ over hidden variables H given the observed data D, 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[1]. Eq. 9 gives the update for the mean μ of the source visual appearance distribution. This is defined in terms of the posterior mean $\mu^j_{\mathbf{v}|\mathbf{y},l}$ of the video appearance given the data D for each frame j and translation l, as inferred during the E step:

$$\mu \leftarrow \sum_{j,l} q(l^j, \mathrm{z}^j|D)\mu^j_{\mathbf{v}|\mathbf{y},l} / \sum_j q(\mathrm{z}^j|D) \qquad (9)$$

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. The scalar precision parameter of background noise is given by Eq. 10, where N_f specifies the number of samples per audio frame.

$$\sigma_i^{-1} \leftarrow \sum_j q(\overline{\mathrm{w}^j}|D)(\mathbf{x}_i^j)^T \mathbf{x}_i^j / N_f \sum_j q(\overline{\mathrm{w}^j}|D) \qquad (10)$$

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 that the source was silent in that frame. In an IID context, the posterior marginals to weight with, e.g. $p(l^j, z^j|D)$, are given visible variable, D^j. In the Markov model context, the posterior marginal given the whole available data set D is used.

3.4 Demonstration

Results for an AV sequence after 25 cycles of EM are illustrated in Fig. 5. In this sequence, the user is initially walking and talking, is then occluded behind another person while continuing to speak, and then continues to walk while remaining silent. Fig. 5(a) illustrates three representative video frames from each of these segments with the inferred data association and location superimposed on each.

Tracking with the IID, pure fusion model

To illustrate tracking behavior, the MAP rather than full location posterior is shown for clarity. In an IID pure fusion model (constrained such that $W, Z = 1$ and with prior π_l instead of transition matrix Γ) the location inference is correct where the multi-modal observations are indeed associated (Fig. 5(c)). The video modality dominates the fusion as it is much higher precision (i.e., the likelihood function is much sharper), and the posterior is still therefore correct during the visible but silent period where peaks in the audio likelihood are spurious. However, while the person in the video foreground is occluded but speaking, the next best match to the learned dark foreground template usually happens to be the filing cabinet in the corner. With pure fusion, the incorrect but still fairly sharp video likelihood still dominates the audio likelihood, resulting in an incorrect posterior.

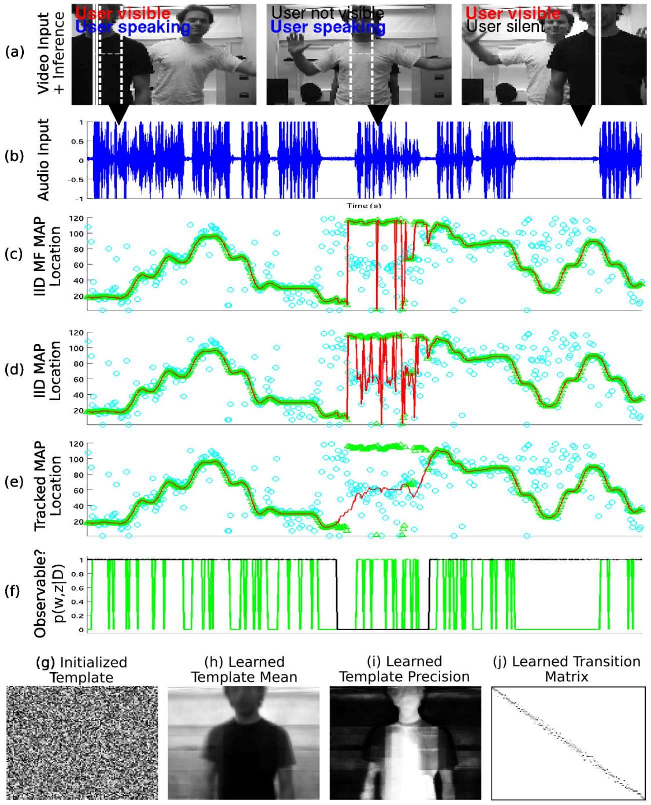


Figure 5: AV data association & inference results. (a) Video samples and (b) audio data from a sequence where the user is first visibly walking and speaking, then occluded but still speaking, and finally visible and walking but silent. (c) Inferred MAP location with IID pure fusion model, (d) IID data association model and (e) full temporal data association model. Inference based on audio observation alone is shown in circles, video observation alone in triangles, and combined inference by the dark line. (f) Posterior probability of visibility (dark) and audibility (light) during the sequence. (g) Initial and (h,i) final video appearance after learning. (j) Final location state transition matrix after learning.

Tracking with IID data association model

In an IID data association model (Fig. 5(d)) the video modality is correctly inferred with high confidence to be disassociated during the occluded period. The final posterior is therefore based mostly on the audio likelihood, and is generally peaked around the correct central region of the azimuth. The outlier points here have two causes. As speech data is intrinsically intermittent, *both* modalities occasionally have low probability of association, during which times the final estimate is still inappropriately attracted to that of the video modality as in the pure fusion case. Others are simply due to the lower inherent precision of the audio modality.

Tracking with smoothed data association model

The data association posterior (W, Z) (Fig. 5(f)) correctly represents the visibility and audibility of the target at the appropriate times, as with the IID case. This enables in-

[1] Complete derivations, video clips & matlab code are available at http://homepages.inf.ed.ac.uk/s0238587/

formation from the appropriate sensor(s) to be used at each time. With the addition of temporal context, tracking based on the noisy and intermittent audio modality is much more reliable in the difficult period of visual occlusion. The user is now reliably and seamlessly tracked during all three domains of the input sequence (Fig. 5(g)). The inferred data association is used to label the frames in (Fig. 5(a)) with the user's speaking/visibility status. To cope with intermittent cues, previous multi-modal machine perception systems in this context have relied on observations of discrepant modalities providing uninformative likelihoods [Perez *et al.*, 2004; Beal *et al.*, 2003]. This may not always be the case, and was not, for example in our video sequence where only the data association models succeeded during the video occlusion. This model retains properties of the inspiring formulation [Beal *et al.*, 2003] which allow most of the expensive E and M step computations in the observation model to be expressed in terms of FFTs. Using 120x100 pixel images and 1000 sample audio frames, our matlab implementation can perform on-line real time (filtered) tracking at 50fps after (smoothed) learning, which proceeds at 10fps.

4 Discussion

In this paper, we introduced a principled formulation of multi-sensor perception and tracking in the framework of Bayesian inference and model selection in probabilistic graphical models. Pure fusion multi-sensor models 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 fused. In many cases, inferring observation association is in itself an important goal for understanding structure in the data. For example, a speech transcription model should not associate nearby background speech of poorly matching template and uncorrelated spatial location with the visible user when he is silent. In our application the model "knows" if observations arise from the source of interest by explicitly inferring association, so it can for example, start recording when the user enters the scene or begins speaking.

In radar tracking and association, some work[Stone *et al.*, 1999] uses similar techniques to ours, however popular methods [Bar-Shalom *et al.*, 2005] tend to be more heuristic, use stronger assumptions and approximations (e.g. Gaussian posteriors) and use highly pre-processed point-input data. One interesting contrast between these candidate detection based approaches and our generative model approach is that we avoids the expensive within-modality data association problem typical of radar. This also enables use of signature or template information in a unified way along with cross-modality correlation during inference, which is exploited to good effect in our AV application.

Investigations of human multi-sensory perception have reported robustness to discrepant cues [Ernst and Banks, 2002] 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. 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 context of machine perception, the type of model described generalizes existing integrative models and provides a principled solution to questions of sensor combination including signature, fusion, fission and association. As our AV application illustrates, computing the exact posterior over source state and data association for real problems, even before applying approximations, is potentially even real-time. The major complicating extension not considered in detail here, is that of multiple sources. 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 objects one at a time in order of best correlation along the lines of [Williams and Titsias, 2004].

References

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

[Bar-Shalom *et al.*, 2005] 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, 2005.

[Beal *et al.*, 2003] Matthew J. Beal, Nebojsa Jojic, and Hagai Attias. A graphical model for audiovisual object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):828–836, July 2003.

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

[Ghahramani and Jordan, 1997] Zoubin Ghahramani and Michael Jordan. Factorial hidden markov models. *Machine Learning*, 29:245–273, 1997.

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

[Perez *et al.*, 2004] Patrick Perez, Jaco Vermaak, and Andrew Blake. Data fusion for visual tracking with particles. *Proceedings of the IEEE*, 92(3):495–513, 2004.

[Shams *et al.*, 2000] Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. Illusions: What you see is what you hear. *Nature*, 408:788, December 2000.

[Stone *et al.*, 1999] Lawrence D. Stone, Carl A. Barlow, and Thomas L. Corwin. *Bayesian Multiple Target Tracking*. Artech House, 1999.

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

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