Bayesian Multisensory Perception and Tracking

Timothy Hospedales, Sethu Vijayakumar — *School of Informatics, The University of Edinburgh, UK*

Introduction

Humans and machines equipped with **multiple sensor modalities** need to combine information from various senses to obtain an accurate, unified perception of the world. Previous research has addressed statistically optimal fusion of multisensory observations of a given object[2,3]. 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. In some domains, these *causal*, association variables may also have critical independent meaning.

Example To fully understand a meeting, two sets of latent object states (*who was there* & *what was said*) and the data association (*who said what*) must be inferred.

Multisensory Structure Inference

These perceptual problems are formalized in a probabilistic generative modelling framework[1] with the following advantages:

- Multisensory integration and segregation are optimal.
- *Bayesian structure inference* or *model selection* is used to infer the optimal data association $p(M|x_1, x_2)$ without heuristics.
  - Bayesian Occam's Razor provides automatic and optimal *complexity control*.

Figure 1: Graphical models for variable structure multisensory perception.

Data Association in Multiple Modalities

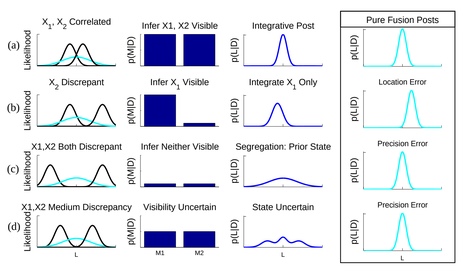

Figure 2: Inference in data association vs Pure fusion models. Left Column: Input: Prior + two observations of unknown association.

The **unknown structure and state problem** is illustrated by the graphical model in Fig 1a for a single object observable with two unreliable/occludeable sensors. For the illustration in Fig. 2, we use 1D linear gaussian prior and observation likelihoods.

Pure Fusion models assume all sensors always observe the source. This can result in incorrect inference (Fig 2, box)

Data Association models infer the model structure as well as source state, obtaining the correct posteriors. (Fig 2a-d)

- But how to choose the prior distribution?
  - Use *temporal context*!

Introducing Temporal Dependencies

In the real world, we also expect objects' states and observability to be correlated in time. We should take advantage of this important prior knowledge. Extended to take into account temporal dependency, our generative model has the form of a *factored hidden Markov model* (FHMM) as illustrated in Fig 3a.

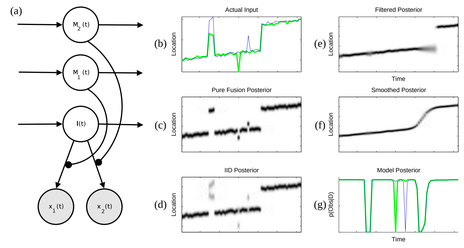

Figure 3: Toy Markov model and inference results

To illustrate, consider tracking a source through time in 1-D using *two* one dimensional linear Gaussian observations (Fig 3b).

Pure Fusion: Model fails with sensor failure/occlusion (Fig 3c).

IID Data Association: Model "knows" something is wrong during sensor failure/occlusion as the sensors do not agree with each other. But without temporal context, it does not know which, if any, sensor to believe (Fig3d).

FHMM Data Association: Model "knows" when source is visible with each sensor (Fig 3g), and hence, is able to base inference on the non-occluded sensor or, if both fail, on temporal history alone. Smoothing (Fig 3e) or filtering (Fig 3f) are possible.

Data Association with Multiple Objects

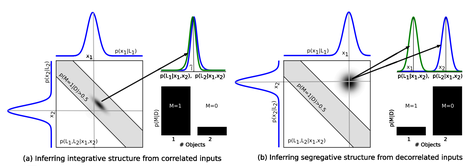

Figure 4: Inferring multiple objects with multiple sensors

Another way of generating multiple observations is for each to be generated by a *different* object of interest. An example of this paradigm occurs during humans psychophysics experiments[2], where the subject must decide if two slightly discrepant multimodal observations (e.g., flashes and beeps) are generated by one or two sources. This is formalized as a one or two source model selection problem (Fig 1b,c).

- The two source model is more *complex*. Additional *degrees of freedom* allow it to fit any data better than the one source model.
- The **maximum likelihood** estimate of the number of objects is therefore *always* the more complex two source explanation.
- In contrast, in our method, the automatic complexity control of *Bayesian Occam's Razor* provides the optimal solution – it optimally weights the explanatory power of the two object model against its increased complexity (Fig 4a,b).
- **Human behaviour** is explained only by the Bayesian solution.

Typical tracking techniques require number of targets to be prespecified or determined heuristically. We can *infer the number of targets* and *track them* all within the same framework.

Audio-Visual Tracking Application

We illustrate the application of these ideas to a real, large scale machine perception problem by considering an unsupervised learning, inference and tracking task with audio-visual input[1]. A more complex generative model is needed to describe the high dimensional data (shown in Fig 5). The basic framework is that of the *Transformed Mixtures of Gaussians* [3], which describes the video as a translated template and multi-microphone audio as phase-offset signals. Position (l) and AV visibility structure (W,Z) are correlated in time as in Fig 3a.

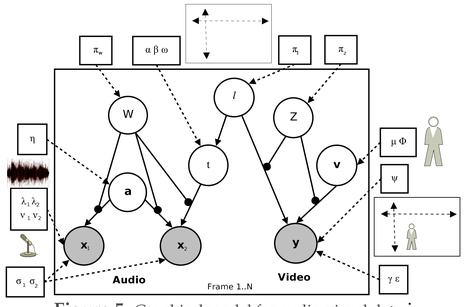

Figure 5: Graphical model for audio-visual data

Inference & Learning

The posterior over hidden variables $H = \{W, Z, l, t, \mathbf{a}, \mathbf{v}\}$ given data $D = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}\}$ is inferred. The filtered posterior $p(W^t, Z^t, l^t|D^{1:t})$ is computed for tracking. Computations are tractable as the $\mathbf{a}$ and $\mathbf{v}$ integrals are solved analytically, and the others are expressible as FFTs. The likelihood of a frame is

$$p(D|w,z,l) \propto \int_{\mathbf{v}} p(\mathbf{y},\mathbf{v}|l,z) \sum_t \int_{\mathbf{a}} p(\mathbf{x}_1,\mathbf{x}_2,\mathbf{a},t|l,w)$$
$$\propto \mathcal{N}(\mathbf{y}|\mu_{\mathbf{y}|l},\nu_{\mathbf{y}|l}) \sum_t p(t|l,D) exp(\mu^T_{\mathbf{a}|t,\mathbf{x}} \nu_{\mathbf{a}|t,\mathbf{x}})$$
$$p(D|\overline{w},\overline{z}) \propto p(\mathbf{x}|\overline{w}) p(\mathbf{y}|\overline{z}) = \mathcal{N}(\mathbf{x}_1|\mathbf{0},\sigma_1\mathbf{I})\mathcal{N}(\mathbf{x}_2|\mathbf{0},\sigma_2\mathbf{I})\mathcal{N}(\mathbf{y}|\gamma\mathbf{1},\epsilon\mathbf{I})$$

EM is used for unsupervised learning of all the parameters θ: $\theta = \{\lambda_{1,2}, \nu_{1,2}, \eta, \alpha, \beta, \omega, \pi_l, \mu, \phi, \Psi, \Gamma, \Theta, \Omega, \pi_w, \pi_z, \gamma, \epsilon, \sigma_{1,2}\}$
Selected examples of parameter updates look like:

$$\mu \leftarrow \sum_{j,l} q(l^j, \mathbf{z}^j|D) \mu^j_{\mathbf{v}|\mathbf{y},l} / \sum_j q(\mathbf{z}^j|D) \quad \sigma_i^{-1} \leftarrow \sum_j q(\overline{\mathbf{w}^j}|D)(\mathbf{x}_i^j)^T \mathbf{x}_i^j / N_f \sum_j q(\overline{\mathbf{w}^j}|D)$$

Results

Using audio visual data (e.g. Fig 6a,b) we wish to learn and track the user through visual and auditory occlusion[1].

Pure Fusion: Inappropriately assuming all data is associated, the location inference is incorrect during visual occlusion (Fig 6c).

IID Association: When audio is available during video occlusion tracking continues, but fails when neither are available (Fig 6d)

FHMM Association: Temporal context allows continuous tracking through audio and visual occlusion (Fig 6e). Structure inference ($p(W, Z|D)$, Fig 6f) allows the labelling of the relational content in the data (AV speaker detection & verification) as is annotated on the images in Fig 6a.

Learning: From a random initialization, the model learns the object (video template and audio spectrum) to be tracked based on sensor correlation only (Fig 6g-i)

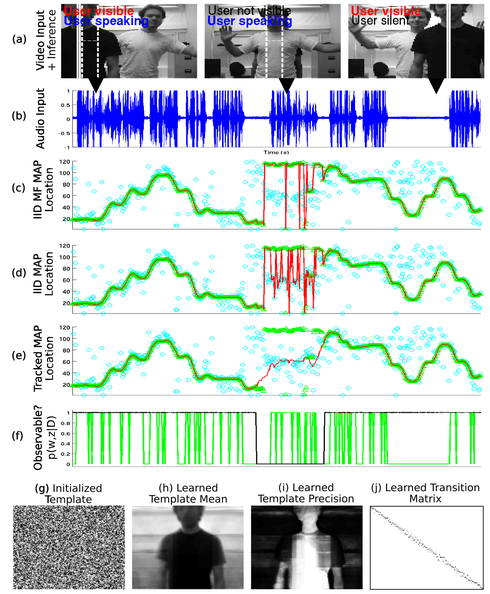

Figure 6: AV Learning, Tracking, Data Association

Discussion

Features and novel contributions include

- Principled formulation of multisensor perception and tracking using *Bayesian model selection*.
- *Automatic learning* of all parameters from data.
- *Real-time* (50fps) inference of structure and state performs simultaneous user detection, tracking and multisensory verification (that detected sound comes from the user).

Outlook: Using sophisticated probabilistic techniques such as those described here, future probabilistic modelling research in machine learning and neuroscience will increasingly be able to deal with higher level existential and relational concepts in data.

Future plans include

- Development of approximate inference techniques to allow extension of audio-visual model to multiple users and more complex relational scenarios.
- Experimental testing of whether human optimal sensor fusion[2] extends to optimal multisensor association.
- Closing the sensorimotor loop. Investigating Bayesian models of active learning & perception.

References

[1] T. Hospedales and S. Vijayakumar, Bayesian Multisensory Perception and Tracking, *IJCAI*, 2007.

[2] L. Shams et al, What you see is what you hear, *Nature*, 2000.

[3] M. Beal, N. Jojic and Hagai Attias, A Graphical Model for Audiovisual Object Tracking *PAMI*, 2003.

Acknowledgements: This research is supported partially by the EPSRC/MRC DTC studentship grant to TH and Royal Society Grant R37647 to SV. Contact Author: Timothy Hospedales (t.hospedales@ed.ac.uk) for further details of the paper.